Introduction to Machine Learning

Feature Selection

Andres Mendez-Vazquez

June 14, 2020

Outline

Outline

What is this?

Main Question

“Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? “

What is this?

Main Question

"Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

Why is important?

1. If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance.
2. if information-rich features are selected, the design of the classifier can be greatly simplified.

1. Large between-class distance.
2. Small within-class variance.

What is this?

Main Question

"Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

Why is important?

1. If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance.
2. if information-rich features are selected, the design of the classifier can be greatly simplified.

Therefore

We want features that lead to

1. Large between-class distance.
2. Small within-class variance.

What is this?

Main Question

"Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

Why is important?

1. If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance.
2. if information-rich features are selected, the design of the classifier can be greatly simplified.

Therefore

We want features that lead to

1. Large between-class distance.
2. Small within-class variance.

What is this?

Main Question

"Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

Why is important?

1. If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance.
2. if information-rich features are selected, the design of the classifier can be greatly simplified.

Therefore

We want features that lead to

1. Large between-class distance.
2. Small within-class variance.

What is this?

Main Question

"Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

Why is important?

1. If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance.
2. if information-rich features are selected, the design of the classifier can be greatly simplified.

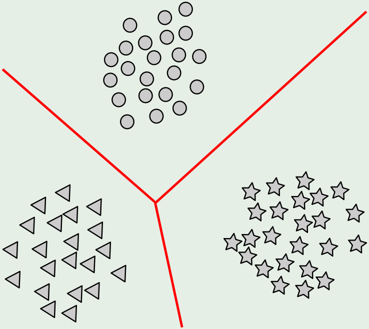
Therefore

We want features that lead to

1. Large between-class distance.
2. Small within-class variance.

Then

Basically, we want nice separated and dense clusters!!!



Outline

However, Before That...

It is necessary to do the following

1. Outlier removal.
2. Data normalization.
3. Deal with missing data.

However, Before That...

It is necessary to do the following

1. Outlier removal.
2. Data normalization.
3. Deal with missing data.

Actually

PREPROCESSING!!!

However, Before That...

It is necessary to do the following

1. Outlier removal.
2. Data normalization.
3. Deal with missing data.

Actually

PREPROCESSING!!!

However, Before That...

It is necessary to do the following

1. Outlier removal.
2. Data normalization.
3. Deal with missing data.

Actually

PREPROCESSING!!!

Outline

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Example

For a normally distributed random

1. A distance of two times the standard deviation covers 95% of the points.
2. A distance of three times the standard deviation covers 99% of the points.

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Example

For a normally distributed random

1. A distance of two times the standard deviation covers 95% of the points.
2. A distance of three times the standard deviation covers 99% of the points.

Note

Points with values very different from the mean value produce large errors during training and may have disastrous effects. These effects are even worse when the outliers, and they are the result of noisy measureme

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Example

For a normally distributed random

1. A distance of two times the standard deviation covers 95% of the points.
2. A distance of three times the standard deviation covers 99% of the points.

Note

Points with values very different from the mean value produce large errors during training and may have disastrous effects. These effects are even worse when the outliers, and they are the result of noisy measureme

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Example

For a normally distributed random

1. A distance of two times the standard deviation covers 95% of the points.
2. A distance of three times the standard deviation covers 99% of the points.

Note

Points with values very different from the mean value produce large errors during training and may have disastrous effects. These effects are even worse when the outliers, and they are the result of noisy measureme

Outliers

Definition

An outlier is defined as a point that lies very far from the mean of the corresponding random variable.

Note: We use the standard deviation

Example

For a normally distributed random

1. A distance of two times the standard deviation covers 95% of the points.
2. A distance of three times the standard deviation covers 99% of the points.

Note

Points with values very different from the mean value produce large errors during training and may have disastrous effects. These effects are even worse when the outliers, and they are the result of noisy measureme

Outlier Removal

Important

Then removing outliers is the biggest importance.

Outlier Removal

Important

Then removing outliers is the biggest importance.

Therefore

You can do the following

1. If you have a small number $\Rightarrow$ discard them!!!
2. Adopt cost functions that are not sensitive to outliers:
   1. For example, possibilistic clustering.
3. For more techniques look at
   1. Huber, P.J. "Robust Statistics," JohnWiley and Sons, 2nd Ed 2009.

Outlier Removal

Important

Then removing outliers is the biggest importance.

Therefore

You can do the following

1. If you have a small number $\Rightarrow$ discard them!!!
2. Adopt cost functions that are not sensitive to outliers:
   1. For example, possibilistic clustering.
3. For more techniques look at
   1. Huber, P.J. "Robust Statistics," JohnWiley and Sons, 2nd Ed 2009.

Outlier Removal

Important

Then removing outliers is the biggest importance.

Therefore

You can do the following

1. If you have a small number $\Rightarrow$ discard them!!!
2. Adopt cost functions that are not sensitive to outliers:
    1. For example, possibilistic clustering.
3. For more techniques look at
    1. Huber, P.J. "Robust Statistics," JohnWiley and Sons, 2nd Ed 2009.

Outlier Removal

Important

Then removing outliers is the biggest importance.

Therefore

You can do the following

1. If you have a small number $\Rightarrow$ discard them!!!
2. Adopt cost functions that are not sensitive to outliers:
    1. For example, possibilistic clustering.
3. For more techniques look at
    1. Huber, P.J. "Robust Statistics," JohnWiley and Sons, 2nd Ed 2009.

Outlier Removal

Important

Then removing outliers is the biggest importance.

Therefore

You can do the following

1. If you have a small number $\Rightarrow$ discard them!!!
2. Adopt cost functions that are not sensitive to outliers:
   1. For example, possibilistic clustering.
3. For more techniques look at
   1. Huber, P.J. “Robust Statistics,” JohnWiley and Sons, 2nd Ed 2009.

Outline

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than

$$\chi_d^2(0.05)$$

4. Return O.

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than
$$\chi_d^2(0.05)$$
4. Return O.

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than

$$\chi_d^2(0.05)$$

4. Return O.

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than

$$\chi^2_d(0.05)$$

4. Return O.

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than

$$\chi_d^2 (0.05)$$

4. Return O.

We can do the following

Algorithm

Input: An $N \times d$ data set $Data$

Output: Candidate Outliers

1. Calculate the sample mean μ and sample covariance matrix Σ.
2. Let M be $N \times 1$ vector consisting of square of the Mahalonobis distance to μ.
3. Find points O in M whose values are greater than

$$\chi_d^2 (0.05)$$

4. Return O.

How?

Get the Sample Mean per feature k

$$\boldsymbol{m}_i = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_{ki}$$

Get the Sample Variance per feature k

$$v_i = \frac{1}{N-1}\sum_{k=1}^{N} \left(\boldsymbol{x}_{ki} - \boldsymbol{m}_i\right)\left(\boldsymbol{x}_{ki} - \boldsymbol{m}_i\right)^T$$

How?

Get the Sample Mean per feature k

$$\boldsymbol{m}_i = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_{ki}$$

Get the Sample Variance per feature k

$$v_i = \frac{1}{N-1}\sum_{k=1}^{N} \left(\boldsymbol{x}_{ki} - \boldsymbol{m}_i\right)\left(\boldsymbol{x}_{ki} - \boldsymbol{m}_i\right)^T$$

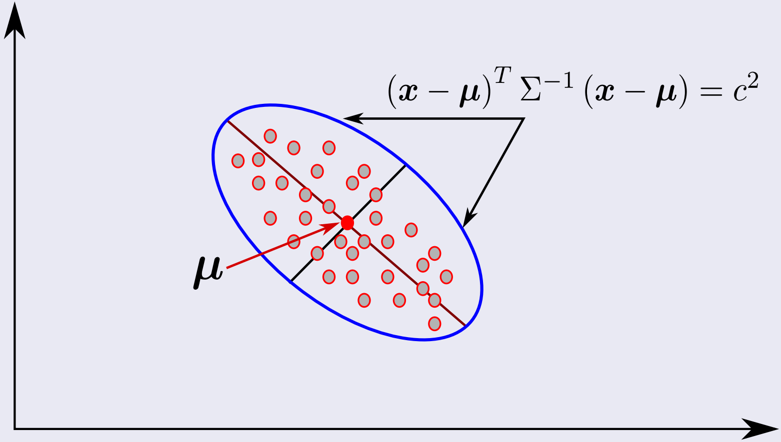
Mahalonobis Distance

We have

$$M\left(\boldsymbol{x}\right)=\sqrt{\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{T}\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}$$

Thus

Setting $M(\boldsymbol{x})$ to a constant c defines a multidimensional ellipsoid with centroid at $\boldsymbol{\mu}$



As Johnson and Wichern (2007, p. 155, Eq. 4-8) state

The solid ellipsoid of x vectors satisfying

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi^2_d(\alpha)$$

has a probability $1-\alpha$.

How?

We know that

χ_d^2 is defined as the distribution of the sum $\sum_{i=1}^{d} Z_i^2$ where $Z_i's$ are independent $N(0,1)$ random variables.

Additionally, if we assume that Σ is positive definite and $\Sigma \in \mathbb{R}^{d\times d}$

$$\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

1. u_i are the orthonormal eigenvectors of Σ
2. λ_i are the corresponding real eigenvectors

How?

We know that

χ_d^2 is defined as the distribution of the sum $\sum_{i=1}^{d} Z_i^2$ where $Z_i's$ are independent $N(0,1)$ random variables.

Additionally, if we assume that Σ is positive definite and $\Sigma \in \mathbb{R}^{d\times d}$

$$\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

1. $u_i$are the orthonormal eigenvectors of Σ
2. λ_i are the corresponding real eigenvectors

Then

Something Notable

$$\Sigma^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

Now, if our data matrix element $X \sim N_d(\mu, \Sigma)$

We have

$$\Sigma^{-1} \boldsymbol{u}_i = \frac{1}{\lambda_i} \boldsymbol{u}_i$$

Then

Something Notable

$$\Sigma^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

Now, if our data matrix element $X \sim N_d(\boldsymbol{\mu}, \Sigma)$

We have

$$\Sigma^{-1} \boldsymbol{u}_i = \frac{1}{\lambda_i} \boldsymbol{u}_i$$

Therefore

We have that

$$\left(X-\boldsymbol{\mu}\right)^{T}\Sigma^{-1}\left(X-\boldsymbol{\mu}\right)=\sum_{i=1}^{d}\frac{1}{\lambda_i}\left(X-\boldsymbol{\mu}\right)^{T}\boldsymbol{u}_i\boldsymbol{u}_i^{T}\left(X-\boldsymbol{\mu}\right)$$

Then

$$\left(X-\boldsymbol{\mu}\right)^{T}\Sigma^{-1}\left(X-\boldsymbol{\mu}\right)=\sum_{i=1}^{d}\left[\frac{1}{\sqrt{\lambda_i}}\boldsymbol{u}_i^{T}\left(X-\boldsymbol{\mu}\right)\right]^{2}=\sum_{i=1}^{d}Z_i^{2}$$

Therefore

We have that

$$\left(X-\boldsymbol{\mu}\right)^T \Sigma^{-1} \left(X-\boldsymbol{\mu}\right) = \sum_{i=1}^{d} \frac{1}{\lambda_i} \left(X-\boldsymbol{\mu}\right)^T \boldsymbol{u}_i \boldsymbol{u}_i^T \left(X-\boldsymbol{\mu}\right)$$

Then

$$\left(X-\boldsymbol{\mu}\right)^T \Sigma^{-1} \left(X-\boldsymbol{\mu}\right) = \sum_{i=1}^{d} \left[\frac{1}{\sqrt{\lambda_i}} \boldsymbol{u}_i^T \left(X-\boldsymbol{\mu}\right)\right]^2 = \sum_{i=1}^{d} Z_i^2$$

Therefore

If we define

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{pmatrix}, A_{d\times d} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1^T \\ \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2^T \\ \vdots \\ \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d^T \end{pmatrix}$$

We know that $(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d(0, \Sigma)$

- Then, we have $\boldsymbol{Z} = A(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d\left(0, A\Sigma A^T\right)$

Therefore

If we define

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{pmatrix}, A_{d\times d} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1^T \\ \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2^T \\ \vdots \\ \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d^T \end{pmatrix}$$

We know that $(X - \boldsymbol{\mu}) \sim N_d(0, \Sigma)$

- Then, we have $\boldsymbol{Z} = A(X - \boldsymbol{\mu}) \sim N_d\left(0, A\Sigma A^T\right)$

Therefore

Something Notable

$$A\Sigma A^T = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1^T \\ \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2^T \\ \vdots \\ \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d^T \end{pmatrix} \left[\sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T\right] \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1 & \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2 & \cdots & \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d \end{pmatrix}$$

Therefore

$$A\Sigma A^T = \begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{u}_1^T \\ \sqrt{\lambda_2}\boldsymbol{u}_2^T \\ \vdots \\ \sqrt{\lambda_d}\boldsymbol{u}_d^T \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1 & \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2 & \cdots & \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d \end{pmatrix} = I$$

Therefore

Something Notable

$$A\Sigma A^T = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1^T \\ \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2^T \\ \vdots \\ \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d^T \end{pmatrix} \left[\sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T\right] \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1 & \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2 & \cdots & \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d \end{pmatrix}$$

Therefore

$$A\Sigma A^T = \begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{u}_1^T \\ \sqrt{\lambda_2}\boldsymbol{u}_2^T \\ \vdots \\ \sqrt{\lambda_d}\boldsymbol{u}_d^T \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\boldsymbol{u}_1 & \frac{1}{\sqrt{\lambda_2}}\boldsymbol{u}_2 & \cdots & \frac{1}{\sqrt{\lambda_d}}\boldsymbol{u}_d \end{pmatrix} = I$$

Therefore

We have that $Z_1, Z_2, \ldots, Z_d$ are independent standard normal variables

- $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ has a χ^2_d-distribution.

Finally, the $P\left((\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq c^2\right)$

- It is the probability assigned to the ellipsoid $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq c^2$ by the density $N_d(\boldsymbol{\mu}, \Sigma)$

Therefore

We have that $Z_1, Z_2, \ldots, Z_d$ are independent standard normal variables

- $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ has a χ^2_d-distribution.

Finally, the $P\left((\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq c^2\right)$

- It is the probability assigned to the ellipsoid $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq c^2$ by the density $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Therefore

We have $P\left((\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi_d^2(\alpha)\right) = 1-\alpha$

Basically $\chi_d^2(\alpha)$ is the the critical chi-square value that makes possible the probability $1-\alpha$

Basically

- We assume that if $1-\alpha = .95$ is the data with probability of not being an outlier!!!

Therefore

We have $P\left((\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi^2_d(\alpha)\right) = 1-\alpha$

Basically $\chi^2_d(\alpha)$ is the the critical chi-square value that makes possible the probability $1-\alpha$

Basically

- We assume that if $1-\alpha = .95$ is the data with probability of not being an outlier!!!

Algorithm

The Partial Code

```
def OutlierRemoval(self, Data):
        SampleMean  = Data.mean(1)
        SampleCov   = Data - SampleMean
        SampleCov   = np.cov(SampleCov.T)
        Mahalonobis = (Data - SampleMean)*
                                        np.inv(SampleCov)*
                                        ((Data - SampleMean).T)

        # Something else here
        # Here you can use chi2.isf(\alpha,dim)
```

Outline

Data Normalization

In the real world

- In many practical situations a designer is confronted with features whose values lie within different dynamic ranges.

For Example

- We can have two features with the following ranges

$$
\begin{aligned}
x_i &\in [0, 100, 000] \\
x_j &\in [0, 0.5]
\end{aligned}
$$

Thus

- Many classification machines will be swamped by the first feature!!!

Data Normalization

In the real world

- In many practical situations a designer is confronted with features whose values lie within different dynamic ranges.

For Example

- We can have two features with the following ranges

$$
\begin{aligned}
x_i &\in [0, 100, 000] \\
x_j &\in [0, 0.5]
\end{aligned}
$$

Thus

- Many classification machines will be swamped by the first feature!!!

Data Normalization

In the real world

- In many practical situations a designer is confronted with features whose values lie within different dynamic ranges.

For Example

- We can have two features with the following ranges

$$
\begin{aligned}
x_i &\in [0, 100, 000] \\
x_j &\in [0, 0.5]
\end{aligned}
$$

Thus

- Many classification machines will be swamped by the first feature!!!

Data Normalization

We have the following situation

- Features with large values may have a larger influence in the cost function than features with small values.

Thus!!!

- This does not necessarily reflect their respective significance in the design of the classifier.

Data Normalization

We have the following situation

- Features with large values may have a larger influence in the cost function than features with small values.

Thus!!!

- This does not necessarily reflect their respective significance in the design of the classifier.

Data Normalization

We have the following situation

- Features with large values may have a larger influence in the cost function than features with small values.

Thus!!!

- This does not necessarily reflect their respective significance in the design of the classifier.

Outline

Min-Max Method

Be Naive

- For each feature $i = 1, ..., d$ obtain the $\max_i$ and the $\min_i$ such that

$$\hat{x}_{ik} = \frac{x_{ik} - \min_i}{\max_i - \min_i} \tag{1}$$

Problem

- This simple normalization will send everything to a unitary sphere thus loosing data resolution!!!

