Introduction to Machine Learning

Introduction to Clustering

Andres Mendez-Vazquez

July 31, 2018

Outline

Outline

Supervised Learning vs. Unsupervised Learning

Supervised learning:

- Discover patterns in the data that relate data attributes with a target (class) attribute.

Unsupervised learning:

The data have no target attribute.

Supervised Learning vs. Unsupervised Learning

Supervised learning:

- Discover patterns in the data that relate data attributes with a target (class) attribute.

Unsupervised learning:

The data have no target attribute.

Outline

Clustering

Clustering

It is a technique for finding similarity groups in data, called clusters.

Clustering

Clustering

It is a technique for finding similarity groups in data, called clusters.

Called

An unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning.

- In fact, association rule mining is also unsupervised

Clustering

Clustering

It is a technique for finding similarity groups in data, called clusters.

Called

An unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning.

Due to historical reasons

Clustering is often considered synonymous with unsupervised learning.

- In fact, association rule mining is also unsupervised.

Clustering

Clustering

It is a technique for finding similarity groups in data, called clusters.

Called

An unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning.

Due to historical reasons

Clustering is often considered synonymous with unsupervised learning.

- In fact, association rule mining is also unsupervised.

Outline

Pattern Recognition

Definition

Search for structure in data

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. Cluster Analysis
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, . . .
3. **Cluster Analysis**
   - Labeling, Validity, . . .
4. Classifier Design
   - Classification, Estimation, Prediction, Control, . . .

Pattern Recognition

Definition

Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

Pattern Recognition

Definition

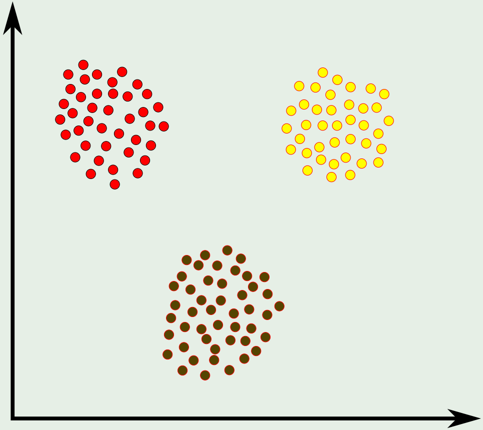
Search for structure in data

Elements of Numerical Pattern Recognition

1. Process Description
   - Feature Nomination, Test Data, Design Data
2. Feature Analysis
   - Preprocessing, Extraction, Selection, ...
3. **Cluster Analysis**
   - Labeling, Validity, ...
4. Classifier Design
   - Classification, Estimation, Prediction, Control, ...

An illustration

The data set has three natural groups of data points, i.e., 3 natural clusters.



Examples

Example 1

Groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

Example 2

In marketing, segment customers according to their similarities.

Examples

Example 1

Groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

Example 2

In marketing, segment customers according to their similarities.

How we create this classes?

For this, we use the following concept

Clustering!!!

Basically

We want to "reveal" the organization of patterns into "sensible" clusters (groups).

Actually

Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day.

How we create this classes?

For this, we use the following concept

Clustering!!!

Basically

We want to “reveal” the organization of patterns into “sensible” clusters (groups).

Actually

Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day.

How we create this classes?

For this, we use the following concept

Clustering!!!

Basically

We want to “reveal” the organization of patterns into “sensible” clusters (groups).

Actually

Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day.

Outline

What is clustering for?

Data Reduction

We can use clustering to reduce the amount of samples in each group to reduce the processing of information.

- Example:
  - In data transmission, a representative for each cluster is defined.
  - Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative.
  - Thus, data compression is achieved.

What is clustering for?

Data Reduction

We can use clustering to reduce the amount of samples in each group to reduce the processing of information.

- Example:
  - In data transmission, a representative for each cluster is defined.
  - Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative.
  - Thus, data compression is achieved.

Hypothesis Generation

In this case we apply cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data.

What is clustering for?

Data Reduction

We can use clustering to reduce the amount of samples in each group to reduce the processing of information.

- Example:
  - In data transmission, a representative for each cluster is defined.
  - Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative.
  - Thus, data compression is achieved.

Hypothesis Generation

In this case we apply cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data.

What is clustering for?

