Introduction to Machine Learning

Introduction to Bayesian Classification

Andres Mendez-Vazquez

June 3, 2020

Outline

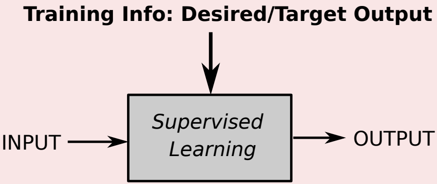
Outline

Classification Problem

Goal

Given $\boldsymbol{x}_{new}$, provide $f(\boldsymbol{x}_{new})$

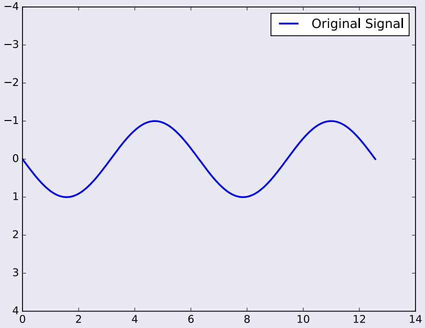
The Machinery in General looks...

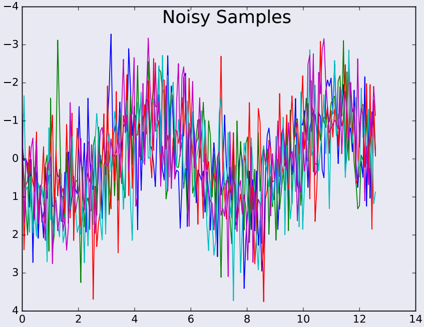


Outline

How do we handle Noise?

Imagine the following signal from $\sin(\theta)$





What if we know the noise?

Given a series of observed samples $\{\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2, ..., \widehat{\boldsymbol{x}}_N\}$ with noise $\epsilon \sim N(0,1)$

We could use our knowledge on the noise, for example additive:

$$\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \epsilon$$

We can use our knowledge of probability to remove such noise

$$E\left[\widehat{\boldsymbol{x}}_i\right] = E\left[\boldsymbol{x}_i + \epsilon\right] = E\left[\boldsymbol{x}_i\right] + E\left[\epsilon\right]$$

Then, because $E\left[\epsilon\right] = 0$

$$E\left[\boldsymbol{x}_i\right] = E\left[\widehat{\boldsymbol{x}}_i\right] \approx \frac{1}{N}\sum_{i=1}^{N} \widehat{\boldsymbol{x}}_i$$

What if we know the noise?

Given a series of observed samples $\left\{\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2, ..., \widehat{\boldsymbol{x}}_N\right\}$ with noise $\epsilon \sim N\left(0,1\right)$

We could use our knowledge on the noise, for example additive:

$$\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \epsilon$$

We can use our knowledge of probability to remove such noise

$$E\left[\widehat{\boldsymbol{x}}_i\right] = E\left[\boldsymbol{x}_i + \epsilon\right] = E\left[\boldsymbol{x}_i\right] + E\left[\epsilon\right]$$

Then, because $E\left[\epsilon\right] = 0$

$$E\left[\boldsymbol{x}_i\right] = E\left[\widehat{\boldsymbol{x}}_i\right] \approx \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{x}}_i$$

What if we know the noise?

Given a series of observed samples $\{\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2, ..., \widehat{\boldsymbol{x}}_N\}$ with noise $\epsilon \sim N(0,1)$

We could use our knowledge on the noise, for example additive:

$$\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \epsilon$$

We can use our knowledge of probability to remove such noise

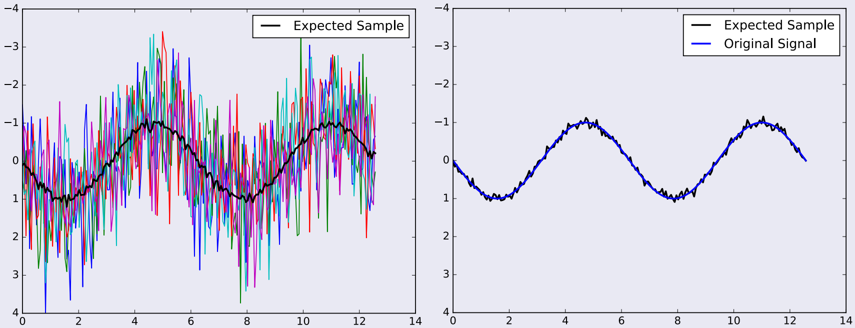
$$E\left[\widehat{\boldsymbol{x}}_i\right] = E\left[\boldsymbol{x}_i + \epsilon\right] = E\left[\boldsymbol{x}_i\right] + E\left[\epsilon\right]$$

Then, because $E\left[\epsilon\right] = 0$

$$E\left[\boldsymbol{x}_i\right] = E\left[\widehat{\boldsymbol{x}}_i\right] \approx \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{x}}_i$$

In our example

We have a nice result



Therefore, we have

The Bayesian Models

- They allow to deal with noise from the samples

Quite different from the deterministic models so far

- Unless Samples are Preprocessed to Reduce the Noise

Something that people in area as Control tend to do

- The importance of Filters as Kalman Filters

Therefore, we have

The Bayesian Models

- They allow to deal with noise from the samples

Quite different from the deterministic models so far

- Unless Samples are Preprocessed to Reduce the Noise

Something that people in area as Control tend to do

- The importance of Filters as Kalman Filters

Therefore, we have

The Bayesian Models

- They allow to deal with noise from the samples

Quite different from the deterministic models so far

- Unless Samples are Preprocessed to Reduce the Noise

Something that people in area as Control tend to do

- The importance of Filters as Kalman Filters

Outline

Example

Given a Spoken Language

The task is to determine the language that someone is speaking

- Therefore, they try to determine the spoken language based on such learning

- Quite easier!!!

Example

Given a Spoken Language

The task is to determine the language that someone is speaking

Generative Models

- They try to learn each language.
- Therefore, they try to determine the spoken language based in such learning.

- Quite easier!!!

Example

Given a Spoken Language

The task is to determine the language that someone is speaking

Generative Models

- They try to learn each language.
- Therefore, they try to determine the spoken language based in such learning.

Discriminative Models

- They try to determine the linguistic differences without learning any language!!!
- Quite easier!!!

Example

Given a Spoken Language

The task is to determine the language that someone is speaking

Generative Models

- They try to learn each language.
- Therefore, they try to determine the spoken language based in such learning.

Discriminative Models

- They try to determine the linguistic differences without learning any language!!!
- Quite easier!!!

Example

Given a Spoken Language

The task is to determine the language that someone is speaking

Generative Models

- They try to learn each language.
- Therefore, they try to determine the spoken language based in such learning.

Discriminative Models

- They try to determine the linguistic differences without learning any language!!!
- Quite easier!!!

Therefore

Generative Methods

1. Model class-conditional pdfs and prior probabilities.
2. "Generative" since sampling can generate synthetic data points.

- Mixtures of Gaussians, Mixtures of Experts, Hidden Markov Models (HMM)
- Sigmoidal Belief Networks, Bayesian Networks, Markov Random Fields

Therefore

Generative Methods

1. Model class-conditional pdfs and prior probabilities.
2. “Generative” since sampling can generate synthetic data points.

Examples

- Gaussians, Naïve Bayes, Mixtures of Multinomials.
- Mixtures of Gaussians, Mixtures of Experts, Hidden Markov Models (HMM).
- Sigmoidal Belief Networks, Bayesian Networks, Markov Random Fields.

Therefore

Generative Methods

1. Model class-conditional pdfs and prior probabilities.
2. "Generative" since sampling can generate synthetic data points.

Examples

- Gaussians, Naïve Bayes, Mixtures of Multinomials.
- Mixtures of Gaussians, Mixtures of Experts, Hidden Markov Models (HMM).
- Sigmoidal Belief Networks, Bayesian Networks, Markov Random Fields.

Therefore

Generative Methods

1. Model class-conditional pdfs and prior probabilities.
2. "Generative" since sampling can generate synthetic data points.

Examples

- Gaussians, Naïve Bayes, Mixtures of Multinomials.
- Mixtures of Gaussians, Mixtures of Experts, Hidden Markov Models (HMM).
- Sigmoidal Belief Networks, Bayesian Networks, Markov Random Fields.

Therefore

Generative Methods

1. Model class-conditional pdfs and prior probabilities.
2. "Generative" since sampling can generate synthetic data points.

Examples

- Gaussians, Naïve Bayes, Mixtures of Multinomials.
- Mixtures of Gaussians, Mixtures of Experts, Hidden Markov Models (HMM).
- Sigmoidal Belief Networks, Bayesian Networks, Markov Random Fields.

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

- Traditional neural networks, Nearest neighbor
- Conditional Random Fields (CRF)

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

Popular models

- Logistic regression, SVMs.
- Traditional neural networks, Nearest neighbor.
- Conditional Random Fields (CRF).

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

Popular models

- Logistic regression, SVMs.
- Traditional neural networks, Nearest neighbor.
- Conditional Random Fields (CRF).

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

Popular models

- Logistic regression, SVMs.
- Traditional neural networks, Nearest neighbor.
- Conditional Random Fields (CRF).