Min-Max Method

Be Naive

- For each feature $i = 1, ..., d$ obtain the $\max_i$ and the $\min_i$ such that

$$\hat{x}_{ik} = \frac{x_{ik} - \min_i}{\max_i - \min_i} \quad (1)$$

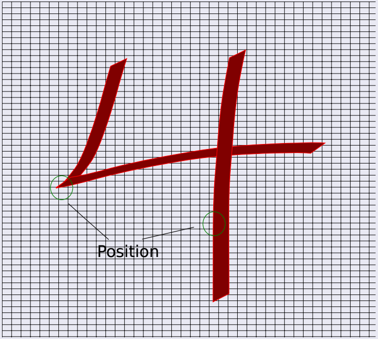
Problem

- This simple normalization will send everything to a unitary sphere thus loosing data resolution!!!

However

Even though this can happens there have been report that it can work...

- When data does not depend of single values as:



Gaussian Method

Use the idea of

Everything is Gaussian...

Gaussian Method

Use the idea of

Everything is Gaussian...

Thus

- For each feature set...
  1. $\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik},\ k = 1, 2, ..., d$
  2. $\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_{ik} - \overline{x}_k\right)^2,\ k = 1, 2, ..., d$

Gaussian Method

Use the idea of

Everything is Gaussian...

Thus

- For each feature set...
  1. $\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik},\ k = 1, 2, ..., d$
  2. $\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ik} - \overline{x}_k\right)^2,\ k = 1, 2, ..., d$

Thus

$$\hat{x}_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \tag{2}$$

Gaussian Method

Use the idea of

Everything is Gaussian...

Thus

- For each feature set...
  1. $\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik},\ \ k = 1, 2, ..., d$
  2. $\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ik} - \overline{x}_k\right)^2,\ \ k = 1, 2, ..., d$

Thus

$$\hat{x}_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \quad (2)$$

Gaussian Method

Use the idea of

Everything is Gaussian...

Thus

- For each feature set...
    1. $\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik},\ \ k = 1, 2, ..., d$
    2. $\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_{ik} - \overline{x}_k\right)^2,\ \ k = 1, 2, ..., d$

Thus

$$\hat{x}_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \tag{2}$$

Gaussian Mehtod

Thus

- All new features have zero mean and unit variance.

Further

- Other linear techniques limit the feature values in the range of $[0, 1]$ or $[-1, 1]$ by proper scaling.

However

- We can non-linear mapping. For example the softmax scaling.

Gaussian Mehtod

Thus

- All new features have zero mean and unit variance.

Further

- Other linear techniques limit the feature values in the range of $[0, 1]$ or $[-1, 1]$ by proper scaling.

However

- We can non-linear mapping. For example the softmax scaling.

Gaussian Mehtod

Thus

- All new features have zero mean and unit variance.

Further

- Other linear techniques limit the feature values in the range of $[0, 1]$ or $[-1, 1]$ by proper scaling.

However

- We can non-linear mapping. For example the softmax scaling.

Soft Max Scaling

Softmax Scaling

- It consists of two steps

First one

$$y_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \qquad (3)$$

Second one

$$\hat{x}_{ik} = \frac{1}{1 + \exp\left\{-y_{ik}\right\}} \qquad (4)$$

Soft Max Scaling

Softmax Scaling

- It consists of two steps

First one

$$y_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \tag{3}$$

Second one

$$\hat{x}_{ik} = \frac{1}{1 + \exp\{-y_{ik}\}} \tag{4}$$

Soft Max Scaling

Softmax Scaling

- It consists of two steps

First one

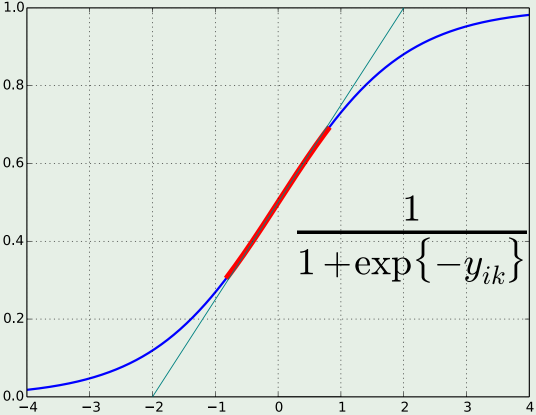
$$y_{ik} = \frac{x_{ik} - \overline{x}_k}{\sigma} \quad (3)$$

Second one

$$\hat{x}_{ik} = \frac{1}{1 + \exp\left\{-y_{ik}\right\}} \quad (4)$$

Explanation

Notice the red area is almost flat!!!



Actually

Thus, we have that

- The red region represents values of y inside of the region defined by the mean and variance (small values of y).
- Then, if we have those values x behaves as a linear function.

And values too away from the mean

- They are squashed by the exponential part of the function.

Actually

Thus, we have that

- The red region represents values of y inside of the region defined by the mean and variance (small values of y).
- Then, if we have those values x behaves as a linear function.

And values too away from the mean

- They are squashed by the exponential part of the function.

If you want a more complex analysis

A more complex analysis

- You can use a Taylor's expansion

$$x = f(y) = f(a) + f'(y)(y-a) + \frac{f''(y)(y-a)^2}{2} + \ldots \tag{5}$$

Outline

Missing Data

This can happen

In practice, certain features may be missing from some feature vectors.

Missing Data

This can happen

In practice, certain features may be missing from some feature vectors.

Examples where this happens

1. Social sciences - incomplete surveys.
2. Remote sensing - sensors go off-line.
3. etc.

Missing Data

This can happen

In practice, certain features may be missing from some feature vectors.

Examples where this happens

1. Social sciences - incomplete surveys.
2. Remote sensing - sensors go off-line.
3. etc.

Note

Completing the missing values in a set of data is also known as imputation.

Missing Data

This can happen

In practice, certain features may be missing from some feature vectors.

Examples where this happens

1. Social sciences - incomplete surveys.
2. Remote sensing - sensors go off-line.
3. etc.

Note

Completing the missing values in a set of data is also known as imputation.

Missing Data

This can happen

In practice, certain features may be missing from some feature vectors.

Examples where this happens

1. Social sciences - incomplete surveys.
2. Remote sensing - sensors go off-line.
3. etc.

Note

Completing the missing values in a set of data is also known as imputation.

Some traditional techniques to solve this problem

Use zeros and risked it!!!

The idea is not to add anything to the features

The sample mean/unconditional mean

Does not matter what distribution you have use the sample mean

$$\overline{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik} \quad (6)$$

Find the distribution of your data

Use the mean from that distribution. For example, if you have a beta distribution

$$\overline{x}_i = \frac{\alpha}{\alpha+\beta} \quad (7)$$

Some traditional techniques to solve this problem

Use zeros and risked it!!!

The idea is not to add anything to the features

The sample mean/unconditional mean

Does not matter what distribution you have use the sample mean

$$\overline{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik} \tag{6}$$

Find the distribution of your data

Use the mean from that distribution. For example, if you have a beta distribution

$$\overline{x}_i = \frac{\alpha}{\alpha+\beta} \tag{7}$$

Some traditional techniques to solve this problem

Use zeros and risked it!!!

The idea is not to add anything to the features

The sample mean/unconditional mean

Does not matter what distribution you have use the sample mean

$$\overline{x}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik} \tag{6}$$

Find the distribution of your data

Use the mean from that distribution. For example, if you have a beta distribution

$$\overline{x}_i = \frac{\alpha}{\alpha+\beta} \tag{7}$$

The MOST traditional

Drop it

- Remove that data
  - Still you need to have a lot of data to have this luxury

Outline

Something more advanced

Split data samples in two set of variables

$$\boldsymbol{x}_{complete} = \begin{pmatrix} \boldsymbol{x}_{observed} \\ \boldsymbol{x}_{missed} \end{pmatrix} \tag{8}$$

Generate the following probability distribution

$$P\left(\boldsymbol{x}_{missed} \middle| \boldsymbol{x}_{observed}, \Theta\right) = \frac{P\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} \middle| \Theta\right)}{P\left(\boldsymbol{x}_{observed} \middle| \Theta\right)} \tag{9}$$

where

$$p\left(\boldsymbol{x}_{observed} \middle| \Theta\right) = \int_{\mathcal{X}} p\left(\boldsymbol{x}_{complete} \middle| \Theta\right) d\boldsymbol{x}_{missed} \tag{10}$$

Something more advanced

Split data samples in two set of variables

$$\boldsymbol{x}_{complete} = \begin{pmatrix} \boldsymbol{x}_{observed} \\ \boldsymbol{x}_{missed} \end{pmatrix} \tag{8}$$

Generate the following probability distribution

$$P\left(\boldsymbol{x}_{missed} | \boldsymbol{x}_{observed}, \Theta\right) = \frac{P\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} | \Theta\right)}{P\left(\boldsymbol{x}_{observed} | \Theta\right)} \tag{9}$$

where

$$p\left(\boldsymbol{x}_{observed} | \Theta\right) = \int_{\mathcal{X}} p\left(\boldsymbol{x}_{complete} | \Theta\right) d\boldsymbol{x}_{missed} \tag{10}$$

Something more advanced

Split data samples in two set of variables

$$\boldsymbol{x}_{complete} = \begin{pmatrix} \boldsymbol{x}_{observed} \\ \boldsymbol{x}_{missed} \end{pmatrix} \tag{8}$$

Generate the following probability distribution

$$P\left(\boldsymbol{x}_{missed} | \boldsymbol{x}_{observed}, \Theta\right) = \frac{P\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} | \Theta\right)}{P\left(\boldsymbol{x}_{observed} | \Theta\right)} \tag{9}$$

where

$$p\left(\boldsymbol{x}_{observed} | \Theta\right) = \int_{\mathcal{X}} p\left(\boldsymbol{x}_{complete} | \Theta\right) d\boldsymbol{x}_{missed} \tag{10}$$

We can use EM

Basically, we use the data to obtain a multivariate version of the data

- Then, we use the α_i in a roulette based algorithm to select a sample
  - Then, we generate $x_{missed} \sim p_j\left(x|\theta\right) + Var\left(x\right)$

This is the most simple

- What about something more complex?

We can use EM

Basically, we use the data to obtain a multivariate version of the data

- Then, we use the α_i in a roulette based algorithm to select a sample
  - Then, we generate $x_{missed} \sim p_j(x|\theta) + Var(x)$

This is the most simple

- What about something more complex?

For this, we can do

We have the following joint probability

$$f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right)$$

Thus, the complete log likelihood

$$\ell\left(\theta\right) = \log f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right)$$

Therefore, we have

$$l_{\boldsymbol{x}_{\boldsymbol{missed}}}\left(\theta\right) = \log \int f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right) d\boldsymbol{x}_{\boldsymbol{missed}}$$

For this, we can do

We have the following joint probability

$$f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right)$$

Thus, the complete log likelihood

$$\ell\left(\theta\right) = \log f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right)$$

Therefore, we have

$$l_{\boldsymbol{x}_{missed}}\left(\theta\right) = \log \int f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right) d\boldsymbol{x}_{missed}$$

For this, we can do

We have the following joint probability

$$f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} | \theta\right)$$

Thus, the complete log likelihood

$$\ell\left(\theta\right) = \log f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} | \theta\right)$$

Therefore, we have

$$l_{\boldsymbol{x}_{missed}}\left(\theta\right) = \log \int f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed} | \theta\right) d\boldsymbol{x}_{missed}$$

Here, it is quite interesting

We have a ratio like this

$$\log \frac{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right)}{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}\theta_t\right)}$$

Basically we can get the Q function

$$\begin{aligned} Q\left(\theta|\theta_t\right) = & E_{\theta_t}\left[\log \frac{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right)}{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta_t\right)}\right] \\ = & \int \log \frac{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta\right)}{f\left(\boldsymbol{x}_{missed}, \boldsymbol{x}_{observed}|\theta_t\right)} f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta_t\right) d\boldsymbol{x}_{observed} \end{aligned}$$

Here, it is quite interesting

We have a ratio like this

$$\log \frac{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right)}{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} \theta_t\right)}$$

Basically we can get the Q function

$$\begin{aligned} Q\left(\theta|\theta_t\right) =& E_{\theta_t}\left[\log \frac{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right)}{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta_t\right)}\right] \\ =& \int \log \frac{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta\right)}{f\left(\boldsymbol{x}_{\boldsymbol{missed}}, \boldsymbol{x}_{\boldsymbol{observed}} | \theta_t\right)} f\left(\boldsymbol{x}_{\boldsymbol{observed}} | \boldsymbol{x}_{\boldsymbol{missed}}, \theta_t\right) d\boldsymbol{x}_{\boldsymbol{observed}} \end{aligned}$$

In this case

Why this ratio?

- Actually, because we want the missing data to be estimated by the observed one

Actually... There is something quite interesting

- Kullback–Leibler Divergence!!!

In this case

Why this ratio?

- Actually, because we want the missing data to be estimated by the observed one

Actually... There is something quite interesting

- Kullback–Leibler Divergence!!!

Actually the Kullback–Leibler Divergence

Definition

- For probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the Kullback–Leibler divergence is defined as

$$KL\left(P\|Q\right) = \int p\left(x\right)\log\left(\frac{p\left(x\right)}{q\left(x\right)}\right)dx$$

Thus, we have that

$$Q\left(\theta|\theta_t\right) = \int \log\frac{f\left(x_{missed}, x_{observed}|\theta\right)}{f\left(x_{missed}, x_{observed}|\theta_t\right)} f\left(x_{observed}|x_{missed}, \theta_t\right) dx_{observed}$$

$$= \int \log\frac{f\left(x_{observed}|x_{missed}, \theta\right) f\left(x_{missed}|\theta\right)}{f\left(x_{observed}|x_{missed}, \theta_t\right) f\left(x_{missed}|\theta_t\right)} f\left(x_{obser}|x_{missed}, \theta_t\right) dx_{obser}$$

Actually the Kullback–Leibler Divergence

Definition

- For probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the Kullback–Leibler divergence is defined as

$$KL\left(P\|Q\right) = \int p\left(x\right)\log\left(\frac{p\left(x\right)}{q\left(x\right)}\right)dx$$

Thus, we have that

$$\begin{aligned} Q\left(\theta|\theta_t\right) &= \int \log\frac{f\left(\boldsymbol{x}_{missed},\boldsymbol{x}_{observed}|\theta\right)}{f\left(\boldsymbol{x}_{missed},\boldsymbol{x}_{observed}|\theta_t\right)} f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)d\boldsymbol{x}_{observed} \\ &= \int \log\frac{f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta\right)f\left(\boldsymbol{x}_{missed}|\theta\right)}{f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)f\left(\boldsymbol{x}_{missed}|\theta_t\right)} f\left(\boldsymbol{x}_{obser}|\boldsymbol{x}_{missed},\theta_t\right)d\boldsymbol{x}_{obser} \end{aligned}$$

Basically, we have

The well known difference and KL Divergence

$$
\begin{aligned}
Q\left(\theta|\theta_t\right) = & \log f\left(\boldsymbol{x}_{missed}|\theta\right)\int f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)d\boldsymbol{x}_{observed} - \dots \\
& \log f\left(\boldsymbol{x}_{missed}|\theta_t\right)\int f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)d\boldsymbol{x}_{observed} + \dots \\
& \int_{\theta_t}\log\frac{f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta\right)}{f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)}f\left(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed},\theta_t\right)d\boldsymbol{x}_{observed}
\end{aligned}
$$

Using a little bit of notation

$$Q\left(\theta|\theta_t\right) = l_y\left(\theta\right) - l_y\left(\theta_t\right) - KL\left(f_{\theta_t}^{\boldsymbol{x}_{missed}} \,\|\, f_{\theta}^{\boldsymbol{x}_{missed}}\right)$$

Basically, we have

The well known difference and KL Divergence

$$
\begin{aligned}
Q(\theta|\theta_t) = & \log f(\boldsymbol{x}_{missed}|\theta) \int f(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta_t)\, d\boldsymbol{x}_{observed} - \dots \\
& \log f(\boldsymbol{x}_{missed}|\theta_t) \int f(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta_t)\, d\boldsymbol{x}_{observed} + \dots \\
& \int_{\theta_t} \log \frac{f(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta)}{f(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta_t)} f(\boldsymbol{x}_{observed}|\boldsymbol{x}_{missed}, \theta_t)\, d\boldsymbol{x}_{observed}
\end{aligned}
$$

Using a little bit of notation

$$
Q(\theta|\theta_t) = l_y(\theta) - l_y(\theta_t) - KL\left(f_{\theta_t}^{\boldsymbol{x}_{missed}} \,\|\, f_{\theta}^{\boldsymbol{x}_{missed}}\right)
$$

KL-divergence is minimized for $\theta = \theta_t$, actually zero!!!

Then when differentiating the Q divergence

$$\left.\frac{\partial Q\left(\theta|\theta_t\right)}{\partial \theta}\right|_{\theta=\theta_y} = \left.\frac{\partial l_{\boldsymbol{x_{missed}}}\left(\theta\right)}{\partial \theta}\right|_{\theta=\theta_y}$$

Thus define the iteration as

$$\theta_{t+1} = \arg\max_{\theta} Q\left(\theta|\theta_t\right)$$

KL-divergence is minimized for $\theta = \theta_t$, actually zero!!!