Data Reduction

We can use clustering to reduce the amount of samples in each group to reduce the processing of information.

- Example:
  - In data transmission, a representative for each cluster is defined.
  - Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative.
  - Thus, data compression is achieved.

Hypothesis Generation

In this case we apply cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data.

What is clustering for?

Data Reduction

We can use clustering to reduce the amount of samples in each group to reduce the processing of information.

- Example:
  - In data transmission, a representative for each cluster is defined.
  - Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative.
  - Thus, data compression is achieved.

Hypothesis Generation

In this case we apply cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - “Big Companies Invest Abroad”

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

What is clustering for?

Hypothesis testing

In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.

- Example:
  - "Big Companies Invest Abroad"

Prediction based on groups

First: We apply cluster analysis to the available data set.

Second: The resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

Third: If we are given an unknown pattern, we can determine the cluster to which it is more likely to belong.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Based in a function

A distance (similarity, or dissimilarity) function.

- Intra-clusters distance $\rightarrow$ minimized.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Based in a function

A distance (similarity, or dissimilarity) function.

Clustering quality

- Inter-clusters distance → maximized.
- Intra-clusters distance → minimized.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Based in a function

A distance (similarity, or dissimilarity) function.

Clustering quality

- Inter-clusters distance $\rightarrow$ maximized.
- Intra-clusters distance $\rightarrow$ minimized.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Based in a function

A distance (similarity, or dissimilarity) function.

Clustering quality

- Inter-clusters distance $\rightarrow$ maximized.
- Intra-clusters distance $\rightarrow$ minimized.

Aspects of clustering

A clustering algorithm - They are Many!!!

- Partition clustering.
- Hierarchical clustering.
- etc.

Based in a function

A distance (similarity, or dissimilarity) function.

Clustering quality

- Inter-clusters distance $\rightarrow$ maximized.
- Intra-clusters distance $\rightarrow$ minimized.

Outline

Introduction

Observation

Clusters are considered as groups containing data objects that are similar to each other.

When they are not

We assume there is some way to measure that difference!!!

Introduction

Observation

Clusters are considered as groups containing data objects that are similar to each other.

When they are not

We assume there is some way to measure that difference!!!

Therefore

It is possible to give a mathematical definition of two important types of clustering

- Partitional Clustering
- Hierarchical Clustering

Therefore

It is possible to give a mathematical definition of two important types of clustering

- Partitional Clustering
- Hierarchical Clustering

Given

Given a set of input patterns $X = \{x_1, x_2, ..., x_N\}$, where $x_j = (x_{1j}, x_{2j}, ..., x_{dj})^T \in \mathbb{R}^d$, with each measure x_{ij} called a feature (attribute, dimension, or variable).

Therefore

It is possible to give a mathematical definition of two important types of clustering

- Partitional Clustering
- Hierarchical Clustering

Given

Given a set of input patterns $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, where $\boldsymbol{x}_j = (x_{1j}, x_{2j}, ..., x_{dj})^T \in \mathbb{R}^d$, with each measure x_{ij} called a feature (attribute, dimension, or variable).

Hard Partitional Clustering

Definition

Hard partitional clustering attempts to seek a K-partition of $\boldsymbol{X}$, $\boldsymbol{C} = \{C_1, C_2, ..., C_K\}$ with $K \leq N$ such that

1. $C_i \neq \emptyset$ for $i = 1, ..., K$.
2. $\cup_{i=1}^{K} C_i = \boldsymbol{X}$.
3. $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$ and $i \neq j$.

Hard Partitional Clustering

Definition

Hard partitional clustering attempts to seek a K-partition of $\boldsymbol{X}$, $\boldsymbol{C} = \{C_1, C_2, ..., C_K\}$ with $K \leq N$ such that

1. $C_i \neq \emptyset$ for $i = 1, ..., K$.
2. $\cup_{i=1}^{K} C_i = \boldsymbol{X}$.
3. $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$ and $i \neq j$.

The third property can be relaxed

- To obtain smoother versions of Hard Clustering

Hard Partitional Clustering

Definition

Hard partitional clustering attempts to seek a K-partition of $\boldsymbol{X}$, $\boldsymbol{C} = \{C_1, C_2, ..., C_K\}$ with $K \leq N$ such that

1. $C_i \neq \emptyset$ for $i = 1, ..., K$.
2. $\cup_{i=1}^{K} C_i = \boldsymbol{X}$.
3. $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$ and $i \neq j$.