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

Popular models

- Logistic regression, SVMs.
- Traditional neural networks, Nearest neighbor.
- Conditional Random Fields (CRF).

Furthermore

Discriminative Methods

1. Directly estimate posterior probabilities.
2. No attempt to model underlying probability distributions.
3. Focus computational resources on given task for better performance.

Popular models

- Logistic regression, SVMs.
- Traditional neural networks, Nearest neighbor.
- Conditional Random Fields (CRF).

Outline

Naive Bayes Model

Task for two classes

Let ω_1, ω_2 be the two classes in which our samples belong.

Naive Bayes Model

Task for two classes

Let ω_1, ω_2 be the two classes in which our samples belong.

There is a prior probability of belonging to that class

- $P(\omega_1)$ for Class 1.
- $P(\omega_2)$ for Class 2.

Naive Bayes Model

Task for two classes

Let ω_1, ω_2 be the two classes in which our samples belong.

There is a prior probability of belonging to that class

- $P(\omega_1)$ for Class 1.
- $P(\omega_2)$ for Class 2.

The Rule for classification is the following one

$$P(\omega_i|x) = \frac{P(x|\omega_i) P(\omega_i)}{P(x)} \tag{1}$$

- Remark: Bayes to the next level.

Naive Bayes Model

Task for two classes

Let ω_1, ω_2 be the two classes in which our samples belong.

There is a prior probability of belonging to that class

- $P(\omega_1)$ for Class 1.
- $P(\omega_2)$ for Class 2.

The Rule for classification is the following one

$$P(\omega_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_i)\,P(\omega_i)}{P(\boldsymbol{x})} \tag{1}$$

Remark: Bayes to the next level.

In Informal English

We have that

$$posterior = \frac{likelihood \times prior\text{-}information}{evidence} \quad (2)$$

In Informal English

We have that

$$posterior = \frac{likelihood \times prior\text{-}information}{evidence} \quad (2)$$

Basically

One: If we can observe x.

Two: we can convert the prior-information into the posterior information.

In Informal English

We have that

$$posterior = \frac{likelihood \times prior\text{-}information}{evidence} \quad (2)$$

Basically

One: If we can observe $\boldsymbol{x}$.

Two: we can convert the prior-information into the posterior information.

We have the following terms...

Likelihood

We call $p\left(\boldsymbol{x}|\omega_i\right)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p\left(\boldsymbol{x}|\omega_i\right)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

We have the following terms...

Likelihood

We call $p\left(\boldsymbol{x}|\omega_i\right)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p\left(\boldsymbol{x}|\omega_i\right)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

Prior Probability

It is the known probability of a given class.

Remark: Because, we lack information about this class, we tend to use the uniform distribution.

However: We can use other tricks for it.

We have the following terms...

Likelihood

We call $p\left(\boldsymbol{x}|\omega_i\right)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p\left(\boldsymbol{x}|\omega_i\right)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

Prior Probability

It is the known probability of a given class.

Remark: Because, we lack information about this class, we tend to use the uniform distribution.

However: We can use other tricks for it.

Evidence

The evidence factor can be seen as a scale factor that guarantees that the posterior probability sum to one.

We have the following terms...

Likelihood

We call $p(\boldsymbol{x}|\omega_i)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p(\boldsymbol{x}|\omega_i)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

Prior Probability

It is the known probability of a given class.

Remark: Because, we lack information about this class, we tend to use the uniform distribution.

However: We can use other tricks for it.

Evidence

The evidence factor can be seen as a scale factor that guarantees that the posterior probability sum to one.

We have the following terms...

Likelihood

We call $p(\boldsymbol{x}|\omega_i)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p(\boldsymbol{x}|\omega_i)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

Prior Probability

It is the known probability of a given class.

Remark: Because, we lack information about this class, we tend to use the uniform distribution.

However: We can use other tricks for it.

Evidence

The evidence factor can be seen as a scale factor that guarantees that the posterior probability sum to one.

We have the following terms...

Likelihood

We call $p(\boldsymbol{x}|\omega_i)$ the likelihood of ω_i given $\boldsymbol{x}$:

- This indicates that given a category ω_i: If $p(\boldsymbol{x}|\omega_i)$ is "large", then ω_i is the "likely" class of $\boldsymbol{x}$.

Prior Probability

It is the known probability of a given class.

Remark: Because, we lack information about this class, we tend to use the uniform distribution.

However: We can use other tricks for it.

Evidence

The evidence factor can be seen as a scale factor that guarantees that the posterior probability sum to one.

The most important term in all this

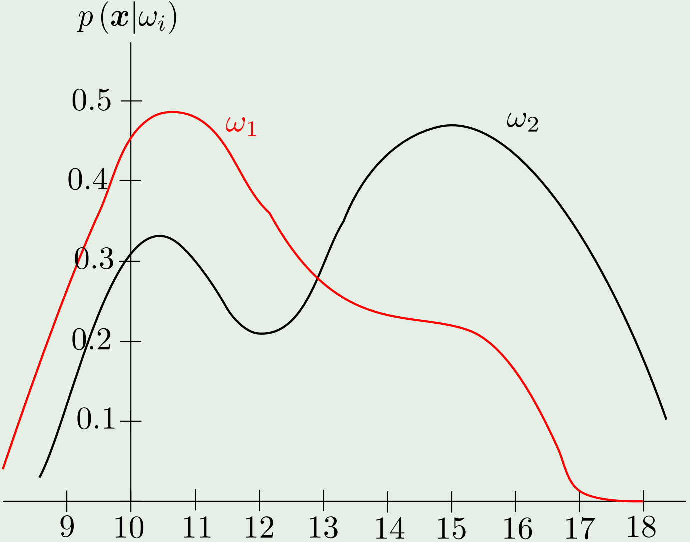
The factor

$$likelihood \times prior\text{-}information \tag{3}$$

Outline

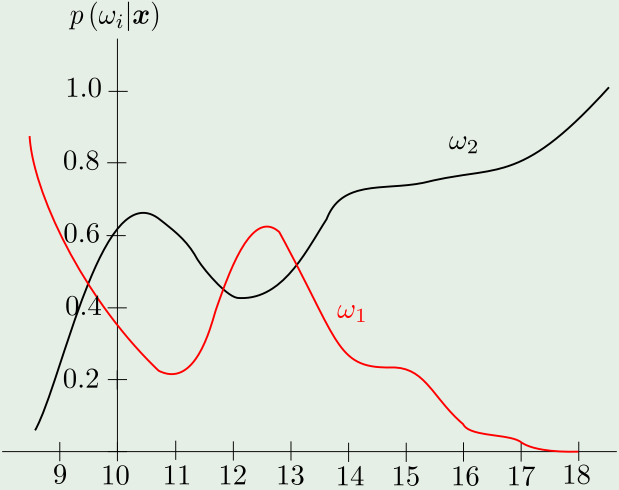
Example

We have the likelihood of two classes



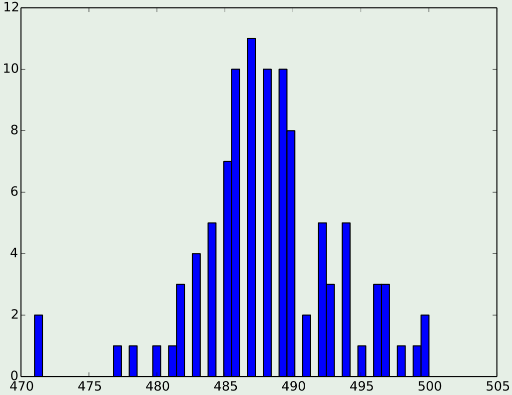
Example

We have the posterior of two classes when $P(\omega_1) = \frac{2}{3}$ and $P(\omega_2) = \frac{1}{3}$



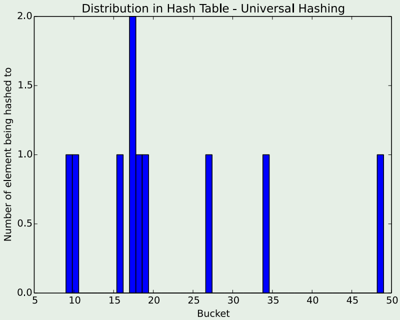
Example of key distribution

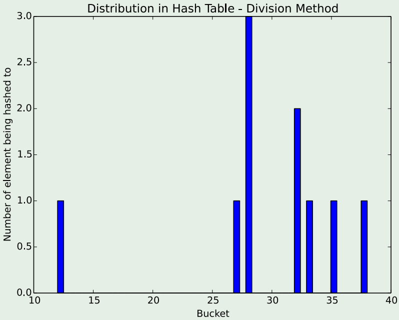
Example, mean = 488.5 and dispersion = 5