Then when differentiating the Q divergence

$$\left.\frac{\partial Q\left(\theta|\theta_t\right)}{\partial \theta}\right|_{\theta=\theta_y} = \left.\frac{\partial l_{\boldsymbol{x_{missed}}}\left(\theta\right)}{\partial \theta}\right|_{\theta=\theta_y}$$

Thus define the iteration as

$$\theta_{t+1} = \arg\max_{\theta} Q\left(\theta|\theta_t\right)$$

It is possible to see that

Something Notable

$$Q\left(\theta_{t+1}|\theta_t\right) + l_y\left(\theta_t\right) + KL\left(f_{\theta_t}^{\boldsymbol{x}_{missed}} \middle\| f_{\theta_t}^{\boldsymbol{x}_{missed}}\right) = l_y\left(\theta_{t+1}\right)$$

Then

$$l_y\left(\theta_{t+1}\right) \geq l_y\left(\theta_t\right) + 0 + 0$$

Thus

- The log-likelihood never decreases after a combined $E - step$ and $M - step$.

It is possible to see that

Something Notable

$$Q\left(\theta_{t+1}|\theta_t\right) + l_y\left(\theta_t\right) + KL\left(f_{\theta_t}^{\boldsymbol{x}_{missed}} \middle\| f_{\theta_t}^{\boldsymbol{x}_{missed}}\right) = l_y\left(\theta_{t+1}\right)$$

Then

$$l_y\left(\theta_{t+1}\right) \geq l_y\left(\theta_t\right) + 0 + 0$$

Thus

- The log-likelihood never decreases after a combined $E-step$ and $M-step$.

It is possible to see that

Something Notable

$$Q\left(\theta_{t+1}|\theta_t\right) + l_y\left(\theta_t\right) + KL\left(f_{\theta_t}^{\boldsymbol{x}_{missed}} \middle\| f_{\theta_t}^{\boldsymbol{x}_{missed}}\right) = l_y\left(\theta_{t+1}\right)$$

Then

$$l_y\left(\theta_{t+1}\right) \geq l_y\left(\theta_t\right) + 0 + 0$$

Thus

- The log-likelihood never decreases after a combined $E - step$ and $M - step$.

Here, everything looks great but...

We need to know to which distribution could come the result

- Thus, we have that we assume that the missing data can come from two distributions!!!

Start from the simple

- We assume a two possible sources of the information for the missing data.

Here, everything looks great but...

We need to know to which distribution could come the result

- Thus, we have that we assume that the missing data can come from two distributions!!!

Start from the simple

- We assume a two possible sources of the information for the missing data.

Thus, we can device the following Likelihood

We can consider a sample $Y = \{Y_1, ..., Y_n\}$ from individual densities

$$f\left(y|\alpha,\mu\right) = \alpha\phi\left(y-\mu\right) + \left(1-\alpha\right)\phi\left(y\right)$$

Where, we have

$$\phi\left(y\right) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{y^2}{2}\right\}$$

- With both α and μ are both unknown, but $0 < \alpha < 1$.

Thus, we can device the following Likelihood

We can consider a sample $Y = \{Y_1, ..., Y_n\}$ from individual densities

$$f\left(y|\alpha, \mu\right) = \alpha\phi\left(y-\mu\right) + \left(1-\alpha\right)\phi\left(y\right)$$

Where, we have

$$\phi\left(y\right) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{y^2}{2}\right\}$$

- With both α and μ are both unknown, but $0 < \alpha < 1$.

Incomplete observation

The likelihood function becomes

$$L_{\boldsymbol{x}_{\boldsymbol{missed}}}\left(\alpha,\mu\right)=\prod_{i=1}^{N}\alpha\phi\left(y_i-\mu\right)+\left(1-\alpha\right)\phi\left(y_i\right)$$

This is a quite unpleasant function

- But suppose we knew which observations came from which population?

Incomplete observation

The likelihood function becomes

$$L_{\boldsymbol{x_{missed}}}(\alpha,\mu)=\prod_{i=1}^{N}\alpha\phi\left(y_i-\mu\right)+(1-\alpha)\,\phi\left(y_i\right)$$

This is a quite unpleasant function

- But suppose we knew which observations came from which population?

What?

Let $X = \{X_1, ..., X_n\}$ be i.i.d. with $P(X_i = 1) = \alpha$

- Then, we play the hierarchical idea

Hierachy

$$Y_i \sim N(\mu, 1) \text{ if } X_i = 1$$
$$Y_i \sim N(0, 1) \text{ if } X_i = 0$$

i.e X_i allows to indicate to which distribution Y_i belongs

- Then we need the marginal distribution of Y.

What?

Let $X = \{X_1, ..., X_n\}$ be i.i.d. with $P(X_i = 1) = \alpha$

- Then, we play the hierarchical idea

Hierachy

$$Y_i \sim N(\mu, 1) \text{ if } X_i = 1$$
$$Y_i \sim N(0, 1) \text{ if } X_i = 0$$

i.e X_i allows to indicate to which distribution Y_i belongs

- Then we need the marginal distribution of Y.

What?

Let $X = \{X_1, ..., X_n\}$ be i.i.d. with $P(X_i = 1) = \alpha$

- Then, we play the hierarchical idea

Hierachy

$$
\begin{aligned}
Y_i \sim & N(\mu, 1) \text{ if } X_i = 1 \\
Y_i \sim & N(0, 1) \text{ if } X_i = 0
\end{aligned}
$$

i.e X_i allows to indicate to which distribution Y_i belongs

- Then we need the marginal distribution of Y.

Thus

The complete Data Likelihood is

$$L_{x,y}(\alpha,\mu) = \prod_{i=1}^{N} \alpha^{x_i}\phi(y_i-\mu)^{x_i}(1-\alpha)^{1-x_i}\phi(y_i)^{1-x_i}$$

Or given that $\phi(y_i)$ does not contain any parameter

$$L_{x,y}(\alpha,\mu) \propto \alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}\prod_{i=1}^{N}\phi(y_i-\mu)^{x_i}$$

Thus

The complete Data Likelihood is

$$L_{x,y}\left(\alpha,\mu\right)=\prod_{i=1}^{N}\alpha^{x_i}\phi\left(y_i-\mu\right)^{x_i}\left(1-\alpha\right)^{1-x_i}\phi\left(y_i\right)^{1-x_i}$$

Or given that $\phi\left(y_i\right)$ does not contain any parameter

$$L_{x,y}\left(\alpha,\mu\right)\propto\alpha^{\sum x_i}\left(1-\alpha\right)^{n-\sum x_i}\prod_{i=1}^{N}\phi\left(y_i-\mu\right)^{x_i}$$

Then taking logarithms

We have that

$$l_{x,y}(\alpha,\mu)=\sum x_i\log\alpha+\left(n-\sum x_i\right)\log(1-\alpha)-\sum\frac{x_i(y_i-\mu)^2}{2}$$

Therefore, if we differentiate

$$\widehat{\alpha}=\frac{1}{x_i}\sum x_i,\ \widehat{\mu}=\frac{\sum x_i y_i}{\sum x_i}$$

We have seen this formulations

- The EM algorithm for the Mixture of Gaussian's

Then taking logarithms

We have that

$$l_{x,y}(\alpha,\mu) = \sum x_i \log\alpha + \left(n - \sum x_i\right)\log(1-\alpha) - \sum \frac{x_i\,(y_i-\mu)^2}{2}$$

Therefore, if we differentiate

$$\widehat{\alpha} = \frac{1}{x_i}\sum x_i, \widehat{\mu} = \frac{\sum x_i y_i}{\sum x_i}$$

We have seen this formulations

- The EM algorithm for the Mixture of Gaussian's

Then taking logarithms

We have that

$$l_{x,y}\left(\alpha,\mu\right)=\sum x_i\log\alpha+\left(n-\sum x_i\right)\log\left(1-\alpha\right)-\sum\frac{x_i\left(y_i-\mu\right)^2}{2}$$

Therefore, if we differentiate

$$\widehat{\alpha}=\frac{1}{x_i}\sum x_i, \widehat{\mu}=\frac{\sum x_i y_i}{\sum x_i}$$

We have seen this formulations

- The EM algorithm for the Mixture of Gaussian's

Outline

Example

We have two matrices

- Data Matrix X
- Missing Data M

$$M_{ij} = \begin{cases} 0 & X_{ij} \text{ is missing} \\ 1 & X_{ij} \text{ is not missing} \end{cases}$$

Therefore, we have

- $X = (X_{obs}, X_{mis})$

This comes from

- "Bayes and multiple imputation" by RJA Little, DB Rubin (2002)

Example

We have two matrices

- Data Matrix X
- Missing Data M

$$M_{ij} = \begin{cases} 0 & X_{ij} \text{ is missing} \\ 1 & X_{ij} \text{ is not missing} \end{cases}$$

Therefore, we have

- $X = (X_{obs}, X_{mis})$

This comes from

- "Bayes and multiple imputation" by RJA Little, DB Rubin (2002)

Example

We have two matrices

- Data Matrix X
- Missing Data M

$$M_{ij} = \begin{cases} 0 & X_{ij} \text{ is missing} \\ 1 & X_{ij} \text{ is not missing} \end{cases}$$

Therefore, we have

- $X = (X_{obs}, X_{mis})$

This comes from

- "Bayes and multiple imputation" by RJA Little, DB Rubin (2002)

We can use the following optimization

We can do the following

$$\min_{M_{ij}=1} \|X - AB\|_F$$

Clearly an initial matrix decomposition, where

$$M_{ij}x_{ij} \approx \sum_{k=1}^{K} a_{ik}b_{kj}$$

So the total error to be minimized is

$$\min_{M_{ij}=1} \|X - AB\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[M_{ij}x_{ij} - \sum_{k=1}^{K} a_{ik}b_{kj}\right]^2}$$

- $K \ll N, M$

We can use the following optimization

We can do the following

$$\min_{M_{ij}=1} \|X - AB\|_F$$

Clearly an initial matrix decomposition, where

$$M_{ij}x_{ij} \approx \sum_{k=1}^{K} a_{ik}b_{kj}$$

So the total error to be minimized is

$$\min_{M_{ij}=1} \|X - AB\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[M_{ij}x_{ij} - \sum_{k=1}^{K} a_{ik}b_{kj}\right]^2}$$

- $K \ll N, M$

We can use the following optimization

We can do the following

$$\min_{M_{ij}=1} \|X - AB\|_F$$

Clearly an initial matrix decomposition, where

$$M_{ij}x_{ij} \approx \sum_{k=1}^{K} a_{ik}b_{kj}$$

So the total error to be minimized is

$$\min_{M_{ij}=1} \|X - AB\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[M_{ij}x_{ij} - \sum_{k=1}^{K} a_{ik}b_{kj}\right]^2}$$

- $K \ll N, M$

This can be regularized

Using the following ideas

$$\min_{M_{ij}=1} \|X - AB\|_F + \lambda \left[\|A\|^2 + \|B\|^2 \right]$$

Therefore, once the minimization is achieved

- We finish with two dense matrices A, B that can be used to obtain the elements with entries $M_{ij} = 0$

This can be regularized

Using the following ideas

$$\min_{M_{ij}=1} \|X - AB\|_F + \lambda \left[\|A\|^2 + \|B\|^2 \right]$$

Therefore, once the minimization is achieved

- We finish with two dense matrices A, B that can be used to obtain the elements with entries $M_{ij} = 0$

There are many other methods for this

For example

- Moritz Hardt. Understanding Alternating Minimization for Matrix Completion. FOCS, pages 651–660, 2014.
- Moritz Hardt, Mary Wootters. Fast matrix completion without the condition number. COLT, pages 638–678, 20
- Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh, Matrix completion from noisy entries, The Journal of Machine Learning Research 99 (2010), 2057–2078.
- Stephen J Wright, Robert D Nowak, and M´ario AT Figueiredo, Sparse reconstruction by separable approximation, Signal Processing, IEEE Transactions on 57 (2009), no. 7, 2479–2493.

Outline

THE PEAKING PHENOMENON

Remeber

Normally, to design a classifier with good generalization performance, we want the number of sample N to be larger than the number of features d.

What?

The intuition, the larger the number of samples vs the number of features, the smaller the error P_e

THE PEAKING PHENOMENON

Remeber

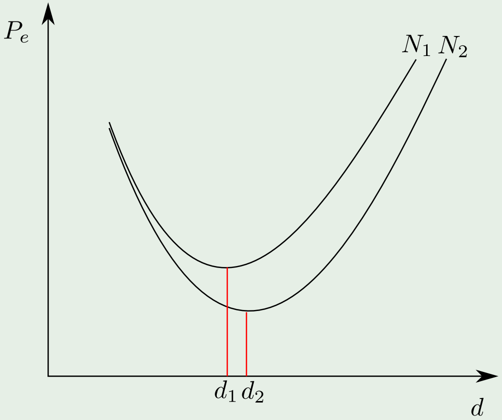
Normally, to design a classifier with good generalization performance, we want the number of sample N to be larger than the number of features d.

What?

The intuition, the larger the number of samples vs the number of features, the smaller the error P_e

Graphically

For $N_2 \gg N_1$



Let us explain

Something Notable

Let's look at the following example from the paper:

- "A Problem of Dimensionality: A Simple Example" by G.A. Trunk

THE PEAKING PHENOMENON

Assume the following problem

We have two classes ω_1, ω_2 such that

$$P(\omega_1) = P(\omega_2) = \frac{1}{2} \tag{11}$$

Both Classes have the following Gaussian distribution

1. $\omega_1 \Rightarrow \mu$ and $\Sigma = I$
2. $\omega_2 \Rightarrow -\mu$ and $\Sigma = I$

Where

$$\mu = \left[1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{d}}\right]$$

THE PEAKING PHENOMENON

Assume the following problem

We have two classes ω_1, ω_2 such that

$$P(\omega_1) = P(\omega_2) = \frac{1}{2} \tag{11}$$

Both Classes have the following Gaussian distribution

1. $\omega_1 \Rightarrow \mu$ and $\Sigma = I$
2. $\omega_2 \Rightarrow -\mu$ and $\Sigma = I$

Where

$$\mu = \left[1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, ..., \frac{1}{\sqrt{d}}\right]$$

THE PEAKING PHENOMENON

Assume the following problem

We have two classes ω_1, ω_2 such that

$$P(\omega_1) = P(\omega_2) = \frac{1}{2} \tag{11}$$

Both Classes have the following Gaussian distribution

1. $\omega_1 \Rightarrow \mu$ and $\Sigma = I$
2. $\omega_2 \Rightarrow -\mu$ and $\Sigma = I$

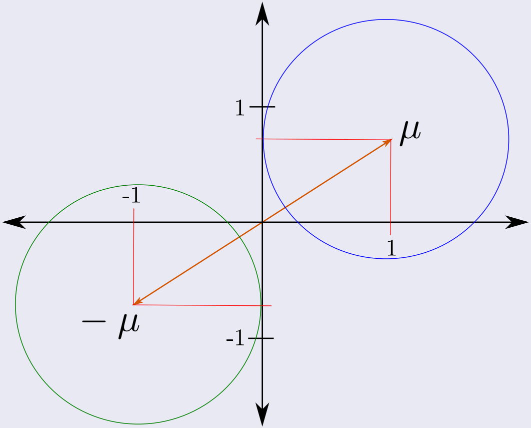
Where

$$\mu = \left[1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, ..., \frac{1}{\sqrt{d}}\right]$$

Example

The μ for $\mathbb{R}^2$

$$\mu = \left(1, \frac{1}{\sqrt{2}}\right)^T$$



THE PEAKING PHENOMENON

Properties of the features

Since the features are jointly Gaussian and $\Sigma = I$,the involved features are statistically independent.

THE PEAKING PHENOMENON

Properties of the features

Since the features are jointly Gaussian and $\Sigma = I$,the involved features are statistically independent.

We use the following rule to classify

if for any vector x, we have that

1. $\|x-\mu\|^2 < \|x+\mu\|^2$ or $z \equiv x^T\mu > 0$ then $x \in \omega_1$.
2. $z \equiv x^T\mu < 0$ then $x \in \omega_2$.

THE PEAKING PHENOMENON

Properties of the features

Since the features are jointly Gaussian and $\Sigma = I$,the involved features are statistically independent.

We use the following rule to classify

if for any vector $\boldsymbol{x}$, we have that

1. $\|\boldsymbol{x} - \boldsymbol{\mu}\|^2 < \|\boldsymbol{x} + \boldsymbol{\mu}\|^2$ or $z \equiv \boldsymbol{x}^T \boldsymbol{\mu} > 0$ then $\boldsymbol{x} \in \omega_1$.
2. $z \equiv \boldsymbol{x}^T \boldsymbol{\mu} < 0$ then $\boldsymbol{x} \in \omega_2$.

THE PEAKING PHENOMENON

Properties of the features

Since the features are jointly Gaussian and $\Sigma = I$,the involved features are statistically independent.

We use the following rule to classify

if for any vector $\boldsymbol{x}$, we have that

1. $\|\boldsymbol{x} - \boldsymbol{\mu}\|^2 < \|\boldsymbol{x} + \boldsymbol{\mu}\|^2$ or $z \equiv \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\mu} > 0$ then $\boldsymbol{x} \in \omega_1$.
2. $z \equiv \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\mu} < 0$ then $\boldsymbol{x} \in \omega_2$.