The third property can be relaxed

- To obtain smoother versions of Hard Clustering

Hard Partitional Clustering

Definition

Hard partitional clustering attempts to seek a K-partition of $\boldsymbol{X}$, $\boldsymbol{C} = \{C_1, C_2, ..., C_K\}$ with $K \leq N$ such that

1. $C_i \neq \emptyset$ for $i = 1, ..., K$.
2. $\cup_{i=1}^{K} C_i = \boldsymbol{X}$.
3. $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$ and $i \neq j$.

The third property can be relaxed

- To obtain smoother versions of Hard Clustering

Hard Partitional Clustering

Definition

Hard partitional clustering attempts to seek a K-partition of $\boldsymbol{X}$, $\boldsymbol{C} = \{C_1, C_2, ..., C_K\}$ with $K \leq N$ such that

1. $C_i \neq \emptyset$ for $i = 1, ..., K$.
2. $\cup_{i=1}^{K} C_i = \boldsymbol{X}$.
3. $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$ and $i \neq j$.

The third property can be relaxed

- To obtain smoother versions of Hard Clustering

For example, Fuzzy and Possibilistic clustering

Given the membership $A_i(\boldsymbol{x}_j) \in [0,1]$

We can have $\sum_{i=1}^{K} A_i(\boldsymbol{x}_j) = 1, \forall j$ and $\sum_{j=1}^{N} A_i(\boldsymbol{x}_j) < N, \forall i$.

Hierarchical Clustering

Definition

Hierarchical clustering attempts to construct a tree-like, nested structure partition of $\boldsymbol{X}$, $\boldsymbol{H} = \{H_1, ..., H_Q\}$ with $Q \leq N$ such that

- If $C_i \in H_m$ and $C_j \in H_l$ and $m > l$ imply $C_i \subset C_j$ or $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$, $i \neq j$ and $m, l = 1, ..., Q$.

Hierarchical Clustering

Definition

Hierarchical clustering attempts to construct a tree-like, nested structure partition of $\boldsymbol{X}$, $\boldsymbol{H} = \{H_1, ..., H_Q\}$ with $Q \leq N$ such that

- If $C_i \in H_m$ and $C_j \in H_l$ and $m > l$ imply $C_i \subset C_j$ or $C_i \cap C_j = \emptyset$ for all $i, j = 1, ..., K$, $i \neq j$ and $m, l = 1, ..., Q$.

Outline

Features

Usually

A data object is described by a set of features or variables, usually represented as a multidimensional vector.

Thus

We tend to collect all that information as a pattern matrix of dimensionality $N \times d$.

Then

A feature can be classified as continuous, discrete, or binary.

Features

Usually

A data object is described by a set of features or variables, usually represented as a multidimensional vector.

Thus

We tend to collect all that information as a pattern matrix of dimensionality $N \times d$.

Then

A feature can be classified as continuous, discrete, or binary.

Features

Usually

A data object is described by a set of features or variables, usually represented as a multidimensional vector.

Thus

We tend to collect all that information as a pattern matrix of dimensionality $N \times d$.

Then

A feature can be classified as continuous, discrete, or binary.

Thus

Continuous

A continuous feature takes values from an uncountably infinite range set.

Discrete

Discrete features only have finite, or at most, a countably infinite number of values.

Binary

Binary or dichotomous features are a special case of discrete features when they have exactly two values.

Thus

Continuous

A continuous feature takes values from an uncountably infinite range set.

Discrete

Discrete features only have finite, or at most, a countably infinite number of values.

Binary

Binary or dichotomous features are a special case of discrete features when they have exactly two values.

Thus

Continuous

A continuous feature takes values from an uncountably infinite range set.

Discrete

Discrete features only have finite, or at most, a countably infinite number of values.

Binary

Binary or dichotomous features are a special case of discrete features when they have exactly two values.

Outline

Measurement Level

Meaning

Another property of features is the measurement level, which reflects the relative significance of numbers.

We have four from lowest to highest

Nominal, ordinal, interval, and ratio.

Measurement Level

Meaning

Another property of features is the measurement level, which reflects the relative significance of numbers.