Example with 10 keys

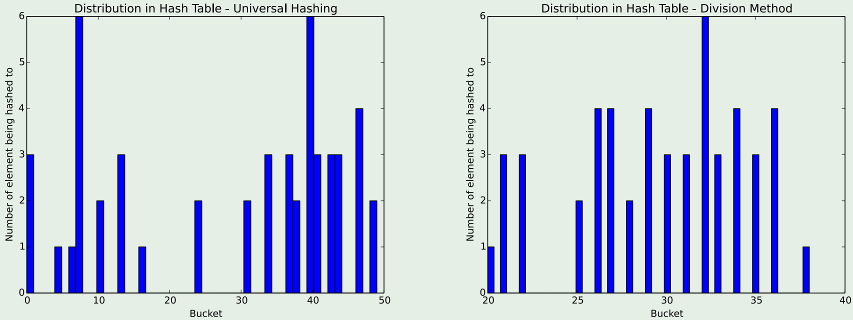
Universal Hashing Vs Division Method





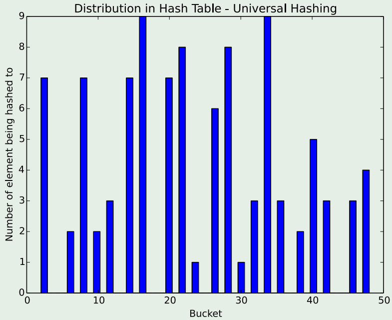
Example with 50 keys

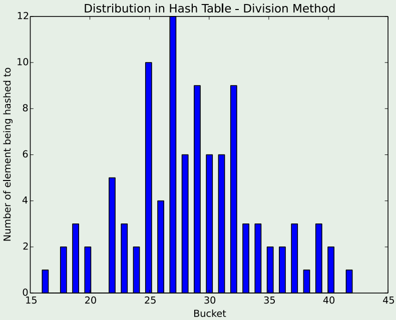
Universal Hashing Vs Division Method



Example with 100 keys

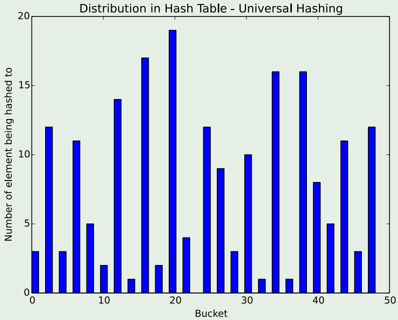
Universal Hashing Vs Division Method

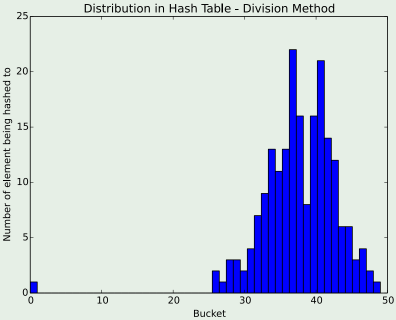




Example with 200 keys

Universal Hashing Vs Division Method





Outline

Naive Bayes Model

In the case of two classes, we can use demarginalization

$$P(\boldsymbol{x}) = \sum_{i=1}^{2} p(\boldsymbol{x}, \omega_i) = \sum_{i=1}^{2} p(\boldsymbol{x}|\omega_i) P(\omega_i) \tag{4}$$

Error in this rule

We have that

$$P\left(error|\boldsymbol{x}\right)=\begin{cases}P\left(\omega_1|\boldsymbol{x}\right) & \text{if we decide } \omega_2\\ P\left(\omega_2|\boldsymbol{x}\right) & \text{if we decide } \omega_1\end{cases} \quad (5)$$

Thus, we have that

$$P\left(error\right)=\int_{-\infty}^{\infty}P\left(error,\boldsymbol{x}\right)d\boldsymbol{x}=\int_{-\infty}^{\infty}P\left(error|\boldsymbol{x}\right)p\left(\boldsymbol{x}\right)d\boldsymbol{x} \quad (6)$$

Error in this rule

We have that

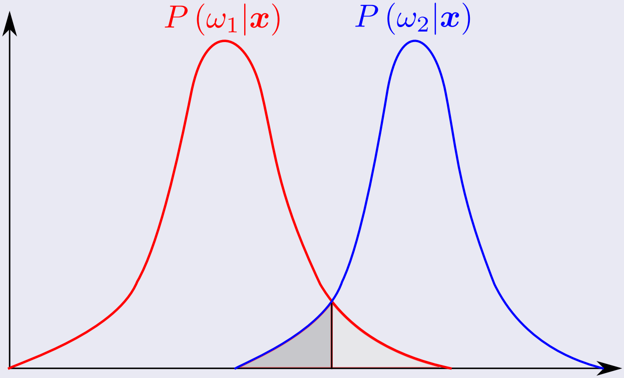
$$P\left(error|\boldsymbol{x}\right)=\begin{cases}P\left(\omega_1|\boldsymbol{x}\right) & \text{if we decide } \omega_2\\ P\left(\omega_2|\boldsymbol{x}\right) & \text{if we decide } \omega_1\end{cases} \tag{5}$$

Thus, we have that

$$P\left(error\right)=\int_{-\infty}^{\infty}P\left(error,\boldsymbol{x}\right)d\boldsymbol{x}=\int_{-\infty}^{\infty}P\left(error|\boldsymbol{x}\right)p\left(\boldsymbol{x}\right)d\boldsymbol{x} \tag{6}$$

Graphically

We have



$$P(error) = \int_{-\infty}^{\infty} P(error, \boldsymbol{x})\, d\boldsymbol{x}$$

Classification Rule

Thus, we have the Bayes Classification Rule

1. If $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$ $\boldsymbol{x}$ is classified to ω_1
2. If $P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x})$ $\boldsymbol{x}$ is classified to ω_2

Classification Rule

Thus, we have the Bayes Classification Rule

1. If $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$ $\boldsymbol{x}$ is classified to ω_1
2. If $P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x})$ $\boldsymbol{x}$ is classified to ω_2

What if we remove the normalization factor?

Remember

$$P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x}) = 1 \tag{7}$$

What if we remove the normalization factor?

Remember

$$P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x}) = 1 \tag{7}$$

We are able to obtain the new Bayes Classification Rule

1. If $P(\boldsymbol{x}|\omega_1)\, p(\omega_1) > P(\boldsymbol{x}|\omega_2)\, P(\omega_2)$ $\boldsymbol{x}$ is classified to ω_1
2. If $P(\boldsymbol{x}|\omega_1)\, p(\omega_1) < P(\boldsymbol{x}|\omega_2)\, P(\omega_2)$ $\boldsymbol{x}$ is classified to ω_2

What if we remove the normalization factor?

Remember

$$P\left(\omega_1|\boldsymbol{x}\right)+P\left(\omega_2|\boldsymbol{x}\right)=1 \tag{7}$$

We are able to obtain the new Bayes Classification Rule

1. If $P\left(\boldsymbol{x}|\omega_1\right)p\left(\omega_1\right)>P\left(\boldsymbol{x}|\omega_2\right)P\left(\omega_2\right)$ $\boldsymbol{x}$ is classified to ω_1
2. If $P\left(\boldsymbol{x}|\omega_1\right)p\left(\omega_1\right)<P\left(\boldsymbol{x}|\omega_2\right)P\left(\omega_2\right)$ $\boldsymbol{x}$ is classified to ω_2

We have several cases

If for some $\boldsymbol{x}$ we have $P(\boldsymbol{x}|\omega_1) = P(\boldsymbol{x}|\omega_2)$

The final decision relies completely from the prior probability.

On the Other hand if $P(\omega_1) = P(\omega_2)$, the "state" is equally probable

In this case the decision is based entirely on the likelihoods $P(\boldsymbol{x}|\omega_i)$.

We have several cases

If for some $\boldsymbol{x}$ we have $P(\boldsymbol{x}|\omega_1) = P(\boldsymbol{x}|\omega_2)$

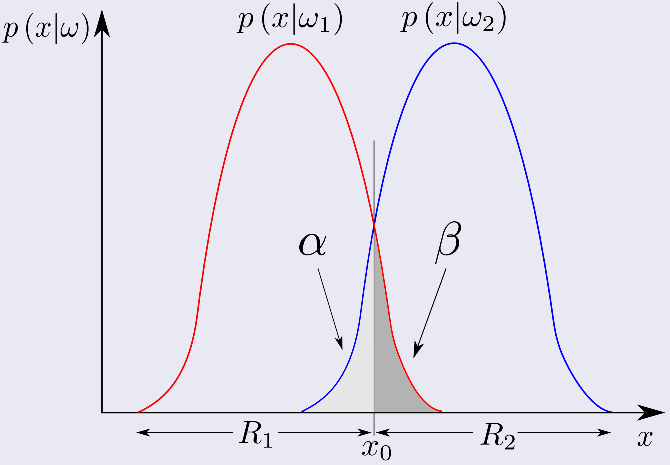
The final decision relies completely from the prior probability.

On the Other hand if $P(\omega_1) = P(\omega_2)$, the "state" is equally probable

In this case the decision is based entirely on the likelihoods $P(\boldsymbol{x}|\omega_i)$.

How the Rule looks like

If $P(\omega_1) = P(\omega_2)$ the Rule depends on the term $p(\boldsymbol{x}|\omega_i)$



Error in Naive Bayes

Error in equiprobable classes $p\left(\omega_1\right) = p\left(\omega_2\right) = \frac{1}{2}$

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

Error in Naive Bayes

Error in equiprobable classes $p\left(\omega_1\right) = p\left(\omega_2\right) = \frac{1}{2}$

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

Error in Naive Bayes

Error in equiprobable classes $p\left(\omega_1\right) = p\left(\omega_2\right) = \frac{1}{2}$

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

Error in Naive Bayes

Error in equiprobable classes $p(\omega_1) = p(\omega_2) = \frac{1}{2}$

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

Error in Naive Bayes

Error in equiprobable classes $p(\omega_1) = p(\omega_2) = \frac{1}{2}$

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

Error in Naive Bayes

Something Notable

Bayesian classifier is optimal with respect to minimizing the classification error probability.