A little bit of algebra

For the first case

$$\left\|\boldsymbol{x}-\boldsymbol{\mu}\right\|^2 < \left\|\boldsymbol{x}+\boldsymbol{\mu}\right\|^2$$

$$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu}) < (\boldsymbol{x}+\boldsymbol{\mu})^T(\boldsymbol{x}+\boldsymbol{\mu})$$

$$\boldsymbol{x}^t\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu} < \boldsymbol{x}^t\boldsymbol{x} + 2\boldsymbol{x}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu}$$

$$0 < \boldsymbol{x}^T\boldsymbol{\mu} \equiv z$$

A little bit of algebra

For the first case

$$\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 < \|\boldsymbol{x}+\boldsymbol{\mu}\|^2$$

$$(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu}) < (\boldsymbol{x}+\boldsymbol{\mu})^T(\boldsymbol{x}+\boldsymbol{\mu})$$

$$x^t x - 2x^T\mu + \mu^T\mu < x^t x + 2x^T\mu + \mu^T\mu$$

$$0 < x^T\mu \equiv z$$

A little bit of algebra

For the first case

$$
\begin{aligned}
\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 &< \|\boldsymbol{x}+\boldsymbol{\mu}\|^2 \\
(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu}) &< (\boldsymbol{x}+\boldsymbol{\mu})^T(\boldsymbol{x}+\boldsymbol{\mu}) \\
\boldsymbol{x}^t\boldsymbol{x}-2\boldsymbol{x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu} &< \boldsymbol{x}^t\boldsymbol{x}+2\boldsymbol{x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu} \\
0 &< x^T\mu \equiv z
\end{aligned}
$$

We have then two cases

1. Known mean value μ.
2. Unknown mean value μ.

A little bit of algebra

For the first case

$$
\begin{aligned}
\left\|\boldsymbol{x}-\boldsymbol{\mu}\right\|^2 &< \left\|\boldsymbol{x}+\boldsymbol{\mu}\right\|^2\\
\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T\left(\boldsymbol{x}-\boldsymbol{\mu}\right) &< \left(\boldsymbol{x}+\boldsymbol{\mu}\right)^T\left(\boldsymbol{x}+\boldsymbol{\mu}\right)\\
\boldsymbol{x}^t\boldsymbol{x}-\boldsymbol{2x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu} &< \boldsymbol{x}^t\boldsymbol{x}+\boldsymbol{2x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu}\\
0 &< \boldsymbol{x}^T\boldsymbol{\mu}\equiv z
\end{aligned}
$$

We have then two cases

1. Known mean value μ.
2. Unknown mean value μ.

A little bit of algebra

For the first case

$$
\begin{aligned}
\left\|\boldsymbol{x}-\boldsymbol{\mu}\right\|^{2} &< \left\|\boldsymbol{x}+\boldsymbol{\mu}\right\|^{2}\\
\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{T}\left(\boldsymbol{x}-\boldsymbol{\mu}\right) &< \left(\boldsymbol{x}+\boldsymbol{\mu}\right)^{T}\left(\boldsymbol{x}+\boldsymbol{\mu}\right)\\
\boldsymbol{x}^{t}\boldsymbol{x}-\boldsymbol{2x}^{\boldsymbol{T}}\boldsymbol{\mu}+\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\mu} &< \boldsymbol{x}^{t}\boldsymbol{x}+\boldsymbol{2x}^{\boldsymbol{T}}\boldsymbol{\mu}+\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{\mu}\\
0 &< \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\mu}\equiv z
\end{aligned}
$$

We have then two cases

1. Known mean value μ.
2. Unknown mean value μ.

A little bit of algebra

For the first case

$$
\begin{aligned}
\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 &< \|\boldsymbol{x}+\boldsymbol{\mu}\|^2 \\
(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu}) &< (\boldsymbol{x}+\boldsymbol{\mu})^T(\boldsymbol{x}+\boldsymbol{\mu}) \\
\boldsymbol{x}^t\boldsymbol{x}-\boldsymbol{2}\boldsymbol{x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu} &< \boldsymbol{x}^t\boldsymbol{x}+\boldsymbol{2}\boldsymbol{x}^T\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\mu} \\
0 &< \boldsymbol{x}^T\boldsymbol{\mu} \equiv z
\end{aligned}
$$

We have then two cases

1. Known mean value μ.
2. Unknown mean value μ.

Known mean value μ

Given that z is a linear combination of independent Gaussian Variables

1. It is a Gaussian variable.
2. $E[z] = \sum_{i=1}^{d} \mu_i E(x_i) = \sum_{i=1}^{d} \frac{1}{\sqrt{i}} \frac{1}{\sqrt{i}} = \sum_{i=1}^{d} \frac{1}{i} = \|\mu\|^2$.
3. $\sigma_z^2 = \|\mu\|^2$.

Known mean value μ

Given that z is a linear combination of independent Gaussian Variables

1. It is a Gaussian variable.
2. $E[z] = \sum_{i=1}^{d} \mu_i E(x_i) = \sum_{i=1}^{d} \frac{1}{\sqrt{i}} \frac{1}{\sqrt{i}} = \sum_{i=1}^{d} \frac{1}{i} = \|\boldsymbol{\mu}\|^2$.
3. $\sigma_z^2 = \|\boldsymbol{\mu}\|^2$.

Known mean value μ

Given that z is a linear combination of independent Gaussian Variables

1. It is a Gaussian variable.
2. $E[z] = \sum_{i=1}^{d} \mu_i E(x_i) = \sum_{i=1}^{d} \frac{1}{\sqrt{i}} \frac{1}{\sqrt{i}} = \sum_{i=1}^{d} \frac{1}{i} = \|\boldsymbol{\mu}\|^2$.
3. $\sigma_z^2 = \|\boldsymbol{\mu}\|^2$.

Known mean value μ

Given that z is a linear combination of independent Gaussian Variables

1. It is a Gaussian variable.
2. $E[z] = \sum_{i=1}^{d} \mu_i E(x_i) = \sum_{i=1}^{d} \frac{1}{\sqrt{i}} \frac{1}{\sqrt{i}} = \sum_{i=1}^{d} \frac{1}{i} = \|\boldsymbol{\mu}\|^2$.
3. $\sigma_z^2 = \|\boldsymbol{\mu}\|^2$.

Why the first statement?

Given that each feature of x

It can be seen as random variable with mean $\frac{1}{\sqrt{i}}$ and variance 1 with no correlation between each other.

What about the variance of z?

Why the first statement?

Given that each feature of $\boldsymbol{x}$

It can be seen as random variable with mean $\frac{1}{\sqrt{i}}$ and variance 1 with no correlation between each other.

What about the variance of $\boldsymbol{z}$?

$$
\begin{aligned}
Var\left(\boldsymbol{z}\right) =& E\left[\left(z-\|\boldsymbol{\mu}\|^{2}\right)^{2}\right] \\
=& E\left[z^{2}-2z\|\boldsymbol{\mu}\|^{2}+\|\boldsymbol{\mu}\|^{4}\right] \\
=& E\left[z^{2}\right]-\|\boldsymbol{\mu}\|^{4} \\
=& E\left[\left(\sum_{i=1}^{d}\mu_{i}x_{i}\right)\left(\sum_{i=1}^{d}\mu_{i}x_{i}\right)\right]-\left(\sum_{i=1}^{d}\frac{1}{i^{2}}+\sum_{j=1}^{d}\sum_{\substack{h=1\\ j\neq h}}^{d}\frac{1}{i}\times\frac{1}{j}\right)
\end{aligned}
$$

Why the first statement?

Given that each feature of x

It can be seen as random variable with mean $\frac{1}{\sqrt{i}}$ and variance 1 with no correlation between each other.

What about the variance of z?

$$
\begin{aligned}
Var\left(\boldsymbol{z}\right) =& E\left[\left(z-\left\|\boldsymbol{\mu}\right\|^{2}\right)^{2}\right]\\
=& E\left[z^{2}-2z\left\|\boldsymbol{\mu}\right\|^{2}+\left\|\boldsymbol{\mu}\right\|^{4}\right]\\
=& E\left[z^{2}\right]-\left\|\boldsymbol{\mu}\right\|^{4}\\
=& E\left[\left(\sum_{i=1}^{d}\mu_{i}x_{i}\right)\left(\sum_{i=1}^{d}\mu_{i}x_{i}\right)\right]-\left(\sum_{i=1}^{d}\frac{1}{i^{2}}+\sum_{j=1}^{d}\sum_{\substack{h=1\\ j\neq h}}^{d}\frac{1}{i}\times\frac{1}{j}\right)
\end{aligned}
$$

Why the first statement?

Given that each feature of x

It can be seen as random variable with mean $\frac{1}{\sqrt{i}}$ and variance 1 with no correlation between each other.

What about the variance of z?

$$
\begin{aligned}
Var(\boldsymbol{z}) =& E\left[\left(z-\|\boldsymbol{\mu}\|^{2}\right)^{2}\right] \\
=& E\left[z^{2}-2z\|\boldsymbol{\mu}\|^{2}+\|\boldsymbol{\mu}\|^{4}\right] \\
=& E\left[\boldsymbol{z}^{2}\right]-\|\boldsymbol{\mu}\|^{4}
\end{aligned}
$$

$$
= E\left[\left(\sum_{i=1}^{d}\mu_i x_i\right)\left(\sum_{i=1}^{d}\mu_i x_i\right)\right]-\left(\sum_{i=1}^{d}\frac{1}{i^2}+\sum_{j=1}^{d}\sum_{\substack{h=1\\ j\neq h}}^{d}\frac{1}{i}\times\frac{1}{j}\right)
$$

Why the first statement?

Given that each feature of $\boldsymbol{x}$

It can be seen as random variable with mean $\frac{1}{\sqrt{i}}$ and variance 1 with no correlation between each other.

What about the variance of $\boldsymbol{z}$?

$$
\begin{aligned}
Var(\boldsymbol{z}) &= E\left[\left(z - \|\boldsymbol{\mu}\|^2\right)^2\right] \\
&= E\left[z^2 - 2z\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|^4\right] \\
&= E\left[\boldsymbol{z}^2\right] - \|\boldsymbol{\mu}\|^4 \\
&= E\left[\left(\sum_{i=1}^{d}\mu_i x_i\right)\left(\sum_{i=1}^{d}\mu_i x_i\right)\right] - \left(\sum_{i=1}^{d}\frac{1}{i^2} + \sum_{\substack{j=1\\ j\neq h}}^{d}\sum_{h=1}^{d}\frac{1}{i}\times\frac{1}{j}\right)
\end{aligned}
$$

Thus

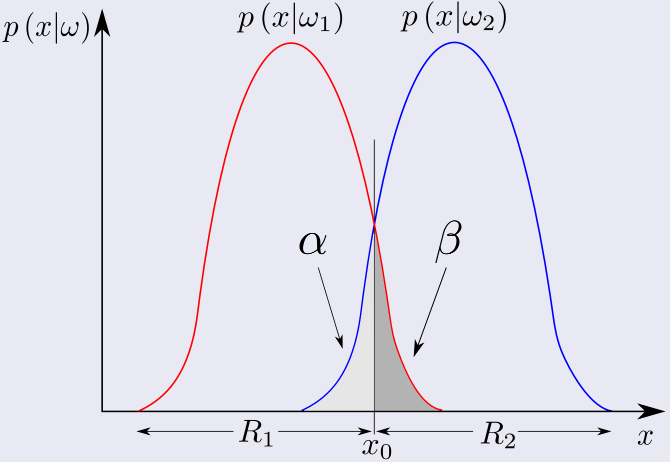
But, given that $x_i^2 \sim \chi_1^2\left(\frac{1}{i}\right)$, with mean

$$E\left[x_i^2\right] = 1 + \frac{1}{i} \quad (12)$$

Remark: The rest is for you to solve so $\sigma_z^2 = \|\boldsymbol{\mu}\|^2$.

Remember the P_e

We have then...



We get the probability of error

We know that the error is coming from the following equation

$$P_e = \frac{1}{2}\int\limits_{-\infty}^{x_0} p\left(z|\omega_2\right)d\boldsymbol{x} + \frac{1}{2}\int\limits_{x_0}^{\infty} p\left(z|\omega_1\right)d\boldsymbol{x} \tag{13}$$

But, we have equiprobable classes

We get the probability of error

We know that the error is coming from the following equation

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p\left(z|\omega_2\right) d\boldsymbol{x} + \frac{1}{2}\int_{x_0}^{\infty} p\left(z|\omega_1\right) d\boldsymbol{x} \tag{13}$$

But, we have equiprobable classes

$$\begin{aligned} P_e =& \frac{1}{2}\int_{-\infty}^{x_0} p\left(z|\omega_2\right) d\boldsymbol{x} + \frac{1}{2}\int_{x_0}^{\infty} p\left(z|\omega_1\right) \\ =& \int_{x_0}^{\infty} p\left(z|\omega_1\right) dx \end{aligned}$$

We get the probability of error

We know that the error is coming from the following equation

$$P_e = \frac{1}{2} \int_{-\infty}^{x_0} p\left(z|\omega_2\right) d\boldsymbol{x} + \frac{1}{2} \int_{x_0}^{\infty} p\left(z|\omega_1\right) d\boldsymbol{x} \tag{13}$$

But, we have equiprobable classes

$$\begin{aligned} P_e =& \frac{1}{2} \int_{-\infty}^{x_0} p\left(z|\omega_2\right) d\boldsymbol{x} + \frac{1}{2} \int_{x_0}^{\infty} p\left(z|\omega_1\right) \\ =& \int_{x_0}^{\infty} p\left(z|\omega_1\right) d\boldsymbol{x} \end{aligned}$$

Thus, we have that

Now, given that z is a sum of Gaussian

$$\text{exp term} = -\frac{1}{2\left\|\boldsymbol{\mu}\right\|^2}\left[\left(z - \left\|\boldsymbol{\mu}\right\|^2\right)^2\right] \tag{14}$$

Because we have the rule

We can do a change of variable to a normalized z

$$P_e = \int_{b_d}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \tag{15}$$

Thus, we have that

Now, given that z is a sum of Gaussian

$$\text{exp term} = -\frac{1}{2\left\|\boldsymbol{\mu}\right\|^2}\left[\left(z - \left\|\boldsymbol{\mu}\right\|^2\right)^2\right] \tag{14}$$

Because we have the rule

We can do a change of variable to a normalized z

$$P_e = \int_{b_d}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \tag{15}$$

Known mean value μ

The probability of error is given by

$$P_e = \int_{b_d}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \quad (16)$$

Known mean value μ

The probability of error is given by

$$P_e = \int_{b_d}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \tag{16}$$

Where

$$b_d = \sqrt{\sum_{i=1}^{d} \frac{1}{i}} \tag{17}$$

How?

Known mean value μ

Thus

When the series b_d tends to infinity as $d \to \infty$, the probability of error tends to **zero** as the number of features increases.

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} s_k \boldsymbol{x}_k \tag{18}$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} s_k \boldsymbol{x}_k \tag{18}$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Now, we have aproblem z is no more a Gaussian variable

Still, if we select d large enough and knowing that $z = \sum x_i \widehat{\mu}_i$, then for the central limit theorem, we can consider z to be Gaussian.

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} s_k \boldsymbol{x}_k \quad (18)$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Now, we have aproblem z is no more a Gaussian variable

Still, if we select d large enough and knowing that $z = \sum x_i \widehat{\mu}_i$, then for the central limit theorem, we can consider z to be Gaussian.

With mean and variance

1. $E[z] = \sum_{i=1}^{d} \frac{1}{i}$.
2. $\sigma_z^2 = \left(1 + \frac{1}{N}\right) \sum_{i=1}^{d} \frac{1}{i} + \frac{d}{N}$.

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} s_k \boldsymbol{x}_k \tag{18}$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Now, we have aproblem z is no more a Gaussian variable

Still, if we select d large enough and knowing that $z = \sum x_i \widehat{\mu}_i$, then for the central limit theorem, we can consider z to be Gaussian.

With mean and variance

1. $E[z] = \sum_{i=1}^{d} \frac{1}{i}$.
2. $\sigma_z^2 = \left(1 + \frac{1}{N}\right)\sum_{i=1}^{d} \frac{1}{i} + \frac{d}{N}$.

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^{N} s_k \boldsymbol{x}_k \tag{18}$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Now, we have aproblem z is no more a Gaussian variable

Still, if we select d large enough and knowing that $z = \sum x_i \widehat{\mu}_i$, then for the central limit theorem, we can consider z to be Gaussian.

With mean and variance

1. $E[z] = \sum_{i=1}^{d} \frac{1}{i}$.
2. $\sigma_z^2 = \left(1 + \frac{1}{N}\right) \sum_{i=1}^{d} \frac{1}{i} + \frac{d}{N}$.

Unknown mean value μ

For This, we use the maximum likelihood

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} s_k \boldsymbol{x}_k \tag{18}$$

where

1. $s_k = 1$ if $\boldsymbol{x}_k \in \omega_1$
2. $s_k = -1$ if $\boldsymbol{x}_k \in \omega_2$

Now, we have aproblem z is no more a Gaussian variable

Still, if we select d large enough and knowing that $z = \sum x_i \widehat{\mu}_i$, then for the central limit theorem, we can consider z to be Gaussian.

With mean and variance

1. $E[z] = \sum_{i=1}^{d} \frac{1}{i}$.
2. $\sigma_z^2 = \left(1 + \frac{1}{N}\right)\sum_{i=1}^{d} \frac{1}{i} + \frac{d}{N}$.

Unknown mean value μ

Thus

$$b_d = \frac{E\left[z\right]}{\sigma_z} \tag{19}$$

Thus, using P_e

- It can now be shown that $b_d \to 0$ as $d \to \infty$ and the probability of error tends to $\frac{1}{2}$ for any finite number N.