We have four from lowest to highest

Nominal, ordinal, interval, and ratio.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Ordinal

- Features at this level are also names, but with a certain order implied.
- However, the difference between the values is again meaningless.
- For example, abdominal distension $\in$ {slight, moderate, high}.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Ordinal

- Features at this level are also names, but with a certain order implied.
- However, the difference between the values is again meaningless.
- For example, abdominal distension $\in$ {slight, moderate, high}.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Ordinal

- Features at this level are also names, but with a certain order implied.
- However, the difference between the values is again meaningless.
- For example, abdominal distension $\in$ {slight, moderate, high}.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Ordinal

- Features at this level are also names, but with a certain order implied.
- However, the difference between the values is again meaningless.
- For example, abdominal distension $\in$ {slight, moderate, high}.

Measurement Level

Nominal

- Features at this level are represented with labels, states, or names.
- The numbers are meaningless in any mathematical sense and no mathematical calculation can be made.
- For example, peristalsis $\in$ {hypermotile, normal, hypomotile, absent}.

Ordinal

- Features at this level are also names, but with a certain order implied.
- However, the difference between the values is again meaningless.
- For example, abdominal distension $\in$ {slight, moderate, high}.

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Ratio

- Features at this level possess all the properties of the above levels.
- There is an absolute zero.
- Thus, we have the meaning of ratio!!!

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Ratio

- Features at this level possess all the properties of the above levels.
- There is an absolute zero.
- Thus, we have the meaning of ratio!!!

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Ratio

- Features at this level possess all the properties of the above levels.
- There is an absolute zero.
- Thus, we have the meaning of ratio!!!

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Ratio

- Features at this level possess all the properties of the above levels.
- There is an absolute zero.
- Thus, we have the meaning of ratio!!!

Measurement Level

Interval

- Features at this level offer a meaningful interpretation of the difference between two values.
- However, there exists no true zero and the ratio between two values is meaningless.
- For example, the concept cold can be seen as an interval.

Ratio

- Features at this level possess all the properties of the above levels.
- There is an absolute zero.
- Thus, we have the meaning of ratio!!!

Outline

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \to \mathbb{R} \tag{1}$$

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \to \mathbb{R} \tag{1}$$

Such that

1. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) \leq s_0$.
2. $s\left(\boldsymbol{x}, \boldsymbol{x}\right) = s_0$ for all $\boldsymbol{x} \in X$.
3. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s\left(\boldsymbol{y}, \boldsymbol{x}\right)$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) s\left(\boldsymbol{y}, \boldsymbol{z}\right) \leq s\left(\boldsymbol{x}, \boldsymbol{z}\right) \left[s\left(\boldsymbol{x}, \boldsymbol{y}\right) + s\left(\boldsymbol{y}, \boldsymbol{z}\right)\right]$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \to \mathbb{R} \tag{1}$$

Such that

1. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) \leq s_0$.
2. $s\left(\boldsymbol{x}, \boldsymbol{x}\right) = s_0$ for all $\boldsymbol{x} \in X$.
3. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s\left(\boldsymbol{y}, \boldsymbol{x}\right)$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) s\left(\boldsymbol{y}, \boldsymbol{z}\right) \leq s\left(\boldsymbol{x}, \boldsymbol{z}\right) \left[s\left(\boldsymbol{x}, \boldsymbol{y}\right) + s\left(\boldsymbol{y}, \boldsymbol{z}\right)\right]$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \rightarrow \mathbb{R} \tag{1}$$

Such that

1. $s\left(\boldsymbol{x},\boldsymbol{y}\right) \leq s_0$.
2. $s\left(\boldsymbol{x},\boldsymbol{x}\right) = s_0$ for all $\boldsymbol{x} \in X$.
3. $s\left(\boldsymbol{x},\boldsymbol{y}\right) = s\left(\boldsymbol{y},\boldsymbol{x}\right)$ for all $\boldsymbol{x},\boldsymbol{y} \in X$.
4. If $s\left(\boldsymbol{x},\boldsymbol{y}\right) = s_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $s\left(\boldsymbol{x},\boldsymbol{y}\right) s\left(\boldsymbol{y},\boldsymbol{z}\right) \leq s\left(\boldsymbol{x},\boldsymbol{z}\right)\left[s\left(\boldsymbol{x},\boldsymbol{y}\right) + s\left(\boldsymbol{y},\boldsymbol{z}\right)\right]$ for all $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in X$.