Proof

Step 1

- R_1 be the region of the feature space in which we decide in favor of ω_1
- R_2 be the region of the feature space in which we decide in favor of ω_2

Proof

Step 1

- R_1 be the region of the feature space in which we decide in favor of ω_1
- R_2 be the region of the feature space in which we decide in favor of ω_2

Step 2

$$P_e = P\left(x \in R_2, \omega_1\right) + P\left(x \in R_1, \omega_2\right) \tag{8}$$

Proof

Step 1

- R_1 be the region of the feature space in which we decide in favor of ω_1
- R_2 be the region of the feature space in which we decide in favor of ω_2

Step 2

$$P_e = P\left(x \in R_2, \omega_1\right) + P\left(x \in R_1, \omega_2\right) \tag{8}$$

Thus

$$\begin{aligned} P_e &= P\left(x \in R_2 | \omega_1\right) P\left(\omega_1\right) + P\left(x \in R_1 | \omega_2\right) P\left(\omega_2\right) \\ &= P\left(\omega_1\right) \int_{R_2} p\left(x | \omega_1\right) dx + P\left(\omega_2\right) \int_{R_1} p\left(x | \omega_2\right) dx \end{aligned}$$

Proof

Step 1

- R_1 be the region of the feature space in which we decide in favor of ω_1
- R_2 be the region of the feature space in which we decide in favor of ω_2

Step 2

$$P_e = P\left(x \in R_2, \omega_1\right) + P\left(x \in R_1, \omega_2\right) \tag{8}$$

Thus

$$\begin{aligned} P_e &= P\left(x \in R_2 | \omega_1\right) P\left(\omega_1\right) + P\left(x \in R_1 | \omega_2\right) P\left(\omega_2\right) \\ &= P\left(\omega_1\right) \int_{R_2} p\left(x|\omega_1\right) dx + P\left(\omega_2\right) \int_{R_1} p\left(x|\omega_2\right) dx \end{aligned}$$

Proof

Step 1

- R_1 be the region of the feature space in which we decide in favor of ω_1
- R_2 be the region of the feature space in which we decide in favor of ω_2

Step 2

$$P_e = P\left(x \in R_2, \omega_1\right) + P\left(x \in R_1, \omega_2\right) \tag{8}$$

Thus

$$\begin{aligned} P_e &= P\left(x \in R_2 | \omega_1\right) P\left(\omega_1\right) + P\left(x \in R_1 | \omega_2\right) P\left(\omega_2\right) \\ &= P\left(\omega_1\right) \int\limits_{R_2} p\left(x|\omega_1\right) dx + P\left(\omega_2\right) \int\limits_{R_1} p\left(x|\omega_2\right) dx \end{aligned}$$

Proof

It is more

$$P_e = P(\omega_1) \int\limits_{R_2} \frac{p(\omega_1, x)}{P(\omega_1)} dx + P(\omega_2) \int\limits_{R_1} \frac{p(\omega_2, x)}{P(\omega_2)} dx \tag{9}$$

Finally

$$P_e = \int\limits_{R_2} p(\omega_1|x)\, p(x)\, dx + \int\limits_{R_1} p(\omega_2|x)\, p(x)\, dx \tag{10}$$

Now, we choose the Bayes Classification Rule

$$\begin{aligned} R_1 &: P(\omega_1|x) > P(\omega_2|x) \\ R_2 &: P(\omega_2|x) > P(\omega_1|x) \end{aligned}$$

Proof

It is more

$$P_e = P(\omega_1)\int\limits_{R_2}\frac{p(\omega_1, x)}{P(\omega_1)}dx + P(\omega_2)\int\limits_{R_1}\frac{p(\omega_2, x)}{P(\omega_2)}dx \qquad (9)$$

Finally

$$P_e = \int\limits_{R_2} p(\omega_1|x)\, p(x)\, dx + \int\limits_{R_1} p(\omega_2|x)\, p(x)\, dx \qquad (10)$$

Now, we choose the Bayes Classification Rule

$$\begin{aligned} R_1 &: P(\omega_1|x) > P(\omega_2|x) \\ R_2 &: P(\omega_2|x) > P(\omega_1|x) \end{aligned}$$

Proof

It is more

$$P_e = P(\omega_1) \int\limits_{R_2} \frac{p(\omega_1, x)}{P(\omega_1)} dx + P(\omega_2) \int\limits_{R_1} \frac{p(\omega_2, x)}{P(\omega_2)} dx \tag{9}$$

Finally

$$P_e = \int\limits_{R_2} p(\omega_1|x)\, p(x)\, dx + \int\limits_{R_1} p(\omega_2|x)\, p(x)\, dx \tag{10}$$

Now, we choose the Bayes Classification Rule

$$\begin{aligned} R_1 &: \quad P(\omega_1|x) > P(\omega_2|x) \\ R_2 &: \quad P(\omega_2|x) > P(\omega_1|x) \end{aligned}$$

Proof

Thus

$$P(\omega_1) = \int\limits_{R_1} p(\omega_1|x)\,p(x)\,dx + \int\limits_{R_2} p(\omega_1|x)\,p(x)\,dx \tag{11}$$

Now, we have...

$$P(\omega_1) - \int\limits_{R_1} p(\omega_1|x)\,p(x)\,dx = \int\limits_{R_2} p(\omega_1|x)\,p(x)\,dx \tag{12}$$

Then

$$P_e = P(\omega_1) - \int\limits_{R_1} p(\omega_1|x)\,p(x)\,dx + \int\limits_{R_1} p(\omega_2|x)\,p(x)\,dx \tag{13}$$

Proof

Thus

$$P(\omega_1) = \int_{R_1} p(\omega_1|x)\, p(x)\, dx + \int_{R_2} p(\omega_1|x)\, p(x)\, dx \tag{11}$$

Now, we have...

$$P(\omega_1) - \int_{R_1} p(\omega_1|x)\, p(x)\, dx = \int_{R_2} p(\omega_1|x)\, p(x)\, dx \tag{12}$$

Then

$$P_e = P(\omega_1) - \int_{R_1} p(\omega_1|x)\, p(x)\, dx + \int_{R_1} p(\omega_2|x)\, p(x)\, dx \tag{13}$$

Proof

Thus

$$P(\omega_1) = \int\limits_{R_1} p(\omega_1|x)\, p(x)\, dx + \int\limits_{R_2} p(\omega_1|x)\, p(x)\, dx \tag{11}$$

Now, we have...

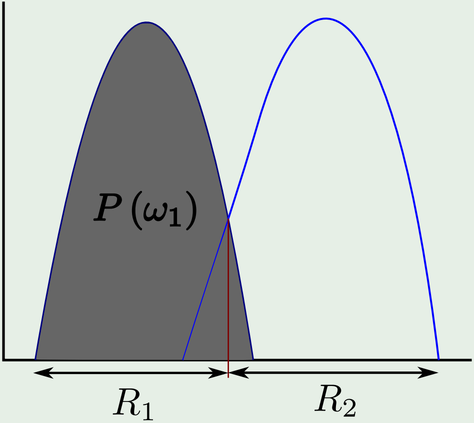
$$P(\omega_1) - \int\limits_{R_1} p(\omega_1|x)\, p(x)\, dx = \int\limits_{R_2} p(\omega_1|x)\, p(x)\, dx \tag{12}$$

Then

$$P_e = P(\omega_1) - \int\limits_{R_1} p(\omega_1|x)\, p(x)\, dx + \int\limits_{R_1} p(\omega_2|x)\, p(x)\, dx \tag{13}$$

Graphically $P(\omega_1)$: Thanks Edith 2013 Class!!!

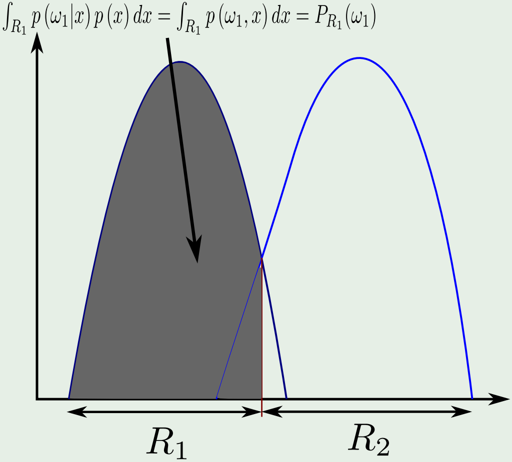
In Gray



Thus we have

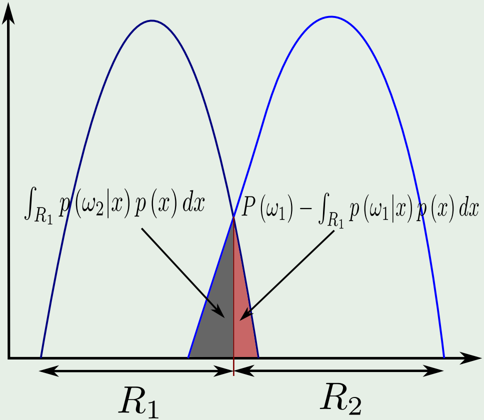
$$\int_{R_1} p(\omega_1|x)\, p(x)\, dx = \int_{R_1} p(\omega_1, x)\, dx = P_{R_1}(\omega_1)$$

Thus



Finally P_e

A great idea Edith!!!