Unknown mean value μ

Thus

$$b_d = \frac{E[z]}{\sigma_z} \tag{19}$$

Thus, using P_e

- It can now be shown that $b_d \to 0$ as $d \to \infty$ and the probability of error tends to $\frac{1}{2}$ for any finite number N.

Finally

Case I

- If for any d the corresponding PDF is known, then we can perfectly discriminate the two classes by arbitrarily increasing the number of features.

Case II

- If the PDF's are not known, then the arbitrary increase of the number of features leads to the maximum possible value of the error rate, that is, $\frac{1}{2}$.

Thus

- Under a limited number of training data we must try to keep the number of features to a relatively low number.

Finally

Case I

- If for any d the corresponding PDF is known, then we can perfectly discriminate the two classes by arbitrarily increasing the number of features.

Case II

- If the PDF's are not known, then the arbitrary increase of the number of features leads to the maximum possible value of the error rate, that is, $\frac{1}{2}$.

Thus

- Under a limited number of training data we must try to keep the number of features to a relatively low number.

Finally

Case I

- If for any d the corresponding PDF is known, then we can perfectly discriminate the two classes by arbitrarily increasing the number of features.

Case II

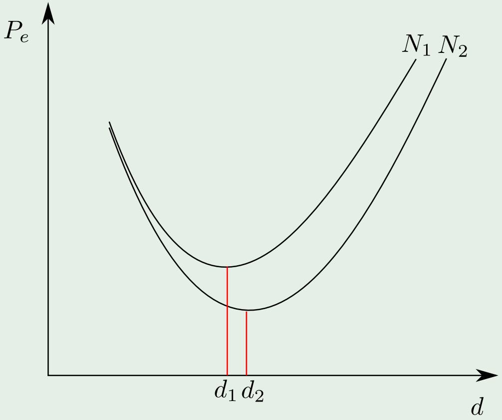
- If the PDF's are not known, then the arbitrary increase of the number of features leads to the maximum possible value of the error rate, that is, $\frac{1}{2}$.

Thus

- Under a limited number of training data we must try to keep the number of features to a relatively low number.

Graphically

For $N_2 \gg N_1$, minimum at $d = \frac{N}{\alpha}$ with $\alpha \in [2, 10]$



Back to Feature Selection

The Goal

1. Select the "optimum" number d of features.
2. Select the "best" d features.

Back to Feature Selection

The Goal

1. Select the “optimum” number d of features.
2. Select the “best” d features.

Why? Large d has a three-fold disadvantage:

- High computational demands.
- Low generalization performance.
- Poor error estimates

Back to Feature Selection

The Goal

1. Select the “optimum” number d of features.
2. Select the “best” d features.

Why? Large d has a three-fold disadvantage:

- High computational demands.
- Low generalization performance.
- Poor error estimates

Back to Feature Selection

The Goal

1. Select the "optimum" number d of features.
2. Select the "best" d features.

Why? Large d has a three-fold disadvantage:

- High computational demands.
- Low generalization performance.
- Poor error estimates

Back to Feature Selection

The Goal

1. Select the "optimum" number d of features.
2. Select the "best" d features.

Why? Large d has a three-fold disadvantage:

- High computational demands.
- Low generalization performance.
- Poor error estimates

Outline

Back to Feature Selection

Given N

d must be large enough to learn what makes classes different and what makes patterns in the same class similar

In addition

d must be small enough not to learn what makes patterns of the same class different

In practice

In practice, $d < N/3$ has been reported to be a sensible choice for a number of cases

Back to Feature Selection

Given N

d must be large enough to learn what makes classes different and what makes patterns in the same class similar

In addition

d must be small enough not to learn what makes patterns of the same class different

In practice

In practice, $d < N/3$ has been reported to be a sensible choice for a number of cases

Back to Feature Selection

Given N

d must be large enough to learn what makes classes different and what makes patterns in the same class similar

In addition

d must be small enough not to learn what makes patterns of the same class different

In practice

In practice, $d < {}^{N}\!/_{3}$ has been reported to be a sensible choice for a number of cases

Thus

Oh!!!

Once d has been decided, choose the d most informative features:

Best: Large between class distance, Small within class variance.

Thus

Oh!!!

Once d has been decided, choose the d most informative features:

Best: Large between class distance, Small within class variance.

The basic philosophy

1. Discard individual features with poor information content.
2. The remaining information rich features are examined jointly as vectors

Thus

Oh!!!

Once d has been decided, choose the d most informative features:

Best: Large between class distance, Small within class variance.

The basic philosophy

1. Discard individual features with poor information content.
2. The remaining information rich features are examined jointly as vectors

Thus

Oh!!!

Once d has been decided, choose the d most informative features:

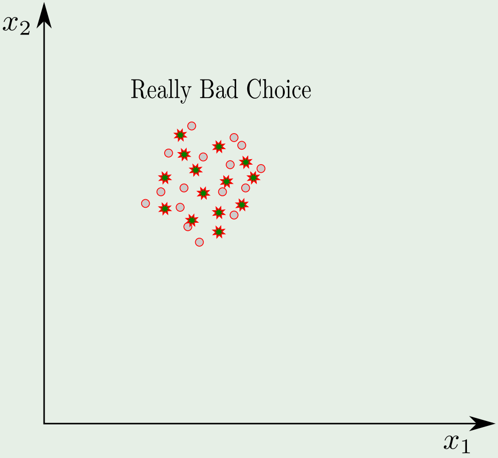
Best: Large between class distance, Small within class variance.

The basic philosophy

1. Discard individual features with poor information content.
2. The remaining information rich features are examined jointly as vectors

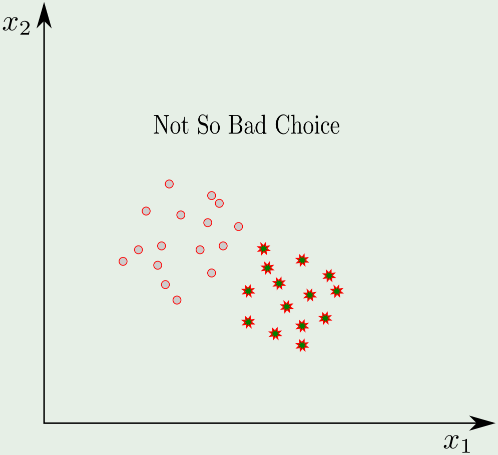
Example

Thus, we want to avoid choices



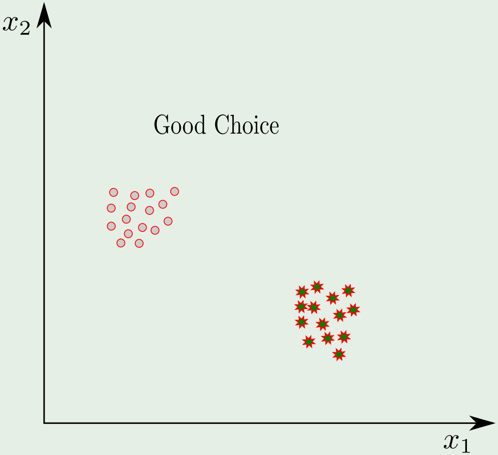
Example

Better Choice



Example

What We Want to Have



Outline

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

For this, we can use the following hypothesis testing

Assume the samples for two classes ω_1, ω_2 are vectors of random variables.

1. H_1: The values of the feature differ significantly
2. H_0: The values of the feature do not differ significantly

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

For this, we can use the following hypothesis testing

Assume the samples for two classes ω_1, ω_2 are vectors of random variables.

1. H_1: The values of the feature differ significantly
2. H_0: The values of the feature do not differ significantly

Meaning

H_0 is known as the null hypothesis and H_1 as the alternative hypothesis.

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

For this, we can use the following hypothesis testing

Assume the samples for two classes ω_1, ω_2 are vectors of random variables.

1. H_1: The values of the feature differ significantly
2. H_0: The values of the feature do not differ significantly

Meaning

H_0 is known as the null hypothesis and H_1 as the alternative hypothesis.

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

For this, we can use the following hypothesis testing

Assume the samples for two classes ω_1, ω_2 are vectors of random variables.

1. H_1: The values of the feature differ significantly
2. H_0: The values of the feature do not differ significantly

Meaning

H_0 is known as the null hypothesis and H_1 as the alternative hypothesis.

Using Statistics

Simplicity First Principles - Marcus Aurelius

- A first step in feature selection is to look at each of the generated features independently.
- Then, test their discriminatory capability for the problem at hand.

For this, we can use the following hypothesis testing

Assume the samples for two classes ω_1, ω_2 are vectors of random variables.

1. H_1: The values of the feature differ significantly
2. H_0: The values of the feature do not differ significantly

Meaning

H_0 is known as the null hypothesis and H_1 as the alternative hypothesis.

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$
\begin{aligned}
H_1 &: \theta \neq \theta_0 \\
H_0 &: \theta = \theta_0
\end{aligned}
$$

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_0 &: \theta = \theta_0 \end{aligned}$$

We want to generate a q

That measures the quality of our answer under our knowledge of the sample features $x_1, x_2, ..., x_N$.

1. Where $\overline{\Omega}$, the complement or critical region, is the region where we reject H_0.

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_0 &: \theta = \theta_0 \end{aligned}$$

We want to generate a q

That measures the quality of our answer under our knowledge of the sample features $x_1, x_2, ..., x_N$.

We ask for

1. Where a D (Acceptance Interval) is an interval where q lies with high probability under hypothesis H_0.
2. Where $\overline{D}$, the complement or critical region, is the region where we reject H_0.

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_0 &: \theta = \theta_0 \end{aligned}$$

We want to generate a q

That measures the quality of our answer under our knowledge of the sample features $x_1, x_2, ..., x_N$.

We ask for

1. Where a D (Acceptance Interval) is an interval where q lies with high probability under hypothesis H_0.
2. Where $\overline{D}$, the complement or critical region, is the region where we reject H_0.

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_0 &: \theta = \theta_0 \end{aligned}$$

We want to generate a q

That measures the quality of our answer under our knowledge of the sample features $x_1, x_2, ..., x_N$.

We ask for

1. Where a D (Acceptance Interval) is an interval where q lies with high probability under hypothesis H_0.
2. Where $\overline{D}$, the complement or critical region, is the region where we reject H_0.

Hypothesis Testing Basics

We need to represent these ideas in a more mathematical way

For this, given an unknown parameter θ:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_0 &: \theta = \theta_0 \end{aligned}$$

We want to generate a q

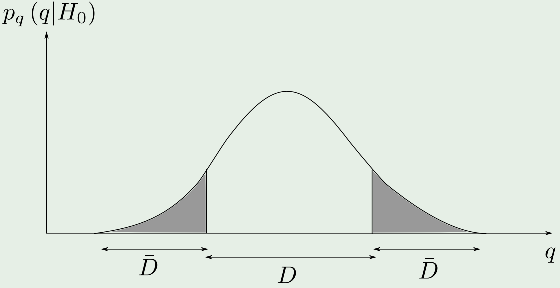
That measures the quality of our answer under our knowledge of the sample features $x_1, x_2, ..., x_N$.

We ask for

1. Where a D (Acceptance Interval) is an interval where q lies with high probability under hypothesis H_0.
2. Where $\overline{D}$, the complement or critical region, is the region where we reject H_0.

Example

Acceptance and critical regions for hypothesis testing. The area of the shaded region is the probability of an erroneous decision.



Known Variance Case

Assume

Be x a random variable and x_i the resulting experimental samples.

Let

1. $E[x] = \mu$
2. $E\left[(x-\mu)^2\right] = \sigma^2$

We can estimate μ using

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{20}$$

Known Variance Case

Assume

Be x a random variable and x_i the resulting experimental samples.

Let

1. $E\left[x\right] = \mu$
2. $E\left[\left(x-\mu\right)^2\right] = \sigma^2$

We can estimate μ using

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{20}$$

Known Variance Case

Assume

Be x a random variable and x_i the resulting experimental samples.

Let

1. $E[x] = \mu$
2. $E\left[(x-\mu)^2\right] = \sigma^2$

We can estimate μ using

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{20}$$

Known Variance Case

It can be proved that the

$\overline{x}$ is an unbiased estimate of the mean of x.

In a similar way

The variance of $\sigma^2_{\overline{x}}$ of $\overline{x}$ is

$$E\left[(\overline{x}-\mu)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)\right)^2\right] \quad (21)$$

Which is the following

$$E\left[(\overline{x}-\mu)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} E\left[(x_i-\mu)^2\right] + \frac{1}{N^2}\sum_{i}\sum_{j\neq i} E\left[(x_i-\mu)(x_j-\mu)\right] \quad (22)$$

Known Variance Case

It can be proved that the

$\overline{x}$ is an unbiased estimate of the mean of x.

In a similar way

The variance of $\sigma^2_{\overline{x}}$ of $\overline{x}$ is

$$E\left[(\overline{x}-\mu)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)\right)^2\right] \quad (21)$$

Which is the following

$$E\left[(\overline{x}-\mu)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} E\left[(x_i-\mu)^2\right] + \frac{1}{N^2}\sum_{i}\sum_{j\neq i} E\left[(x_i-\mu)(x_j-\mu)\right] \quad (22)$$

Known Variance Case

It can be proved that the

$\overline{x}$ is an unbiased estimate of the mean of x.

In a similar way

The variance of $\sigma^2_{\overline{x}}$ of $\overline{x}$ is

$$E\left[(\overline{x}-\mu)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)\right)^2\right] \quad (21)$$

Which is the following

$$E\left[(\overline{x}-\mu)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} E\left[(x_i-\mu)^2\right] + \frac{1}{N^2}\sum_{i}\sum_{j\neq i} E\left[(x_i-\mu)(x_j-\mu)\right] \quad (22)$$

Known Variance Case

Because independence

$$E\left[(x_i-\mu)((x_j-\mu)\right]=E\left[x_i-\mu\right]E\left[x_j-\mu\right]=0 \tag{23}$$

Thus

$$\sigma_{\overline{x}}^2=\frac{1}{N}\sigma^2 \tag{24}$$

Note: the larger the number of measurement samples, the smaller the variance of $\overline{x}$ around the true mean.

Known Variance Case

Because independence

$$E\left[(x_i-\mu)((x_j-\mu)\right]=E\left[x_i-\mu\right]E\left[x_j-\mu\right]=0 \tag{23}$$

Thus

$$\sigma_{\overline{x}}^2=\frac{1}{N}\sigma^2 \tag{24}$$

Note: the larger the number of measurement samples, the smaller the variance of $\overline{x}^{-}$ around the true mean.

What to do with it

Now, you are given a $\widehat{\mu}$ the estimated parameter (In our case the mean sample)

Thus:

$$\begin{aligned} H_1 &: E[x] \neq \widehat{\mu} \\ H_0 &: E[x] = \widehat{\mu} \end{aligned}$$

We define q

$$q = \frac{\overline{x} - \widehat{\mu}}{\frac{\sigma}{N}} \tag{25}$$

Recalling the central limit theorem

The probability density function of $\overline{x}$ under H_0 is approx Gaussian $N\left(\widehat{\mu}, \frac{\sigma}{N}\right)$

What to do with it

Now, you are given a $\widehat{\mu}$ the estimated parameter (In our case the mean sample)

Thus:

$$\begin{aligned} H_1 &: E[x] \neq \widehat{\mu} \\ H_0 &: E[x] = \widehat{\mu} \end{aligned}$$

We define q

$$q = \frac{\overline{x} - \widehat{\mu}}{\frac{\sigma}{N}} \tag{25}$$

Recalling the central limit theorem

The probability density function of $\overline{x}$ under H_0 is approx Gaussian $N\left(\widehat{\mu}, \frac{\sigma}{N}\right)$

What to do with it

Now, you are given a $\widehat{\mu}$ the estimated parameter (In our case the mean sample)

Thus:

$$\begin{aligned} H_1 &: E[x] \neq \widehat{\mu} \\ H_0 &: E[x] = \widehat{\mu} \end{aligned}$$

We define q

$$q = \frac{\overline{x} - \widehat{\mu}}{\frac{\sigma}{N}} \tag{25}$$

Recalling the central limit theorem

The probability density function of $\overline{x}$ under H_0 is approx Gaussian $N\left(\widehat{\mu}, \frac{\sigma}{N}\right)$

Thus

Thus

q under H_0 is approx $N(0,1)$

Then

We can choose an acceptance level ρ with interval $D = [-x_\rho, x_\rho]$ such that q lies on it with probability $1-\rho$.

Thus

Thus

q under H_0 is approx $N(0,1)$

Then

We can choose an acceptance level ρ with interval $D = [-x_\rho, x_\rho]$ such that q lies on it with probability $1-\rho$.

Final Process

First Step

- Given the N experimental samples of x, compute $\overline{x}$ and then q.

Second One

- Choose the significance level ρ.

Third One

- Compute from the corresponding tables for $N(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ with probability $1-\rho$.

Final Process

First Step

- Given the N experimental samples of x, compute $\overline{x}$ and then q.

Second One

- Choose the significance level ρ.

Third One

- Compute from the corresponding tables for $N(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ with probability $1-\rho$.

Final Process

First Step

- Given the N experimental samples of x, compute $\overline{x}$ and then q.

Second One

- Choose the significance level ρ.

Third One

- Compute from the corresponding tables for $N(0,1)$ the acceptance interval $D = [-x_\rho, x_\rho]$ with probability $1-\rho$.