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \to \mathbb{R} \tag{1}$$

Such that

1. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) \leq s_0$.
2. $s\left(\boldsymbol{x}, \boldsymbol{x}\right) = s_0$ for all $\boldsymbol{x} \in X$.
3. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s\left(\boldsymbol{y}, \boldsymbol{x}\right)$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) s\left(\boldsymbol{y}, \boldsymbol{z}\right) \leq s\left(\boldsymbol{x}, \boldsymbol{z}\right)\left[s\left(\boldsymbol{x}, \boldsymbol{y}\right) + s\left(\boldsymbol{y}, \boldsymbol{z}\right)\right]$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Similarity Measures

Definition

A Similarity Measure s on X is a function

$$s : X \times X \rightarrow \mathbb{R} \tag{1}$$

Such that

1. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) \leq s_0$.
2. $s\left(\boldsymbol{x}, \boldsymbol{x}\right) = s_0$ for all $\boldsymbol{x} \in X$.
3. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s\left(\boldsymbol{y}, \boldsymbol{x}\right)$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $s\left(\boldsymbol{x}, \boldsymbol{y}\right) = s_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $s\left(\boldsymbol{x}, \boldsymbol{y}\right) s\left(\boldsymbol{y}, \boldsymbol{z}\right) \leq s\left(\boldsymbol{x}, \boldsymbol{z}\right)\left[s\left(\boldsymbol{x}, \boldsymbol{y}\right) + s\left(\boldsymbol{y}, \boldsymbol{z}\right)\right]$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Outline

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \to \mathbb{R} \tag{2}$$

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \to \mathbb{R} \tag{2}$$

Such that

1. $d(\boldsymbol{x}, \boldsymbol{y}) \geq d_0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$ for all $\boldsymbol{x} \in X$.
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \to \mathbb{R} \tag{2}$$

Such that

1. $d(\boldsymbol{x}, \boldsymbol{y}) \geq d_0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$ for all $\boldsymbol{x} \in X$.
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \to \mathbb{R} \tag{2}$$

Such that

1. $d(\boldsymbol{x}, \boldsymbol{y}) \geq d_0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$ for all $\boldsymbol{x} \in X$.
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \rightarrow \mathbb{R} \tag{2}$$

Such that

1. $d(\boldsymbol{x}, \boldsymbol{y}) \geq d_0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0$ for all $\boldsymbol{x} \in X$.
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Dissimilarity Measures

Defintion

A Dissimilarity Measure d on X is a function

$$d : X \times X \to \mathbb{R} \tag{2}$$

Such that

1. $d\left(\boldsymbol{x}, \boldsymbol{y}\right) \geq d_0$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
2. $d\left(\boldsymbol{x}, \boldsymbol{x}\right) = d_0$ for all $\boldsymbol{x} \in X$.
3. $d\left(\boldsymbol{x}, \boldsymbol{y}\right) = d\left(\boldsymbol{y}, \boldsymbol{x}\right)$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.
4. If $d\left(\boldsymbol{x}, \boldsymbol{y}\right) = d_0 \iff \boldsymbol{x} = \boldsymbol{y}$.
5. $d\left(\boldsymbol{x}, \boldsymbol{z}\right) \leq d\left(\boldsymbol{x}, \boldsymbol{y}\right) + d\left(\boldsymbol{y}, \boldsymbol{z}\right)$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$.

Outline

Outline

Dissimilarity Measures

The Hamming distance for $\{0,1\}$

The vectors here are collections of zeros and ones. Thus,

$$d_H(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{d} (x_i - y_i)^2 \tag{3}$$

Euclidean distance

$$d(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^{\frac{1}{2}} \right)^2 \tag{4}$$

Dissimilarity Measures

The Hamming distance for $\{0,1\}$

The vectors here are collections of zeros and ones. Thus,

$$d_H\left(\boldsymbol{x},\boldsymbol{y}\right)=\sum_{i=1}^{d}\left(x_i-y_i\right)^2 \tag{3}$$

Euclidean distance

$$d\left(\boldsymbol{x},\boldsymbol{y}\right)=\left(\sum_{i=1}^{d}\left|x_i-y_i\right|^{\frac{1}{2}}\right)^2 \tag{4}$$

Dissimilarity Measures

The weighted l_p metric DMs

$$d_p\left(\boldsymbol{x},\boldsymbol{y}\right)=\left(\sum_{i=1}^{d} w_i\left|x_i-y_i\right|^p\right)^{\frac{1}{p}} \tag{5}$$

Where

Where x_i, y_i are the ith coordinates of x and y, $i = 1, \ldots, d$ and $w_i \geq 0$ is the ith weight coefficient.