Thus

Finally

$$P_e = P(\omega_1) - \int_{R_1} \left[p(\omega_1|x) - p(\omega_2|x) \right] p(x)\, dx \tag{14}$$

Thus

The probability of error is minimized at the region of space in which $R_1 : P(\omega_1|x) > P(\omega_2|x)$.

Thus

Finally

$$P_e = P(\omega_1) - \int_{R_1} \left[p(\omega_1|x) - p(\omega_2|x) \right] p(x)\, dx \tag{14}$$

Thus

The probability of error is minimized at the region of space in which $R_1 : P(\omega_1|x) > P(\omega_2|x)$.

Finally

Similarly

$$P_e = P(\omega_2) - \int_{R_2} \left[p(\omega_2|x) - p(\omega_1|x)\right] p(x)\, dx \tag{15}$$

Thus

The probability of error is minimized at the region of space in which $R_2 : P(\omega_2|x) > P(\omega_1|x)$.

Thus

The Naive Bayes Rule minimizes the error.

Finally

Similarly

$$P_e = P(\omega_2) - \int_{R_2} \left[p(\omega_2|x) - p(\omega_1|x)\right] p(x)\, dx \tag{15}$$

Thus

The probability of error is minimized at the region of space in which $R_2 : P(\omega_2|x) > P(\omega_1|x)$.

Thus

The Naive Bayes Rule minimizes the error.

Finally

Similarly

$$P_e = P(\omega_2) - \int_{R_2} \left[p(\omega_2|x) - p(\omega_1|x)\right] p(x)\, dx \tag{15}$$

Thus

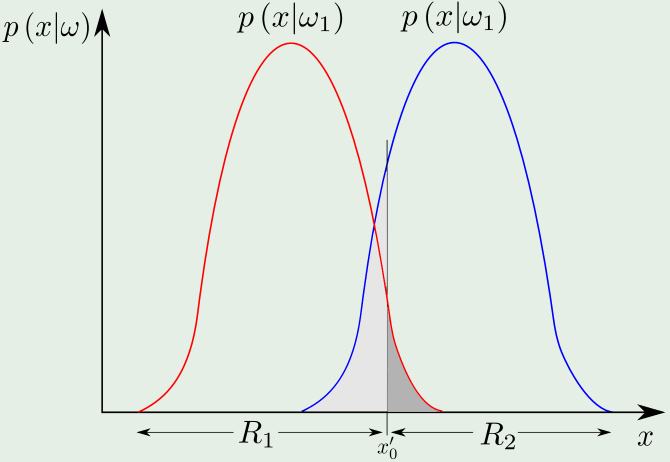
The probability of error is minimized at the region of space in which $R_2 : P(\omega_2|x) > P(\omega_1|x)$.

Thus

The Naive Bayes Rule minimizes the error.

After all!!!

If you choose any other x_0'



Outline

For M classes $\omega_1, \omega_2, ..., \omega_M$

We have that vector $\boldsymbol{x}$ is in ω_i

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \;\; \forall j \neq i \tag{16}$$

Something Notable

It turns out that such a choice also minimizes the classification error probability.

For M classes $\omega_1, \omega_2, ..., \omega_M$

We have that vector $\boldsymbol{x}$ is in ω_i

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \;\; \forall j \neq i \tag{16}$$

Something Notable

It turns out that such a choice also minimizes the classification error probability.

Outline

Decision Surface

Because the R_1 and R_2 are contiguous

The separating surface between both of them is described by

$$P\left(\omega_1|x\right) - P\left(\omega_2|x\right) = 0 \tag{17}$$

Thus, we define the decision function as

$$g_{12}\left(x\right) = P\left(\omega_1|x\right) - P\left(\omega_2|x\right) = 0 \tag{18}$$

Decision Surface

Because the R_1 and R_2 are contiguous

The separating surface between both of them is described by

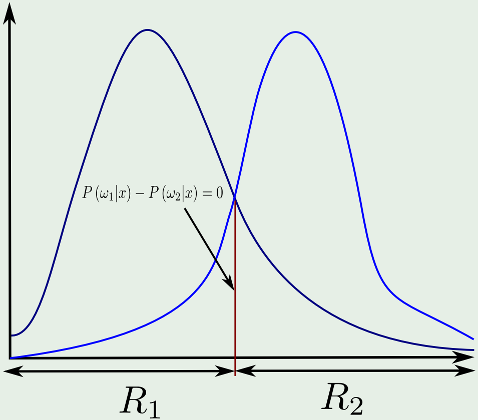
$$P(\omega_1|x) - P(\omega_2|x) = 0 \tag{17}$$

Thus, we define the decision function as

$$g_{12}(x) = P(\omega_1|x) - P(\omega_2|x) = 0 \tag{18}$$

Which decision function for the Naive Bayes

A single number in this case



In general

First

Instead of working with probabilities, we work with an equivalent function of them $g_i(\boldsymbol{x}) = f(P(\omega_i|\boldsymbol{x}))$.

- Classic Example the Monotonically increasing $f(P(\omega_i|\boldsymbol{x})) = \ln P(\omega_i|\boldsymbol{x})$.

In general

First

Instead of working with probabilities, we work with an equivalent function of them $g_i(\boldsymbol{x}) = f(P(\omega_i|\boldsymbol{x}))$.

- Classic Example the Monotonically increasing $f(P(\omega_i|\boldsymbol{x})) = \ln P(\omega_i|\boldsymbol{x})$.

The decision test is now

$$\text{classify } x \text{ in } \omega_i \text{ if } g_i(x) > g_j(x) \ \forall j \neq i.$$

In general

First

Instead of working with probabilities, we work with an equivalent function of them $g_i(\boldsymbol{x}) = f(P(\omega_i|\boldsymbol{x}))$.

- Classic Example the Monotonically increasing $f(P(\omega_i|\boldsymbol{x})) = \ln P(\omega_i|\boldsymbol{x})$.

The decision test is now

classify $\boldsymbol{x}$ in ω_i if $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})\ \forall j \neq i$.

The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(\boldsymbol{x}) = g_i(\boldsymbol{x}) - g_j(\boldsymbol{x})\ i,j = 1,2,...,M\ i \neq j$$

In general

First

Instead of working with probabilities, we work with an equivalent function of them $g_i(\boldsymbol{x}) = f(P(\omega_i|\boldsymbol{x}))$.

- Classic Example the Monotonically increasing $f(P(\omega_i|\boldsymbol{x})) = \ln P(\omega_i|\boldsymbol{x})$.

The decision test is now

$$\text{classify } \boldsymbol{x} \text{ in } \omega_i \text{ if } g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \ \forall j \neq i.$$

The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(\boldsymbol{x}) = g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) \ i,j = 1,2,...,M \ i \neq j$$

Outline

Gaussian Distribution

We can use the Gaussian distribution

$$p\left(\boldsymbol{x}|\boldsymbol{\omega_i}\right)=\frac{1}{\left(2\pi\right)^{l/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu_i}\right)^T\Sigma_i^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu_i}\right)\right\} \tag{19}$$

Example

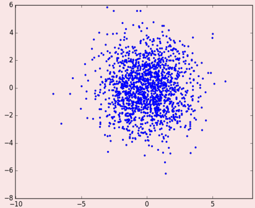
Gaussian Distribution

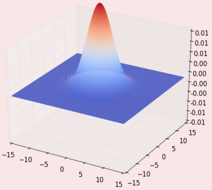
We can use the Gaussian distribution

$$p\left(\boldsymbol{x}|\boldsymbol{\omega_i}\right) = \frac{1}{\left(2\pi\right)^{l/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu_i}\right)^T \Sigma_i^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu_i}\right)\right\} \quad (19)$$

Example

$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$





Some Properties

About Σ

It is the covariance matrix between variables.

Thus

- It is positive semi-definite.
- Symmetric.
- The inverse exists.

Some Properties

About Σ

It is the covariance matrix between variables.

Thus

- It is positive semi-definite.
- Symmetric.
- The inverse exists.