Final Process

Final Step

If $q \in D$ decide H_0 , if not decide H_1.

Final Process

Final Step

If $q \in D$ decide H_0 , if not decide H_1.

Second one

- Basically, all we say is that we expect the resulting value q to lie in the high-percentage $1-\rho$ interval.
- If it does not, then we decide that this is because the assumed mean value is not "correct."

Final Process

Final Step

If $q \in D$ decide H_0 , if not decide H_1.

Second one

- Basically, all we say is that we expect the resulting value q to lie in the high-percentage $1 - \rho$ interval.
- If it does not, then we decide that this is because the assumed mean value is not "correct."

Outline

Example

Let us consider an experiment with a random variable x of $\sigma = 0.23$

- Assume N to be equal to 16 and $\overline{x} = 1.35$
- Adopt $\rho = 0.05$

We will test if the hypothesis $\widehat{\mu} = 1.4$ is true

$$P\left\{-1.97 < \frac{\overline{x} - \widehat{\mu}}{0.23/4} < 1.97\right\} = 0.95$$

Therefore, we accept the hypothesis

- We have $1.237 < \widehat{\mu} < 1.463$

Example

Let us consider an experiment with a random variable x of $\sigma = 0.23$

- Assume N to be equal to 16 and $\overline{x} = 1.35$
- Adopt $\rho = 0.05$

We will test if the hypothesis $\widehat{\mu} = 1.4$ is true

$$P\left\{-1.97 < \frac{\overline{x} - \widehat{\mu}}{0.23/4} < 1.97\right\} = 0.95$$

Therefore, we accept the hypothesis

- We have $1.237 < \widehat{\mu} < 1.463$

Example

Let us consider an experiment with a random variable x of $\sigma = 0.23$

- Assume N to be equal to 16 and $\overline{x} = 1.35$
- Adopt $\rho = 0.05$

We will test if the hypothesis $\widehat{\mu} = 1.4$ is true

$$P\left\{-1.97 < \frac{\overline{x} - \widehat{\mu}}{0.23/4} < 1.97\right\} = 0.95$$

Therefore, we accept the hypothesis

- We have $1.237 < \widehat{\mu} < 1.463$

Outline

Application of the t -Test in Feature Selection

Very Simple

Use the difference $\mu_1 - \mu_2$ for the testing.

Note: Each μ correspond to a class ω_1, ω_2

Application of the t -Test in Feature Selection

Very Simple

Use the difference $\mu_1 - \mu_2$ for the testing.

Note Each μ correspond to a class ω_1, ω_2

Thus, What is the logic?

Basically, if we have two classes... we must see different $\mu's$.

Application of the t -Test in Feature Selection

Very Simple

Use the difference $\mu_1 - \mu_2$ for the testing.

Note Each μ correspond to a class ω_1, ω_2

Thus, What is the logic?

Basically, if we have two classes... we must see different $\mu' s$.

Assume that the variance of the feature values is the same in both

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \tag{26}$$

Application of the t -Test in Feature Selection

Very Simple

Use the difference $\mu_1 - \mu_2$ for the testing.

Note Each μ correspond to a class ω_1, ω_2

Thus, What is the logic?

Basically, if we have two classes... we must see different $\mu' s$.

Assume that the variance of the feature values is the same in both

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \tag{26}$$

What is the Hypothesis?

A very simple one

$$
\begin{aligned}
H_1 &: \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$

The new random variable is

$$z = x - y \tag{27}$$

where x, y denote the random variables corresponding to the values of the feature in the two classes.

Properties

- $E[z] = \mu_1 - \mu_2$
- $\sigma_z^2 = 2\sigma^2$

What is the Hypothesis?

A very simple one

$$
\begin{aligned}
H_1 &: \quad \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \quad \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$

The new random variable is

$$z = x - y \tag{27}$$

where x, y denote the random variables corresponding to the values of the feature in the two classes.

Properties

- $E[z] = \mu_1 - \mu_2$
- $\sigma_z^2 = 2\sigma^2$

What is the Hypothesis?

A very simple one

$$
\begin{aligned}
H_1 &: \quad \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \quad \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$

The new random variable is

$$z = x - y \tag{27}$$

where x, y denote the random variables corresponding to the values of the feature in the two classes.

Properties

- $E[z] = \mu_1 - \mu_2$
- $\sigma_z^2 = 2\sigma^2$

Then

It is possible to prove that z follows the distribution

$$N\left(\mu_1 - \mu_2, \frac{2\sigma^2}{N}\right) \tag{28}$$

So

We can use the following

$$q = \frac{(\overline{x} - \overline{y}) - (\mu_1 - \mu_2)}{s_z\sqrt{\frac{2}{N}}} \tag{29}$$

where

$$s_z^2 = \frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2 + \sum_{i=1}^{N}(y_i - \overline{y})^2\right) \tag{30}$$

Then

It is possible to prove that z follows the distribution

$$N\left(\mu_1-\mu_2,\frac{2\sigma^2}{N}\right) \tag{28}$$

So

We can use the following

$$q=\frac{(\overline{x}-\overline{y})-(\mu_1-\mu_2)}{s_z\sqrt{\frac{2}{N}}} \tag{29}$$

where

$$s_z^2=\frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i-\overline{x})^2+\sum_{i=1}^{N}(y_i-\overline{y})^2\right) \tag{30}$$

Then

It is possible to prove that z follows the distribution

$$N\left(\mu_1-\mu_2, \frac{2\sigma^2}{N}\right) \tag{28}$$

So

We can use the following

$$q=\frac{(\overline{x}-\overline{y})-(\mu_1-\mu_2)}{s_z\sqrt{\frac{2}{N}}} \tag{29}$$

where

$$s_z^2=\frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i-\overline{x})^2+\sum_{i=1}^{N}(y_i-\overline{y})^2\right) \tag{30}$$

Now

It can be shown that $\frac{s_z^2(2N-2)}{\sigma^2}$ follows

- A Chi-Square distribution with $2N-2$ degrees of freedom.

Testing

- q turns out to follow a Chi-Square distribution with $2N-2$ degrees of freedom

Now

It can be shown that $\frac{s_z^2(2N-2)}{\sigma^2}$ follows

- A Chi-Square distribution with $2N-2$ degrees of freedom.

Testing

- q turns out to follow a Chi-Square distribution with $2N-2$ degrees of freedom

Outline

We have two classes

The sample measurements of a feature in two classes are

class ω_1	3.5	3.7	3.9	4.1	3.4	3.5	4.1	3.8	3.6	3.7
class ω_2	3.2	3.6	3.1	3.4	3.0	3.4	2.8	3.1	3.3	3.6

Now, we want to know if the feature is informative enough

$$H_1 : \Delta\mu = \mu_1 - \mu_2 \neq 0$$
$$H_0 : \Delta\mu = \mu_1 - \mu_2 = 0$$

Again, we choose $\rho = 0.05$

$$\omega_1 : \overline{x} = 3.73,\ \hat{\sigma}_1^2 = 0.0601$$
$$\omega_2 : \overline{y} = 3.25,\ \hat{\sigma}_2^2 = 0.0672$$

We have two classes

The sample measurements of a feature in two classes are

class ω_1	3.5	3.7	3.9	4.1	3.4	3.5	4.1	3.8	3.6	3.7
class ω_2	3.2	3.6	3.1	3.4	3.0	3.4	2.8	3.1	3.3	3.6

Now, we want to know if the feature is informative enough

$$
\begin{aligned}
H_1 &: \quad \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \quad \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$

Again, we choose $\rho = 0.05$

$$
\begin{aligned}
\omega_1 &: \overline{x} = 3.73, \; \hat{\sigma}_1^2 = 0.0601 \\
\omega_2 &: \overline{y} = 3.25, \; \hat{\sigma}_2^2 = 0.0672
\end{aligned}
$$

We have two classes

The sample measurements of a feature in two classes are

class ω_1	3.5	3.7	3.9	4.1	3.4	3.5	4.1	3.8	3.6	3.7
class ω_2	3.2	3.6	3.1	3.4	3.0	3.4	2.8	3.1	3.3	3.6

Now, we want to know if the feature is informative enough

$$
\begin{aligned}
H_1 &: \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$

Again, we choose $\rho = 0.05$

$$
\begin{aligned}
\omega_1 &: \overline{x} = 3.73,\ \widehat{\sigma}_1^2 = 0.0601 \\
\omega_2 &: \overline{y} = 3.25,\ \widehat{\sigma}_2^2 = 0.0672
\end{aligned}
$$

Then

For $N = 10$

- $s_z^2 = \frac{1}{2}\left(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2\right)$
- $q = \frac{(\overline{x} - \overline{y} - 0)}{s_z\sqrt{\frac{2}{N}}}$

We have $q = 4.25$

- We have 20-2 = 18 degrees of freedom and significance level 0.05

Then, $D = [-2.10, 2.10]$

- $q = 4.25$ is outside of D, we decide $H_1 : \Delta\mu = \mu_1 - \mu_2 \neq 0$

Then

For $N = 10$

- $s_z^2 = \frac{1}{2}\left(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2\right)$
- $q = \frac{(\overline{x} - \overline{y} - 0)}{s_z\sqrt{\frac{2}{N}}}$

We have $q = 4.25$

- We have 20-2 $=$ 18 degrees of freedom and significance level 0.05

Then, $D = [-2.10, 2.10]$

- $q = 4.25$ is outside of D, we decide $H_1 : \Delta\mu = \mu_1 - \mu_2 \neq 0$

Then

For $N = 10$

- $s_z^2 = \frac{1}{2}\left(\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2\right)$
- $q = \frac{(\overline{x} - \overline{y} - 0)}{s_z\sqrt{\frac{2}{N}}}$

We have $q = 4.25$

- We have 20-2 $=$ 18 degrees of freedom and significance level 0.05

Then, $D = [-2.10, 2.10]$

- $q = 4.25$ is outside of D, we decide $H_1 : \Delta\mu = \mu_1 - \mu_2 \neq 0$

Finally

The means μ_1 and μ_2 are significantly different with $\alpha = 0.05$

- The Feature is selected

Outline

Considering Feature Sets

Something Notable

- The emphasis so far was on individually considered features.

But

- That is, two features may be rich in information, but if they are highly correlated we need not consider both of them.

Then

- Combine features to search for the "best" combination after features have been discarded.

Considering Feature Sets

Something Notable

- The emphasis so far was on individually considered features.

But

- That is, two features may be rich in information, but if they are highly correlated we need not consider both of them.

Then

- Combine features to search for the "best" combination after features have been discarded.

Considering Feature Sets

Something Notable

- The emphasis so far was on individually considered features.

But

- That is, two features may be rich in information, but if they are highly correlated we need not consider both of them.

Then

- Combine features to search for the "best" combination after features have been discarded.

What to do?

Possible

- Use different feature combinations to form the feature vector.
- Train the classifier, and choose the combination resulting in the best classifier performance.

What to do?

Possible

- Use different feature combinations to form the feature vector.
- Train the classifier, and choose the combination resulting in the best classifier performance.

However

- A major disadvantage of this approach is the high complexity.
- Also, local minimum may give misleading results.

What to do?

Possible

- Use different feature combinations to form the feature vector.
- Train the classifier, and choose the combination resulting in the best classifier performance.

However

- A major disadvantage of this approach is the high complexity.
- Also, local minimum may give misleading results.

Better

- Adopt a class separability measure and choose the best feature combination against this cost.

What to do?

Possible

- Use different feature combinations to form the feature vector.
- Train the classifier, and choose the combination resulting in the best classifier performance.

However

- A major disadvantage of this approach is the high complexity.
- Also, local minimum may give misleading results.

Better

- Adopt a class separability measure and choose the best feature combination against this cost.

What to do?

Possible

- Use different feature combinations to form the feature vector.
- Train the classifier, and choose the combination resulting in the best classifier performance.

However

- A major disadvantage of this approach is the high complexity.
- Also, local minimum may give misleading results.

Better

- Adopt a class separability measure and choose the best feature combination against this cost.

Outline

Scatter Matrices

Definition

- These are used as a measure of the way data are scattered in the respective feature space.

Scatter Matrices

Definition

- These are used as a measure of the way data are scattered in the respective feature space.

Within-class Scatter Matrix

$$S_w = \sum_{i=1}^{C} P_i S_i \tag{31}$$

- where C is the number of classes.

- P_i the a priori probability of class ω_i defined as $P_i \cong n_i/N$.
  - n_i is the number of samples in class ω_i.

Scatter Matrices

Definition

- These are used as a measure of the way data are scattered in the respective feature space.

Within-class Scatter Matrix

$$S_w = \sum_{i=1}^{C} P_i S_i \tag{31}$$

- where C is the number of classes.

where

1. $S_i = E\left[\left(\boldsymbol{x} - \boldsymbol{\mu_i}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_i}\right)^T\right]$
2. P_i the a priori probability of class ω_i defined as $P_i \cong n_i/N$.
   1. n_i is the number of samples in class ω_i.

Scatter Matrices

Definition

- These are used as a measure of the way data are scattered in the respective feature space.

Within-class Scatter Matrix

$$S_w = \sum_{i=1}^{C} P_i S_i \tag{31}$$

- where C is the number of classes.

where

1. $S_i = E\left[\left(\boldsymbol{x} - \boldsymbol{\mu_i}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_i}\right)^T\right]$
2. P_i the a priori probability of class ω_i defined as $P_i \cong n_i/N$.
3. n_i is the number of samples in class ω_i.

Scatter Matrices

Definition

- These are used as a measure of the way data are scattered in the respective feature space.

Within-class Scatter Matrix

$$S_w = \sum_{i=1}^{C} P_i S_i \tag{31}$$

- where C is the number of classes.

where

1. $S_i = E\left[(\boldsymbol{x} - \boldsymbol{\mu_i})(\boldsymbol{x} - \boldsymbol{\mu_i})^T\right]$
2. P_i the a priori probability of class ω_i defined as $P_i \cong {}^{n_i}/_{N}$.
    1. n_i is the number of samples in class ω_i.

Scatter Matrices

Between-class scatter matrix

$$S_b = \sum_{i=1}^{C} P_i \left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T \quad (32)$$

Where

$$\boldsymbol{\mu_0} = \sum_{i=1}^{C} P_i \boldsymbol{\mu_i} \quad (33)$$

The global mean.

Mixture scatter matrix

$$S_m = E\left[\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T\right] \quad (34)$$

Note: it can be proved that $S_m = S_w + S_b$

Scatter Matrices

Between-class scatter matrix

$$S_b = \sum_{i=1}^{C} P_i \left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T \tag{32}$$

Where

$$\boldsymbol{\mu_0} = \sum_{i=1}^{C} P_i \boldsymbol{\mu}_i \tag{33}$$

The global mean.

Mixture scatter matrix

$$S_m = E\left[\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T\right] \tag{34}$$

Note: it can be proved that $S_m = S_w + S_b$

Scatter Matrices

Between-class scatter matrix

$$S_b = \sum_{i=1}^{C} P_i \left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T \tag{32}$$

Where

$$\boldsymbol{\mu_0} = \sum_{i=1}^{C} P_i \boldsymbol{\mu}_i \tag{33}$$

The global mean.

Mixture scatter matrix

$$S_m = E\left[\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_0}\right)^T\right] \tag{34}$$

Note: it can be proved that $S_m = S_w + S_b$

Criterion's

First One

$$J_1 = \frac{trace\left\{S_m\right\}}{trace\left\{S_w\right\}} \quad (35)$$

- It takes takes large values when samples in the d-dimensional space are well clustered around their mean, within each class, and the clusters of the different classes are well separated.

Criterion's

First One

$$J_1 = \frac{trace\left\{S_m\right\}}{trace\left\{S_w\right\}} \quad (35)$$

- It takes takes large values when samples in the d-dimensional space are well clustered around their mean, within each class, and the clusters of the different classes are well separated.

Other Criteria are

1. $J_2 = \frac{|S_m|}{|S_w|}$
2. $J_3 = trace\left\{S_w^{-1} S_m\right\}$

Criterion's

First One

$$J_1 = \frac{trace\{S_m\}}{trace\{S_w\}} \tag{35}$$

- It takes takes large values when samples in the d-dimensional space are well clustered around their mean, within each class, and the clusters of the different classes are well separated.

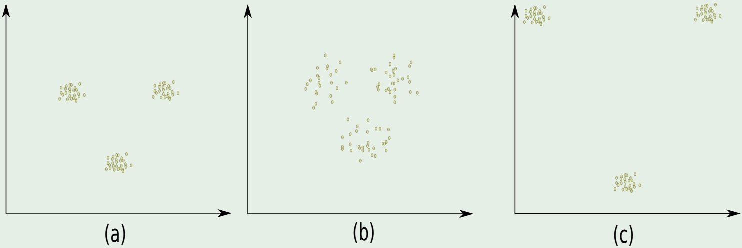
Other Criteria are

1. $J_2 = \frac{|S_m|}{|S_w|}$
2. $J_3 = trace\{S_w^{-1} S_m\}$

Example

We have

- Classes with
  - (a) small within-class variance and small between-class distances,
  - (b) large within- class variance and small between-class distances,
  - (c) small within-class variance and large between-class distances.