Outline

Influence of the Covariance Σ

Look at the following Covariance

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

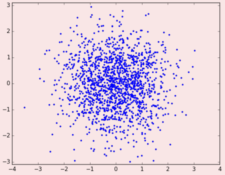
It simple the unit Gaussian with mean μ

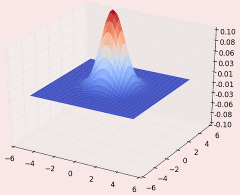
Influence of the Covariance Σ

Look at the following Covariance

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

It simple the unit Gaussian with mean μ





The Covariance Σ as a Rotation

Look at the following Covariance

$$\Sigma = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix}$$

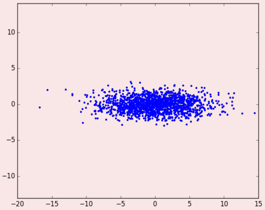
Actually, it flatten the circle through the $x - axis$

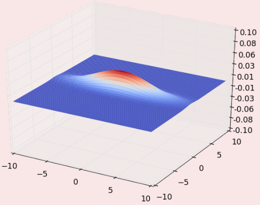
The Covariance Σ as a Rotation

Look at the following Covariance

$$\Sigma = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix}$$

Actually, it flatten the circle through the $x - axis$





Influence of the Covariance Σ

Look at the following Covariance

$$\Sigma_a = R \Sigma_b R^T \text{ with } R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

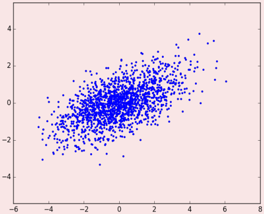
It allows to rotate the axises

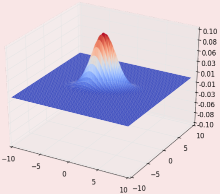
Influence of the Covariance Σ

Look at the following Covariance

$$\Sigma_a = R\Sigma_b R^T \text{ with } R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

It allows to rotate the axises





Now For Two Classes

Then, we use the following trick for two Classes $i = 1, 2$

We know that the pdf of correct classification is
$p\left(x, \omega_1\right) = p\left(x|\omega_i\right) P\left(\omega_i\right)$!!!

Thus

It is possible to generate the following decision function:

$$g_i\left(x\right) = \ln\left[p\left(x|\omega_i\right) P\left(\omega_i\right)\right] = \ln p\left(x|\omega_i\right) + \ln P\left(\omega_i\right) \tag{20}$$

Thus

$$g_i\left(x\right) = -\frac{1}{2}\left(x - \mu_i\right)^T \Sigma_i^{-1}\left(x - \mu_i\right) + \ln P\left(\omega_i\right) + c_i \tag{21}$$

Now For Two Classes

Then, we use the following trick for two Classes $i = 1, 2$

We know that the pdf of correct classification is
$p(x, \omega_1) = p(x|\omega_i) P(\omega_i)$!!!

Thus

It is possible to generate the following decision function:

$$g_i(\boldsymbol{x}) = \ln\left[p(x|\omega_i) P(\omega_i)\right] = \ln p(x|\omega_i) + \ln P(\omega_i) \tag{20}$$

Thus

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i) + c_i \tag{21}$$

Now For Two Classes

Then, we use the following trick for two Classes $i = 1, 2$

We know that the pdf of correct classification is
$p(x, \omega_1) = p(x|\omega_i) P(\omega_i)$!!!

Thus

It is possible to generate the following decision function:

$$g_i(\boldsymbol{x}) = \ln\left[p(x|\omega_i) P(\omega_i)\right] = \ln p(x|\omega_i) + \ln P(\omega_i) \tag{20}$$

Thus

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) + \ln P(\omega_i) + c_i \tag{21}$$

Outline

We can work one of the possible decision surfaces

Assume first that $\Sigma_i = \sigma^2 I$

- The features are statistically independent
- Each feature has the same variance

We can work one of the possible decision surfaces

Assume first that $\Sigma_i = \sigma^2 I$

- The features are statistically independent
- Each feature has the same variance

Therefore

- The samples fall in equal size spherical clusters!!!
- Each Cluster centered at mean vector μ_i.

We can work one of the possible decision surfaces

Assume first that $\Sigma_i = \sigma^2 I$

- The features are statistically independent
- Each feature has the same variance

Therefore

- The samples fall in equal size spherical clusters!!!
- Each Cluster centered at mean vector μ_i.

We can work one of the possible decision surfaces

Assume first that $\Sigma_i = \sigma^2 I$

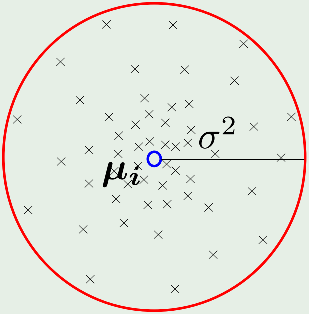
- The features are statistically independent
- Each feature has the same variance

Therefore

- The samples fall in equal size spherical clusters!!!
- Each Cluster centered at mean vector μ_i.

For Example

We have



Now

We have that

$$|\Sigma_i| = \sigma^{2d} \text{ and } \Sigma_i^{-1} = \left(\frac{1}{\sigma^2}\right) I$$

Something Notable

- Gaussian Multivariate function after the log

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \ln P(\omega_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$$

The term $-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$

It is unimportant therefore it can be ignored!!!

Now

We have that

$$|\Sigma_i| = \sigma^{2d} \text{ and } \Sigma_i^{-1} = \left(\frac{1}{\sigma^2}\right) I$$

Something Notable

- Gaussian Multivariate function after the $\log$

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^T \Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu_i}) + \ln P(\omega_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$$

The term $-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$

It is unimportant therefore it can be ignored!!!

Now

We have that

$$|\Sigma_i| = \sigma^{2d} \text{ and } \Sigma_i^{-1} = \left(\frac{1}{\sigma^2}\right) I$$

Something Notable

- Gaussian Multivariate function after the $\log$

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) + \ln P(\omega_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$$

The term $-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i|$

It is unimportant therefore it can be ignored!!!

Then

We have the following discriminant functions

$$g_i(\boldsymbol{x}) = -\frac{\overbrace{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T(\boldsymbol{x}-\boldsymbol{\mu}_i)}^{\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2}}{2\sigma^2} + \ln P(\omega_i) \tag{22}$$

Then, we have that

$$g_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\left[\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] + \ln P(\omega_i)$$

Then

We have the following discriminant functions

$$g_i(\boldsymbol{x}) = -\frac{\overbrace{(\boldsymbol{x}-\boldsymbol{\mu_i})^T(\boldsymbol{x}-\boldsymbol{\mu_i})}^{\|\boldsymbol{x}-\boldsymbol{\mu_i}\|^2}}{2\sigma^2} + \ln P(\omega_i) \tag{22}$$

Then, we have that

$$g_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\left[\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{\mu_i}^T\boldsymbol{x} + \boldsymbol{\mu_i}^T\boldsymbol{\mu_i}\right] + \ln P(\omega_i)$$

We can then...

Do you notice that $\boldsymbol{x}^T\boldsymbol{x}$ is actually the same for all g_i?

Then, we can ignore that term thus, we get

$$g_i(\boldsymbol{x}) = \underbrace{\frac{1}{\sigma^2}\boldsymbol{\mu_i}}_{\boldsymbol{w}_i^T}{}^T\boldsymbol{x} - \frac{1}{2\sigma^2}\boldsymbol{\mu_i}^T\underbrace{\boldsymbol{\mu_i}}_{w_{i0}} + \ln P(\omega_i)$$

Or if you want

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

We can then...

Do you notice that $\boldsymbol{x}^T\boldsymbol{x}$ is actually the same for all g_i?

Then, we can ignore that term thus, we get

$$g_i(\boldsymbol{x}) = \overbrace{\frac{1}{\sigma^2}\boldsymbol{\mu_i}^T}^{\boldsymbol{w}_i^T}\boldsymbol{x} \overbrace{- \frac{1}{2\sigma^2}\boldsymbol{\mu_i}^T\boldsymbol{\mu_i} + \ln P(\omega_i)}^{w_{i0}}$$

Or if you want

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$$

Outline

Given a series of classes $\omega_1, \omega_2, ..., \omega_M$

We assume for each class ω_j

The samples are drawn independently according to the probability law $p(\boldsymbol{x}|\omega_j)$

We call those samples as

i.i.d. — independent identically distributed random variables.

We assume in addition

$p(\boldsymbol{x}|\omega_j)$ has a known parametric form with vector θ_j of parameters.

Given a series of classes $\omega_1, \omega_2, ..., \omega_M$

We assume for each class ω_j

The samples are drawn independently according to the probability law $p(\boldsymbol{x}|\omega_j)$

We call those samples as

i.i.d. — independent identically distributed random variables.

We assume in addition

$p(\boldsymbol{x}|\omega_j)$ has a known parametric form with vector θ_j of parameters.

Given a series of classes $\omega_1, \omega_2, ..., \omega_M$

We assume for each class ω_j

The samples are drawn independently according to the probability law $p(\boldsymbol{x}|\omega_j)$

We call those samples as

i.i.d. — independent identically distributed random variables.

We assume in addition

$p(\boldsymbol{x}|\omega_j)$ has a known parametric form with vector $\boldsymbol{\theta}_j$ of parameters.

Given a series of classes $\omega_1, \omega_2, ..., \omega_M$

For example

$$p(\boldsymbol{x}|\omega_j) \sim N\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{23}$$

In our case

We will assume that there is no dependence between classes!!!

Given a series of classes $\omega_1, \omega_2, ..., \omega_M$

For example

$$p\left(\boldsymbol{x}|\omega_j\right) \sim N\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{23}$$

In our case

We will assume that there is no dependence between classes!!!