Outline

What to do with it

We want to avoid

High Complexities

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\binom{m}{l}$$

with $l = 1, 2, ..., m$

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\begin{pmatrix} m \\ l \end{pmatrix}$$

with $l = 1, 2, ..., m$

We can do better

1. Sequential Backward Selection
2. Sequential Forward Selection
3. Floating Search Methods

However these are sub-optimal methods

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\binom{m}{l}$$

with $l = 1, 2, ..., m$

We can do better

1. Sequential Backward Selection
2. Sequential Forward Selection
3. Floating Search Methods

However these are sub-optimal methods

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\binom{m}{l}$$

with $l = 1, 2, ..., m$

We can do better

1. Sequential Backward Selection
2. Sequential Forward Selection
3. Floating Search Methods

However these are sub-optimal methods

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\begin{pmatrix} m \\ l \end{pmatrix}$$

with $l = 1, 2, ..., m$

We can do better

1. Sequential Backward Selection
2. Sequential Forward Selection
3. Floating Search Methods

However these are sub-optimal methods

What to do with it

We want to avoid

High Complexities

As for example

1. Select a class separability
2. Then, get all possible combinations of features

$$\begin{pmatrix} m \\ l \end{pmatrix}$$

with $l = 1, 2, ..., m$

We can do better

1. Sequential Backward Selection
2. Sequential Forward Selection
3. Floating Search Methods

However these are sub-optimal methods

Outline

For example: Sequential Backward Selection

We have the following example

Given x_1, x_2, x_3, x_4 and we wish to select two of them

Step 1

Adopt a class separability criterion, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]^T$.

Step 2

Eliminate one feature, you get

$$[x_1, x_2, x_3]^T, [x_1, x_2, x_4]^T, [x_1, x_3, x_4]^T, [x_2, x_3, x_4]^T,$$

For example: Sequential Backward Selection

We have the following example

Given x_1, x_2, x_3, x_4 and we wish to select two of them

Step 1

Adopt a class separability criterion, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]^T$.

Step 2

Eliminate one feature, you get

$$[x_1, x_2, x_3]^T, [x_1, x_2, x_4]^T, [x_1, x_3, x_4]^T, [x_2, x_3, x_4]^T,$$

For example: Sequential Backward Selection

We have the following example

Given x_1, x_2, x_3, x_4 and we wish to select two of them

Step 1

Adopt a class separability criterion, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]^T$.

Step 2

Eliminate one feature, you get

$$[x_1, x_2, x_3]^T, [x_1, x_2, x_4]^T, [x_1, x_3, x_4]^T, [x_2, x_3, x_4]^T,$$

For example: Sequential Backward Selection

You use your criterion C

Thus the winner is $[x_1, x_2, x_3]^T$

Step 3

Now, eliminate a feature and generate $[x_1, x_2]^T, [x_1, x_3]^T, [x_2, x_3]^T$,

Use criterion C

To select the best one

For example: Sequential Backward Selection

You use your criterion C

Thus the winner is $[x_1, x_2, x_3]^T$

Step 3

Now, eliminate a feature and generate $[x_1, x_2]^T, [x_1, x_3]^T, [x_2, x_3]^T$,

Use criterion C

To select the best one

For example: Sequential Backward Selection

You use your criterion C

Thus the winner is $[x_1, x_2, x_3]^T$

Step 3

Now, eliminate a feature and generate $[x_1, x_2]^T, [x_1, x_3]^T, [x_2, x_3]^T,$

Use criterion C

To select the best one

Complexity of the Method

Complexity

Thus, starting from m, at each step we drop out one feature from the "best" combination until we obtain a vector of l features.

Complexity of the Method

Complexity

Thus, starting from m, at each step we drop out one feature from the "best" combination until we obtain a vector of l features.

Thus, we need

$1 + 1/2((m+1)m - l(l+1))$ combinations

- It suffers of the so-called nesting effect
  - Once a feature is discarded, there is no way to reconsider that feature again.

Complexity of the Method

Complexity

Thus, starting from m, at each step we drop out one feature from the "best" combination until we obtain a vector of l features.

Thus, we need

$1 + 1/2((m+1)m - l(l+1))$ combinations

However

- The method is sub-optimal
- It suffers of the so called nesting-effect
  - Once a feature is discarded, there is no way to reconsider that feature again.

Complexity of the Method

Complexity

Thus, starting from m, at each step we drop out one feature from the "best" combination until we obtain a vector of l features.

Thus, we need

$1 + 1/2((m+1)m - l(l+1))$ combinations

However

- The method is sub-optimal
- It suffers of the so called nesting-effect
  - Once a feature is discarded, there is no way to reconsider that feature again.

Complexity of the Method

Complexity

Thus, starting from m, at each step we drop out one feature from the "best" combination until we obtain a vector of l features.

Thus, we need

$1 + 1/2((m+1)m - l(l+1))$ combinations

However

- The method is sub-optimal
- It suffers of the so called nesting-effect
  - Once a feature is discarded, there is no way to reconsider that feature again.

Similar Problem

For

- Sequential Forward Selection

We can overcome this by using

- Floating Search Methods

A more elegant methods are the ones based on

- Dynamic Programming
- Branch and Bound

Similar Problem

For

- Sequential Forward Selection

We can overcome this by using

- Floating Search Methods

A more elegant methods are the ones based on

- Dynamic Programming
- Branch and Bound

Similar Problem

For

- Sequential Forward Selection

We can overcome this by using

- Floating Search Methods

A more elegant methods are the ones based on

- Dynamic Programming
- Branch and Bound

Outline

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest

- Subset Selection produces a model that is interpretable,
- It possibly produces lower prediction error than the full model.

- It does not reduce the prediction error of the full model.

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest

- Subset Selection produces a model that is interpretable,
- It possibly produces lower prediction error than the full model.

However given process

- it often exhibits high variance,
- It does not reduce the prediction error of the full model.

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest

- Subset Selection produces a model that is interpretable,
- It possibly produces lower prediction error than the full model.

However given process

- it often exhibits high variance,
- It does not reduce the prediction error of the full model.

Therefore

- Shrinkage methods are more continuous avoiding high variability.

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest

- Subset Selection produces a model that is interpretable,
- It possibly produces lower prediction error than the full model.

However given process

- it often exhibits high variance,
- It does not reduce the prediction error of the full model.

Therefore

- Shrinkage methods are more continuous avoiding high variability.

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest

- Subset Selection produces a model that is interpretable,
- It possibly produces lower prediction error than the full model.

However given process

- it often exhibits high variance,
- It does not reduce the prediction error of the full model.

Therefore

- Shrinkage methods are more continuous avoiding high variability.

Outline

The house example

Imagine the following data set

Now assume that we use LSE

For the fitting

$$\frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{w}}\left(x_i\right)-y_i\right)^2$$

We can then run one of our machine to see what minimize better the previous equation

Question: Did you notice that I did not impose any structure to $h_w(x)$?

Now assume that we use LSE

For the fitting

$$\frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{w}}\left(x_i\right)-y_i\right)^2$$

We can then run one of our machine to see what minimize better the previous equation

Question: Did you notice that I did not impose any structure to $h_{\boldsymbol{w}}\left(x\right)$?

Then, First fitting

What about using $h_1(x) = w_0 + w_1 x + w_2 x^2$?

Second fitting

What about using $h_2(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$?

Therefore, we have a problem

We get weird overfitting effects!!!

What do we do? What about minimizing the influence of w_3, w_4, w_5?

How do we do that?

$$\min_{w} \frac{1}{2}\sum_{i=1}^{N}\left(h_w\left(x_i\right)-y_i\right)^2$$

What about integrating those values to the cost function? Ideas

Therefore, we have a problem

We get weird overfitting effects!!!

What do we do? What about minimizing the influence of w_3, w_4, w_5?

How do we do that?

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} \left(h_{\boldsymbol{w}}\left(x_i\right) - y_i\right)^2$$

What about integrating those values to the cost function? Ideas

Outline

We have

Regularization intuition is as follow

Small values for parameters $w_0, w_1, w_2, ..., w_n$

It implies

1. "Simpler" function
2. Less prone to overfitting

We have

Regularization intuition is as follow

Small values for parameters $w_0, w_1, w_2, ..., w_n$

It implies

1. "Simpler" function
2. Less prone to overfitting

We can do the previous idea for the other parameters

We can do the same for the other parameters

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{w}}\left(x_i\right)-y_i\right)^2+\sum_{i=1}^{d}\lambda_i w_i^2 \tag{36}$$

However handling such many parameters can be so difficult

Combinatorial problem in reality!!!

We can do the previous idea for the other parameters

We can do the same for the other parameters

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{w}}\left(x_i\right)-y_i\right)^2+\sum_{i=1}^{d}\lambda_i w_i^2 \tag{36}$$

However handling such many parameters can be so difficult

Combinatorial problem in reality!!!

Better, we can

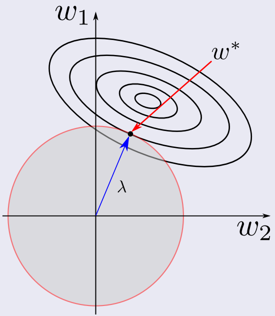
We better use the following

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} \left(h_{\boldsymbol{w}}\left(x_i\right) - y_i\right)^2 + \lambda \sum_{i=1}^{d} w_i^2 \tag{37}$$

Graphically

Geometrically Equivalent to

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}\right)^2 + \lambda\sum_{i=1}^{d+1} w_i^2$$



Outline

Ridge Regression

Equation

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j-1}^{d} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right\}$$

Here

- $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage

Ridge Regression

Equation

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j-1}^{d} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right\}$$

Here

- $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage

Therefore

The Larger $\lambda \geq 0$

- The coefficients are shrunk toward zero (and each other).

This is also used in Neural Networks

- where it is known as weight decay

Therefore

The Larger $\lambda \geq 0$

- The coefficients are shrunk toward zero (and each other).

This is also used in Neural Networks

- where it is known as weight decay

This is also can be written

Optimization Solution

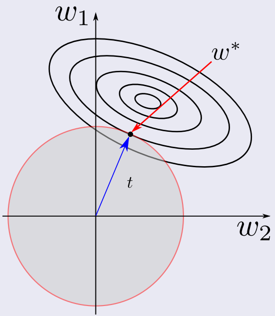
$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j-1}^{d} x_{ij} w_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{d} w_j^2 < t$$

Graphically

Geometrically Equivalent to

$$\begin{array}{l} \arg\min \ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2 \\ \text{subject to } \sum_{i=1}^{d+1} w_i^2 < t \end{array}$$



Outline

Important

as a number

We have

The ridge solutions are not equivariant under scaling of the inputs.

Thus, the need to standardize the input data

Before Solving:

$$\arg\min_{w} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{d} w_j^2 < t$$

Important

as a number

We have

The ridge solutions are not equivariant under scaling of the inputs.

Thus, the need to standardize the input data

Before Solving:

$$\arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j-1}^{d} x_{ij} w_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{d} w_j^2 < t$$

Here

Notice that w_0 is not being penalized

- Penalizing w_0 would make the procedure depend on the origin chosen for y_i.

Adding a constant c to each of the targets y_i

- It would not simply result in a shift of the predictioas a numberns by the same amount c.

Here

Notice that w_0 is not being penalized

- Penalizing w_0 would make the procedure depend on the origin chosen for y_i.

Adding a constant c to each of the targets y_i

- It would not simply result in a shift of the predictioas a numberns by the same amount c.

Thus

First

- each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$.

Then, we estimate w_0

$$w_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Thus

First

- each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$.

Then,we estimate w_0

$$w_0 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Thus after centering

Now the data matrix $\boldsymbol{X}$ has d dimensions

$$RSS(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda \boldsymbol{w}^T \boldsymbol{w}$$

We have seen that the Ridge Regression solution is equivalent to

$$\widehat{w}^{Ridge} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Thus after centering

Now the data matrix $\boldsymbol{X}$ has d dimensions

$$RSS\left(\lambda\right) = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)^{T}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right) + \lambda\boldsymbol{w}^{T}\boldsymbol{w}$$

We have seen that the Ridge Regression solution is equivalent to

$$\widehat{\boldsymbol{w}}^{Ridge} = \left(\boldsymbol{X}^{T}\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{y}.$$

Outline

Now

as a number

We can define the degree of freedom by looking at the SVD, $\boldsymbol{X}$ $N \times d$

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

With orthogonal matrices

1. The columns of $\boldsymbol{U}$ span the column space of $\boldsymbol{X}$
2. The columns of $\boldsymbol{V}$ span the row space of $\boldsymbol{X}$

And with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ singular values

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_d \end{pmatrix}$$

Now

as a number

We can define the degree of freedom by looking at the SVD, $\boldsymbol{X}$ $N \times d$

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

With orthogonal matrices

1. The columns of $\boldsymbol{U}$ span the column space of $\boldsymbol{X}$
2. The columns of $\boldsymbol{V}$ span the row space of $\boldsymbol{X}$

And with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ singular values

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_d \end{pmatrix}$$

Now

as a number

We can define the degree of freedom by looking at the SVD, $\boldsymbol{X}$ $N \times d$

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

With orthogonal matrices

1. The columns of $\boldsymbol{U}$ span the column space of $\boldsymbol{X}$
2. The columns of $\boldsymbol{V}$ span the row space of $\boldsymbol{X}$

And with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ singular values

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_d \end{pmatrix}$$

Therefore, for the Ridge Regression

We have that

$$\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Thus, we have

$$\begin{aligned}\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} &= UDV^T\left(VDU^TUDV^T + \lambda VIV^T\right)VDU^T\boldsymbol{y} \\ &= UD\left(D^2 + \lambda I\right)^{-1}DU^T\boldsymbol{y}\end{aligned}$$

Finally

$$\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} = \sum_{i=1}^{d}\frac{\lambda_i^2}{\lambda_i^2 + \lambda}\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{y}$$

Therefore, for the Ridge Regression

We have that

$$\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Thus, we have

$$\begin{aligned}\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + \lambda\boldsymbol{V}I\boldsymbol{V}^T\right)\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\\ &= \boldsymbol{U}\boldsymbol{D}\left(\boldsymbol{D}^2 + \lambda I\right)^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\end{aligned}$$

Finally

$$X\widehat{w}^{Ridge} = \sum_{i=1}^{d}\frac{\lambda_i^2}{\lambda_i^2+\lambda}u_iu_i^Ty$$

Therefore, for the Ridge Regression

We have that

$$\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Thus, we have

$$\begin{aligned}\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + \lambda\boldsymbol{V}\boldsymbol{I}\boldsymbol{V}^T\right)\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\\ &= \boldsymbol{U}\boldsymbol{D}\left(\boldsymbol{D}^2 + \lambda I\right)^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\end{aligned}$$

Finally

$$\boldsymbol{X}\widehat{\boldsymbol{w}}^{Ridge} = \sum_{i=1}^{d}\frac{\lambda_i^2}{\lambda_i^2 + \lambda}\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{y}$$

Therefore

We have that given $\lambda \geq 0$

$$\frac{\lambda_i^2}{\lambda_i^2 + \lambda} \leq 1$$

Thus, like Linear Regression

- Ridge Regression computes the coordinates of y with respect to the orthonormal basis U.

Then, it shrinks the coordinates by a factor of $\frac{\lambda_i^2}{\lambda_i^2+\lambda}$

- Meaning the smaller is a λ_j the larger shrinkage you have!!!

Therefore

We have that given $\lambda \geq 0$

$$\frac{\lambda_i^2}{\lambda_i^2 + \lambda} \leq 1$$

Thus, like Linear Regression

- Ridge Regression computes the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$.

Then, it shrinks the coordinates by a factor of $\frac{\lambda_i^2}{\lambda_i^2 + \lambda}$

- Meaning the smaller is a λ_i the larger shrinkage you have!!!

Therefore

We have that given $\lambda \geq 0$

$$\frac{\lambda_i^2}{\lambda_i^2 + \lambda} \leq 1$$

Thus, like Linear Regression

- Ridge Regression computes the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$.

Then, it shrinks the coordinates by a factor of $\frac{\lambda_i^2}{\lambda_i^2+\lambda}$

- Meaning the smaller is a λ_j the larger shrinkage you have!!!

Therefore

This behaves has what we know as Principal Component Analysis

- We will look at this later...