Now

Suppose that ω_j contains n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$

$$p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n|\boldsymbol{\theta}_j\right) = \prod_{j=1}^{n} p\left(\boldsymbol{x}_j|\boldsymbol{\theta}_j\right) \tag{24}$$

We can see then the function $p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n|\boldsymbol{\theta}_j\right)$ as a function of

$$L\left(\boldsymbol{\theta}_j\right) = \prod_{j=1}^{n} p\left(\boldsymbol{x}_j|\boldsymbol{\theta}_j\right) \tag{25}$$

Now

Suppose that ω_j contains n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$

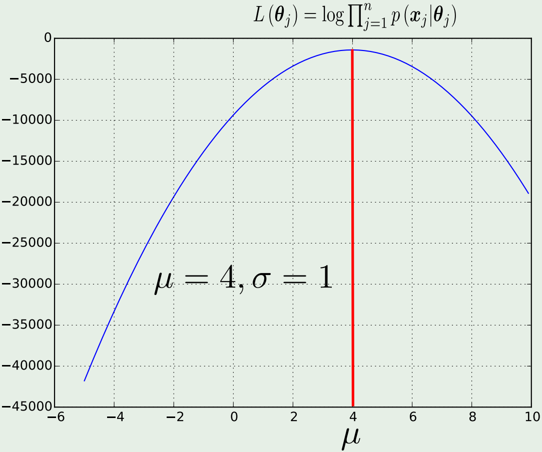
$$p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n | \boldsymbol{\theta}_j\right) = \prod_{j=1}^{n} p\left(\boldsymbol{x}_j | \boldsymbol{\theta}_j\right) \tag{24}$$

We can see then the function $p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n | \boldsymbol{\theta}_j\right)$ as a function of

$$L\left(\boldsymbol{\theta}_j\right) = \prod_{j=1}^{n} p\left(\boldsymbol{x}_j | \boldsymbol{\theta}_j\right) \tag{25}$$

Example

$L\left(\boldsymbol{\theta}_j\right) = \log \prod_{j=1}^{n} p\left(\boldsymbol{x}_j | \boldsymbol{\theta}_j\right)$



Outline

Maximum Likelihood on a Gaussian

Then, using the log!!!

$$\ln L\left(\omega_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\left[\sum_{j=1}^{n}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\Sigma_i^{-1}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\right] + c_2 \quad (26)$$

We know that

$$\frac{dx^T Ax}{dx} = Ax + A^T x,\ \ \frac{dAx}{dx} = A \quad (27)$$

Thus, we expand equation26

$$-\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[x_j{}^T\Sigma_i^{-1}x_j - 2x_j{}^T\Sigma_i^{-1}\mu_i + \mu_i{}^T\Sigma_i^{-1}\mu_i\right] + c_2 \quad (28)$$

Maximum Likelihood on a Gaussian

Then, using the log!!!

$$\ln L\left(\omega_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\left[\sum_{j=1}^{n}\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)^T \Sigma_i^{-1}\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)\right] + c_2 \tag{26}$$

We know that

$$\frac{d\boldsymbol{x}^T A\boldsymbol{x}}{d\boldsymbol{x}} = Ax + A^T x,\ \frac{dA\boldsymbol{x}}{d\boldsymbol{x}} = A \tag{27}$$

Thus, we expand equation26

$$-\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[x_j{}^T\Sigma_i^{-1}x_j - 2x_j{}^T\Sigma_i^{-1}\mu_i + \mu_i{}^T\Sigma_i^{-1}\mu_i\right] + c_2 \tag{28}$$

Maximum Likelihood on a Gaussian

Then, using the log!!!

$$\ln L(\omega_i) = -\frac{n}{2}\ln|\Sigma_i| - \frac{1}{2}\left[\sum_{j=1}^{n}(\boldsymbol{x_j} - \boldsymbol{\mu_i})^T \Sigma_i^{-1}(\boldsymbol{x_j} - \boldsymbol{\mu_i})\right] + c_2 \quad (26)$$

We know that

$$\frac{d\boldsymbol{x}^T A\boldsymbol{x}}{d\boldsymbol{x}} = Ax + A^T x,\ \frac{dA\boldsymbol{x}}{d\boldsymbol{x}} = A \quad (27)$$

Thus, we expand equation26

$$-\frac{n}{2}\ln|\Sigma_i| - \frac{1}{2}\sum_{j=1}^{n}\left[\boldsymbol{x_j}^T\Sigma_i^{-1}\boldsymbol{x_j} - 2\boldsymbol{x_j}^T\Sigma_i^{-1}\boldsymbol{\mu_i} + \boldsymbol{\mu_i}^T\Sigma_i^{-1}\boldsymbol{\mu_i}\right] + c_2 \quad (28)$$

Maximum Likelihood

Then

$$
\begin{aligned}
\frac{\partial \ln L(\omega_i)}{\partial \boldsymbol{\mu}_i} &= \sum_{j=1}^{n} \Sigma_i^{-1} \left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right) = 0 \\
n\Sigma_i^{-1}\left[-\boldsymbol{\mu}_i + \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j\right] &= 0 \\
\hat{\boldsymbol{\mu}}_i &= \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j
\end{aligned}
$$

Maximum Likelihood

Then

$$
\begin{aligned}
\frac{\partial \ln L(\omega_i)}{\partial \boldsymbol{\mu}_i} &= \sum_{j=1}^{n} \Sigma_i^{-1} \left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right) = 0 \\
n\Sigma_i^{-1}\left[-\boldsymbol{\mu}_i + \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j\right] &= 0 \\
\hat{\boldsymbol{\mu}}_i &= \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j
\end{aligned}
$$

Maximum Likelihood

Then

$$\begin{aligned}
\frac{\partial \ln L(\omega_i)}{\partial \boldsymbol{\mu}_i} &= \sum_{j=1}^{n} \Sigma_i^{-1} \left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right) = 0 \\
n\Sigma_i^{-1} \left[-\boldsymbol{\mu}_i + \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j \right] &= 0 \\
\hat{\boldsymbol{\mu}}_i &= \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j
\end{aligned}$$

Maximum Likelihood

Then, we derive with respect to Σ_i

For this we use the following tricks:

1. $\frac{\partial \log|\Sigma|}{\partial \Sigma^{-1}} = -\frac{1}{|\Sigma|} \cdot |\Sigma| \left(\Sigma\right)^T = -\Sigma$
2. $\frac{\partial Tr[AB]}{\partial A} = \frac{\partial Tr[BA]}{\partial A} = B^T$
3. Trace(of a number)=the number
4. $Tr(A^T B) = Tr\left(BA^T\right)$

Thus

$$f\left(\Sigma_i\right) = -\frac{n}{2}\ln\left|\Sigma_I\right| - \frac{1}{2}\sum_{j=1}^{n}\left[\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)^T \Sigma_i^{-1}\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)\right] + c_1 \tag{29}$$

Maximum Likelihood

Thus

$$f\left(\Sigma_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[Trace\left\{\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\Sigma_i^{-1}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\right\}\right] + c_1 \quad (30)$$

Tricks!!!

$$f\left(\Sigma_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[Trace\left\{\Sigma_i^{-1}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\right\}\right] + c_1 \quad (31)$$

Maximum Likelihood

Thus

$$f\left(\Sigma_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[Trace\left\{\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\Sigma_i^{-1}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\right\}\right] + c_1 \tag{30}$$

Tricks!!!

$$f\left(\Sigma_i\right) = -\frac{n}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\sum_{j=1}^{n}\left[Trace\left\{\Sigma_i^{-1}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\right\}\right] + c_1 \tag{31}$$

Maximum Likelihood

Derivative with respect to Σ

$$\frac{\partial f\left(\Sigma_i\right)}{\partial \Sigma_i} = \frac{n}{2}\Sigma_i - \frac{1}{2}\sum_{j=1}^{n}\left[\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)^T\right]^T \quad (32)$$

Thus, when making it equal to zero

$$\hat{\Sigma}_i = \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j} - \boldsymbol{\mu_i}\right)^T \quad (33)$$

Maximum Likelihood

Derivative with respect to Σ

$$\frac{\partial f\left(\Sigma_i\right)}{\partial \Sigma_i}=\frac{n}{2}\Sigma_i-\frac{1}{2}\sum_{j=1}^{n}\left[\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T\right]^T \tag{32}$$

Thus, when making it equal to zero

$$\hat{\Sigma}_i=\frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)\left(\boldsymbol{x_j}-\boldsymbol{\mu_i}\right)^T \tag{33}$$

Therefore

Step 1 - Assume a Gaussian Distribution over each class

- The So Called Model Selection

Step 2

- Adjust the Gaussian Distribution, for each class, using the previous Maximum Likelihood

Step 3

$$
\begin{aligned}
R_1 &: P(\omega_1|x) > P(\omega_2|x) \\
R_2 &: P(\omega_2|x) > P(\omega_1|x)
\end{aligned}
$$

Therefore

Step 1 - Assume a Gaussian Distribution over each class

- The So Called Model Selection

Step 2

- Adjust the Gaussian Distribution, for each class, using the previous Maximum Likelihood

Step 3

$$
\begin{aligned}
R_1 &: P(\omega_1|x) > P(\omega_2|x) \\
R_2 &: P(\omega_2|x) > P(\omega_1|x)
\end{aligned}
$$

Therefore

Step 1 - Assume a Gaussian Distribution over each class

- The So Called Model Selection

Step 2

- Adjust the Gaussian Distribution, for each class, using the previous Maximum Likelihood

Step 3

$$
\begin{aligned}
R_1 &: P(\omega_1|x) > P(\omega_2|x) \\
R_2 &: P(\omega_2|x) > P(\omega_1|x)
\end{aligned}
$$

Outline

In the case of Bayesian Model

We have

$$P\left(Y_n = i|\boldsymbol{x}_n\right) = \frac{P\left(\boldsymbol{x}_n|Y_n = i\right) P\left(Y_n = i\right)}{P\left(\boldsymbol{x}_n\right)}$$

In the Generative Model

- We model two distribution $P\left(x_n|Y_n = 1\right)$ and $P\left(Y_n = i\right)$

In the Discriminative Model

- We model a single distribution $P\left(Y_n = i\right)$

In the case of Bayesian Model

We have

$$P\left(Y_n = i|\boldsymbol{x}_n\right) = \frac{P\left(\boldsymbol{x}_n|Y_n = i\right) P\left(Y_n = i\right)}{P\left(\boldsymbol{x}_n\right)}$$

In the Generative Model

- We model two distribution $P\left(\boldsymbol{x}_n|Y_n = 1\right)$ and $P\left(Y_n = i\right)$

In the Discriminative Model

- We model a single distribution $P\left(Y_n = i\right)$

In the case of Bayesian Model

We have

$$P\left(Y_n = i|\boldsymbol{x}_n\right) = \frac{P\left(\boldsymbol{x}_n|Y_n = i\right)P\left(Y_n = i\right)}{P\left(\boldsymbol{x}_n\right)}$$

In the Generative Model

- We model two distribution $P\left(\boldsymbol{x}_n|Y_n = 1\right)$ and $P\left(Y_n = i\right)$

In the Discriminative Model

- We model a single distribution $P\left(Y_n = i\right)$

Therefore

We have

- In the Generative Model, we discover the distribution from X and Y

Therefore

Although discriminative models tend to be faster and less complex, they cannot model the joint $P(X,Y)$.

Thus

- We have a decision problem
  - Do we want to know the joint distribution?

Therefore

We have

- In the Generative Model, we discover the distribution from X and Y

Therefore

Although discriminative models tend to be faster and less complex, they cannot model the joint $P(X, Y)$.

Thus

- We have a decision problem
  - Do we want to know the joint distribution?

Therefore

We have

- In the Generative Model, we discover the distribution from X and Y

Therefore

Although discriminative models tend to be faster and less complex, they cannot model the joint $P(X, Y)$.

Thus

- We have a decision problem
  - Do we want to know the joint distribution?

Outline

Introduction

We go back to the Bayesian Rule

$$p\left(\Theta|\mathcal{X}\right)=\frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{p\left(\mathcal{X}\right)} \tag{34}$$

We now seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p\left(\Theta|\mathcal{X}\right)$

Introduction

We go back to the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\, p(\Theta)}{p(\mathcal{X})} \tag{34}$$

We now seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} p\left(\Theta|\mathcal{X}\right)$$

$$= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)}$$

$$\approx \underset{\Theta}{\text{argmax}} p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)$$

$$= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\ &= \underset{\Theta}{\text{argmax}}\frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\ &\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\ &= \underset{\Theta}{\text{argmax}}\prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right) \\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\operatorname{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\operatorname{argmax}} p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right) \\
&= \underset{\Theta}{\operatorname{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\ &= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\ &\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\ &= \underset{\Theta}{\text{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

We can make this easier

Use logarithms

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\mathrm{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i|\Theta\right) + \log p\left(\Theta\right) \right] \tag{35}$$

Outline

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\, p(\Theta)}{p(\mathcal{X})} \quad (36)$$

We seek that value for Θ, called $\hat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p\left(\Theta|\mathcal{X}\right)=\frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{p\left(\mathcal{X}\right)} \tag{36}$$

We seek that value for Θ, called $\hat{\Theta}_{MAP}$

It allows to maximize the posterior $p\left(\Theta|\mathcal{X}\right)$

What can we do?

We can specify a distribution

Then, learn the parameters

Remember the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\, p(\Theta)}{p(\mathcal{X})} \tag{36}$$

We seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

Therefore

We can use this idea of maximizing the posterior

To obtain the distribution through the Maximum a Posteriori

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} p\left(\Theta|\mathcal{X}\right)$$

$$= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}$$

$$\approx \underset{\Theta}{\text{argmax}} p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)$$

$$= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} \, p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\operatorname{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\operatorname{argmax}} \, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\operatorname{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\operatorname{argmax}} p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\operatorname{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\operatorname{argmax}} p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\operatorname{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}} p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\text{argmax}} p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right) \\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

We can make this easier

Use logarithms

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i|\Theta\right) + \log p\left(\Theta\right) \right] \tag{37}$$

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{38}$$

The log likelihood function

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{38}$$

The log likelihood function

$$\begin{aligned}
\log\ p\left(\mathcal{X}|q\right) &= \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\
&= \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\
& \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\
&= n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right)
\end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} i = 1, ..., N \right\} \tag{38}$$

The log likelihood function

$$\begin{aligned} \log\ p(\mathcal{X}|q) &= \sum_{i=1}^{N} \log\ p(x_i|q) \\ &= \sum_{i} \log\ p(x_i = PRI|q) + ... \\ &\quad \sum_{i} \log\ p(x_i = PAN|1-q) \\ &= n_{PRI} \log(q) + (N - n_{PRI}) \log(1-q) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{38}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ =& \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ =& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{38}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ =& \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ =& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

We use our classic tricks

By setting

$$\mathcal{L} = \log \; p(\mathcal{X}|q) \tag{39}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{40}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} = 0 \tag{41}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log \; p(\mathcal{X}|q) \tag{39}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{40}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} = 0 \tag{41}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log\ p\left(\mathcal{X}|q\right) \tag{39}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{40}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} = 0 \tag{41}$$

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \tag{42}$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \tag{42}$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha - 1} (1 - q)^{\beta - 1}. \tag{43}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \tag{44}$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p\left(q\right) = \frac{1}{B\left(\alpha, \beta\right)} q^{\alpha-1}\left(1-q\right)^{\beta-1}. \tag{43}$$

Where

We have $B\left(\alpha, \beta\right) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha-1}{\alpha+\beta-2}. \tag{44}$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha - 1} (1 - q)^{\beta - 1}. \tag{43}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \tag{44}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{45}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{45}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{45}$$

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var(q) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var(q) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{46}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{47}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1} \right) \tag{48}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i|q\right) + \log p\left(q\right) \right] \tag{46}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{47}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha,\beta\right)} q^{\alpha-1}\left(1-q\right)^{\beta-1}\right) \tag{48}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{46}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{47}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1} \right) \tag{48}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{49}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{50}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{51}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1-q) - \log B(\alpha, \beta) \tag{49}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} - \frac{\beta - 1}{1-q} + \frac{\alpha - 1}{q} = 0 \tag{50}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{51}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{49}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{50}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{51}$$

Now

With $N = 20$ with $n_{PRI} = 12$ and $\alpha = \beta = 5$

$$\widehat{q}_{MAP} = 0.571$$

Outline

Properties

First

MAP estimation "pulls" the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

First

MAP estimation "pulls" the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

First

MAP estimation "pulls" the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

Third

In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Properties

Third

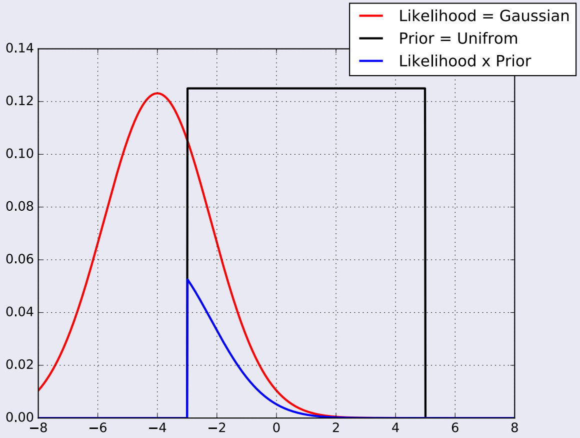
In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Basically the MAP

It is using the power of Likelihood $\times$ Prior to obtain more information from the data



Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the likelihood $\times$ the prior?

For example when we have something like $|\theta_i|$.

We can try the following

EM + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the likelihood $\times$ the prior?

For example when we have something like $|\theta_i|$.

We can try the following

EM + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the likelihood $\times$ the prior?

For example when we have something like $|\theta_i|$.

We can try the following

EM $+$ MAP to be able to estimate the sought parameters.

Outline

Exercises

Duda and Hart

Chapter 3

- 3.1, 3.2, 3.3, 3.13

Theodoridis

Chapter 2

- 2.5, 2.7, 2.10, 2.12, 2.14, 2.17

Exercises

Duda and Hart

Chapter 3

- 3.1, 3.2, 3.3, 3.13

Theodoridis

Chapter 2

- 2.5, 2.7, 2.10, 2.12, 2.14, 2.17