Outline

Thus

Using Our Singular Value Decomposition

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Therefore the Sample Variance, for centered data, is defined as

$$C_X = \frac{1}{N}X^T X$$

Becomes which is called an eigen decomposition

$$C_X = \frac{1}{N}VD^2V^T$$

Thus

Using Our Singular Value Decomposition

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Therefore the Sample Variance, for centered data, is defined as

$$C_X = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$$

Becomes which is called an eigen decomposition

$$C_X = \frac{1}{N}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Thus

Using Our Singular Value Decomposition

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Therefore the Sample Variance, for centered data, is defined as

$$C_X = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$$

Becomes which is called an eigen decomposition

$$C_X = \frac{1}{N}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$$

Goal of SVD

Find the best transformation with the minimal noise and redundancy

$$Y = \boldsymbol{X} A$$

Thus, we are looking by a orthonormal basis vectors

- Grouped as A

Covariance matrix captures all the information about X

- Only true for exponential family distributions

Goal of SVD

Find the best transformation with the minimal noise and redundancy

$$Y = \boldsymbol{X}A$$

Thus, we are looking by a orthonormal basis vectors

- Grouped as A

Covariance matrix captures all the information about X

- Only true for exponential family distributions

Goal of SVD

Find the best transformation with the minimal noise and redundancy

$$Y = \boldsymbol{X}A$$

Thus, we are looking by a orthonormal basis vectors

- Grouped as A

Covariance matrix captures all the information about $\boldsymbol{X}$

- Only true for exponential family distributions

First

Find the Covariance of Y

$$
\begin{aligned}
C_Y &= \frac{1}{N} Y^T Y \\
&= \frac{1}{N} (\boldsymbol{X} A)^T (\boldsymbol{X} A) \\
&= \frac{1}{N} A^T \boldsymbol{X}^T \boldsymbol{X} A
\end{aligned}
$$

Therefore

Find the direction for which the variance is maximized

$$v_1 = \arg\max_{v_1} \; var\left(Xv_1\right)$$
$$\text{s.t.} \;\; v_1^T v_1 = 1$$

We use the sample variance

$$var\left(Xv_1\right) = \frac{1}{N}\left(Xv_1\right)^T\left(Xv_1\right) = \frac{1}{N} v_1^T X^T X v_1 = v_1^T C_X v_1$$

Therefore

Find the direction for which the variance is maximized

$$\boldsymbol{v}_1 = \arg\max_{v_1} \ var\left(\boldsymbol{X}\boldsymbol{v}_1\right)$$

$$\text{s.t. } \boldsymbol{v}_1^T\boldsymbol{v}_1 = 1$$

We use the sample variance

$$var\left(\boldsymbol{X}v_1\right) = \frac{1}{N}\left(\boldsymbol{X}v_1\right)^T\left(\boldsymbol{X}v_1\right) = \frac{1}{N}v_1^T\boldsymbol{X}^T\boldsymbol{X}v_1 = v_1^T C_X v_1$$

Therefore

Find the direction for which the variance is maximized

$$\boldsymbol{v}_1 = \arg\max_{v_1} \; var\left(\boldsymbol{X}\boldsymbol{v}_1\right)$$
$$\text{s.t. } \boldsymbol{v}_1^T\boldsymbol{v}_1 = 1$$

We use the sample variance

$$var\left(\boldsymbol{X}\boldsymbol{v}_1\right) = \frac{1}{N}\left(\boldsymbol{X}\boldsymbol{v}_1\right)^T\left(\boldsymbol{X}\boldsymbol{v}_1\right) = \frac{\mathbf{1}}{\boldsymbol{N}}\boldsymbol{v}_1^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}_1 = \boldsymbol{v}_1^T C_X \boldsymbol{v}_1$$

Thus

We have the Lagrangian

$$L\left(v_1, \lambda_1\right) = \boldsymbol{v}_1^T C_X \boldsymbol{v}_1 + \lambda_1 \left(1 - \boldsymbol{v}_1^T \boldsymbol{v}_1\right)$$

Thus, as in the PCA, $\boldsymbol{v}_1$ is an eigenvector of C_X

$$C_X \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$$

With Variance

$$var\left(\boldsymbol{X}\boldsymbol{v}_1\right) = \boldsymbol{v}_1^T \frac{1}{N} V \boldsymbol{D}^2 V^T \boldsymbol{v}_1$$

Thus

We have the Lagrangian

$$L\left(v_1, \lambda_1\right) = \boldsymbol{v}_1^T C_X \boldsymbol{v}_1 + \lambda_1 \left(1 - \boldsymbol{v}_1^T \boldsymbol{v}_1\right)$$

Thus, as in the PCA, $\boldsymbol{v}_1$ is an eigenvector of C_X

$$C_X \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$$

With Variance

$$var\left(X\boldsymbol{v}_1\right) = \boldsymbol{v}_1^T \frac{1}{N} V D^2 V^T \boldsymbol{v}_1$$

Thus

We have the Lagrangian

$$L\left(v_1, \lambda_1\right) = \boldsymbol{v}_1^T C_X \boldsymbol{v}_1 + \lambda_1 \left(1 - \boldsymbol{v}_1^T \boldsymbol{v}_1\right)$$

Thus, as in the PCA, $\boldsymbol{v}_1$ is an eigenvector of C_X

$$C_X \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$$

With Variance

$$var\left(\boldsymbol{X}\boldsymbol{v}_1\right) = \boldsymbol{v}_1^T \frac{1}{N} \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T \boldsymbol{v}_1$$

Therefore

We have

$$var\left(X\boldsymbol{v}_1\right) = \frac{1}{N}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}\begin{bmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d^2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then

$$var\left(X\boldsymbol{v}_1\right) = \frac{1}{N}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}\begin{bmatrix} \lambda_1^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \frac{\lambda_1^2}{N}$$

Therefore

We have

$$var\left(X\boldsymbol{v}_1\right) = \frac{1}{N}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}\begin{bmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d^2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then

$$var\left(X\boldsymbol{v}_1\right) = \frac{1}{N}\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}\begin{bmatrix} \lambda_1^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \frac{\lambda_1^2}{N}$$

Meaning

The First Principal Component Achieves maximum variance

- When the associated constant to the Sample Variance is equal to $\frac{\lambda_1^2}{N}$

In fact

We have that

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 = \lambda_1 \boldsymbol{u}_1$$

This variable $\boldsymbol{z}_1$ is called the first principal component of $\boldsymbol{X}$

- Therefore $\boldsymbol{u}_1$ is called the normalized first principal component!!!

In fact

We have that

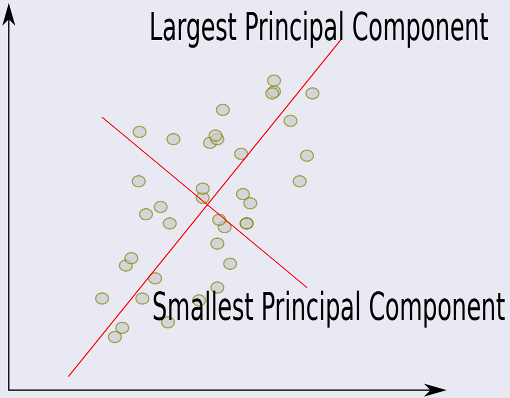
$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 = \lambda_1 \boldsymbol{u}_1$$

This variable $\boldsymbol{z}_1$ is called the first principal component of $\boldsymbol{X}$

- Therefore $\boldsymbol{u}_1$ is called the normalized first principal component!!!

Geometrically

We have



We can define the following function

Effective Degrees of Freedom

About the Regularization Parameter λ

$$
\begin{aligned}
\text{df}(\lambda) =& tr\left[\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}+\lambda I\right)^{-1}\boldsymbol{X}^T\right]\\
=& tr\left[\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T+\lambda I\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\right]
\end{aligned}
$$

Therefore, the inner matrix

$$
\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T+\lambda I\right)^{-1}=\begin{pmatrix}
\frac{1}{\lambda_1^2+\lambda} & 0 & \cdots & 0\\
0 & \frac{1}{\lambda_2^2+\lambda} & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots\\
0 & 0 & \cdots & \frac{1}{\lambda_d^2+\lambda}
\end{pmatrix}
$$

We can define the following function

Effective Degrees of Freedom

About the Regularization Parameter λ

$$\begin{aligned}\mathsf{df}(\lambda) =& tr\left[\boldsymbol{X}\left(\boldsymbol{X}^{T}\boldsymbol{X}+\lambda I\right)^{-1}\boldsymbol{X}^{T}\right]\\ =& tr\left[\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}+\lambda I\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{T}\right]\end{aligned}$$

Therefore, the inner matrix

$$\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{T}+\lambda I\right)^{-1}=\begin{pmatrix}\frac{1}{\lambda_1^2+\lambda} & 0 & \cdots & 0\\ 0 & \frac{1}{\lambda_2^2+\lambda} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\lambda_d^2+\lambda}\end{pmatrix}$$

Finally

We have

$$\mathrm{df}(\lambda) = tr\left[\boldsymbol{D}^2\left(\boldsymbol{D}^2+\lambda I\right)^{-1}\right] = tr\begin{pmatrix} \frac{\lambda_1^2}{\lambda_1^2+\lambda} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2^2}{\lambda_2^2+\lambda} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_d^2}{\lambda_d^2+\lambda} \end{pmatrix}$$

Therefore

$$\mathrm{df}(\lambda) = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i^2+\lambda}$$

Finally

We have

$$\text{df}(\lambda) = tr\left[\boldsymbol{D}^2\left(\boldsymbol{D}^2 + \lambda I\right)^{-1}\right] = tr\begin{pmatrix} \frac{\lambda_1^2}{\lambda_1^2+\lambda} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2^2}{\lambda_2^2+\lambda} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_d^2}{\lambda_d^2+\lambda} \end{pmatrix}$$

Therefore

$$\text{df}(\lambda) = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i^2 + \lambda}$$

Degrees of Freedom in Linear Regression

Usually in a linear-regression fit with p variables

- The degrees-of-freedom of the fit is $d =$ number of features

This is important

- We assume all d coefficients in a ridge fit will be non-zero.
    - They are fit in a restricted fashion controlled by λ.

We have the following cases

- If $\mathrm{df}(\lambda) = d$ when $\lambda = 0$.
- If $\mathrm{df}(\lambda) \to 0$ as $\lambda \to \infty$

Degrees of Freedom in Linear Regression

Usually in a linear-regression fit with p variables

- The degrees-of-freedom of the fit is $d =$ number of features

This is important

- We assume all d coefficients in a ridge fit will be non-zero.
  - They are fit in a restricted fashion controlled by λ.

We have the following cases

- If $\text{df}(\lambda) = d$ when $\lambda = 0$.
- If $\text{df}(\lambda) \to 0$ as $\lambda \to \infty$

Degrees of Freedom in Linear Regression

Usually in a linear-regression fit with p variables

- The degrees-of-freedom of the fit is $d =$ number of features

This is important

- We assume all d coefficients in a ridge fit will be non-zero.
  - They are fit in a restricted fashion controlled by λ.

We have the following cases

- If df $(\lambda) = d$ when $\lambda = 0$.
- If df $(\lambda) \rightarrow 0$ as $\lambda \rightarrow \infty$

From Hastie et. al page 63

Cancer Data using a Linear Model and df $(\lambda) = 5$

	LSE	Subset Selection	Ridge
Test Error	0.521	0.492	0.492
Std Error	0.179	0.143	0.1645

Outline

Least Absolute Shrinkage and Selection Operator (LASSO)

It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's nonnegative garrote

$$\widehat{\boldsymbol{w}}^{garrote} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2 + N\lambda \sum_{j=1}^{d} w_j$$

s.t. $w_j > 0 \ \forall j$

This is quite derivable

However, Tibshirani realized that you could get a more flexible model by using the absolute value at the constraint!!!

Robert Tibshirani proposed the use of the L_1 norm

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|$$

Least Absolute Shrinkage and Selection Operator (LASSO)

It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's nonnegative garrote

$$\widehat{\boldsymbol{w}}^{garrote} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2 + N\lambda \sum_{j=1}^{d} w_j$$

s.t. $w_j > 0 \ \forall j$

This is quite derivable

However, Tibshirani realized that you could get a more flexible model by using the absolute value at the constraint!!!

Robert Tibshirani proposed the use of the L_1 norm

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|$$

Least Absolute Shrinkage and Selection Operator (LASSO)

It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's nonnegative garrote

$$\widehat{\boldsymbol{w}}^{garrote} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2 + N\lambda \sum_{j=1}^{d} w_j$$

s.t. $w_j > 0 \ \forall j$

This is quite derivable

However, Tibshirani realized that you could get a more flexible model by using the absolute value at the constraint!!!

Robert Tibshirani proposed the use of the L_1 norm

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |w_i|$$

The Final Optimization Problem

LASSO

$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2$$

$$\text{s.t. } \sum_{i=1}^{d} |w_i| \leq t$$

This is not derivable

More advanced methods are necessary to solve this problem!!!

The Final Optimization Problem

LASSO

$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2$$

$$\text{s.t.} \ \sum_{i=1}^{d} |w_i| \leq t$$

This is not derivable

More advanced methods are necessary to solve this problem!!!

Outline

The Lagrangian Version

The Lagrangian

$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \right\}$$

However

You have other regularizations as $\|w\|_2 = \sqrt{\sum_{i=1}^{d} |w_i|^2}$

The Lagrangian Version

The Lagrangian

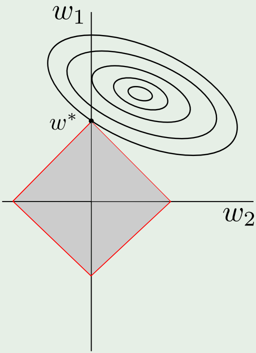
$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i| \right\}$$

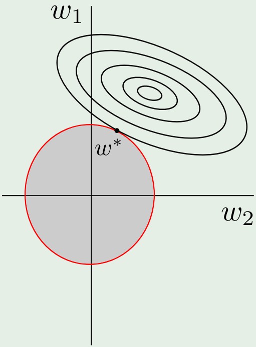
However

You have other regularizations as $\|\boldsymbol{w}\|_2 = \sqrt{\sum_{i=1}^{d} |w_i|^2}$

Graphically

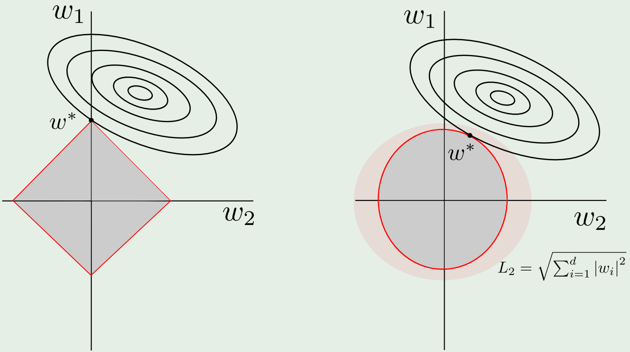
The first area correspond to the L_1 regularization and the second one?





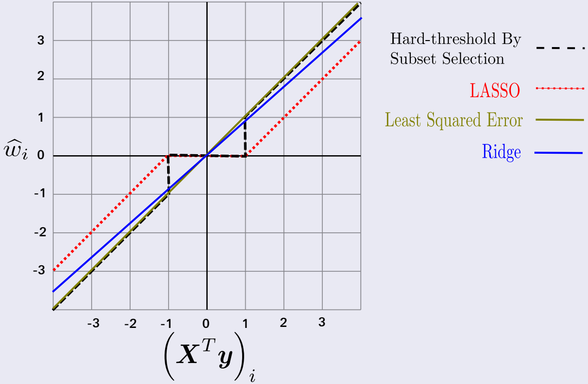
Graphically

Yes the circle defined as $\|\boldsymbol{w}\|_2 = \sqrt{\sum_{i=1}^{d} |w_i|^2}$



For Example

In the Case of $\boldsymbol{X}$ is a Orthogonal Matrix



The seminal paper by Robert Tibshirani

An initial study of this regularization can be seen in

"Regression Shrinkage and Selection via the LASSO" by Robert Tibshirani - 1996

This out the scope of this class

However, it is worth noticing that the most efficient method for solving LASSO problems is

"Pathwise Coordinate Optimization" By Jerome Friedman, Trevor Hastie, Holger Ho and Robert Tibshirani

Nevertheless

It will be a great seminar paper!!!

Generalization

We can generalize ridge regression and the lasso, and view them as Bayes estimates

$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i|^q \right\} \text{ with } q \geq 0$$

This out the scope of this class

However, it is worth noticing that the most efficient method for solving LASSO problems is

"Pathwise Coordinate Optimization" By Jerome Friedman, Trevor Hastie, Holger Ho and Robert Tibshirani

Nevertheless

It will be a great seminar paper!!!

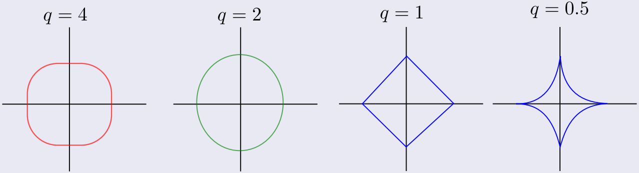
Generalization

We can generalize ridge regression and the lasso, and view them as Bayes estimates

$$\widehat{\boldsymbol{w}}^{LASSO} = \arg\min_{\boldsymbol{w}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w} \right)^2 + \lambda \sum_{i=1}^{d} |w_i|^q \right\} \text{ with } q \geq 0$$

For Example

We have when $d = 2$

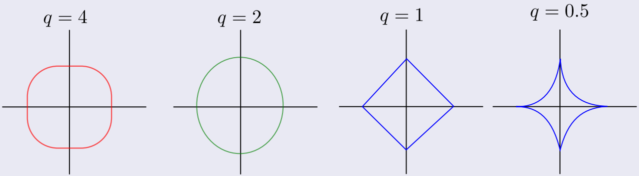


Here, when $q > 1$

- You are having a derivable Lagrangian, but you lose the LASSO properties

For Example

We have when $d = 2$



Here, when $q > 1$

- You are having a derivable Lagrangian, but you lose the LASSO properties

Therefore

Zou and Hastie (2005) introduced the elastic- net penalty

$$\lambda \sum_{i=1}^{d} \left\{ \alpha w_i^2 + (1-\alpha) |w_i| \right\}$$

This is Basically

- A Compromise Between the Ridge and LASSO.

Therefore

Zou and Hastie (2005) introduced the elastic- net penalty

$$\lambda \sum_{i=1}^{d} \left\{ \alpha w_i^2 + (1-\alpha) |w_i| \right\}$$

This is Basically

- A Compromise Between the Ridge and LASSO.