Important

A well-known representative of the latter category of measures is the Euclidean distance.

Dissimilarity Measures

The weighted l_p metric DMs

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{d} w_i \left| x_i - y_i \right|^p \right)^{\frac{1}{p}} \tag{5}$$

Where

Where x_i, y_i are the ith coordinates of x and y, $i = 1, ..., d$ and $w_i \geq 0$ is the ith weight coefficient.

Important

A well-known representative of the latter category of measures is the Euclidean distance.

Dissimilarity Measures

The weighted l_p metric DMs

$$d_p\left(\boldsymbol{x},\boldsymbol{y}\right)=\left(\sum_{i=1}^{d} w_i \left|x_i-y_i\right|^p\right)^{\frac{1}{p}} \tag{5}$$

Where

Where x_i, y_i are the ith coordinates of x and y, $i=1,...,d$ and $w_i \geq 0$ is the ith weight coefficient.

Important

A well-known representative of the latter category of measures is the Euclidean distance.

Dissimilarity Measures

The weighted l_2 metric DM can be further generalized as follows

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T B (\boldsymbol{x} - \boldsymbol{y})} \tag{6}$$

Where B is a symmetric, positive definite matrix

A special case

This includes the Mahalanobis distance as a special case

Dissimilarity Measures

The weighted l_2 metric DM can be further generalized as follows

$$d\left(\boldsymbol{x}, \boldsymbol{y}\right) = \sqrt{\left(\boldsymbol{x} - \boldsymbol{y}\right)^T B \left(\boldsymbol{x} - \boldsymbol{y}\right)} \tag{6}$$

Where B is a symmetric, positive definite matrix

A special case

This includes the Mahalanobis distance as a special case

Dissimilarity Measures

Special l_p metric DMs that are also encountered in practice are

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} w_i |x_i - y_i| \quad \text{(Weighted Manhattan Norm)}$$

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le i \le d} w_i |x_i - y_i| \quad \text{(Weighted } l_\infty \text{ Norm)}$$

Dissimilarity Measures

Special l_p metric DMs that are also encountered in practice are

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} w_i |x_i - y_i| \text{ (Weighted Manhattan Norm)}$$

$$d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \leq i \leq d} w_i |x_i - y_i| \text{ (Weighted } l_\infty \text{ Norm)}$$

Outline

Similarity Measures

The inner product

$$s_{inner}\left(\boldsymbol{x},\boldsymbol{y}\right)=\boldsymbol{x}^{T}\boldsymbol{y}=\sum_{i=1}^{d}x_i y_i \tag{7}$$

Closely related to the inner product is the cosine similarity measure

$$s_{cosine}\left(\boldsymbol{x},\boldsymbol{y}\right)=\frac{\boldsymbol{x}^{T}\boldsymbol{y}}{\left\Vert \boldsymbol{x}\right\Vert \left\Vert \boldsymbol{y}\right\Vert} \tag{8}$$

Where $\left\Vert \boldsymbol{x}\right\Vert=\sqrt{\sum_{i=1}^{d}x_i^2}$ and $\left\Vert \boldsymbol{y}\right\Vert=\sqrt{\sum_{i=1}^{d}y_i^2}$, the length of the vectors!!!

Similarity Measures

The inner product

$$s_{inner}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{d} x_i y_i \tag{7}$$

Closely related to the inner product is the cosine similarity measure

$$s_{cosine}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \tag{8}$$

Where $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{d} x_i^2}$ and $\|\boldsymbol{y}\| = \sqrt{\sum_{i=1}^{d} y_i^2}$, the length of the vectors!!!

Similarity Measures

Pearson's correlation coefficient

$$r_{\text{Pearson}}\left(\boldsymbol{x},\boldsymbol{y}\right) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \tag{9}$$

Similarity Measures

Pearson's correlation coefficient

$$r_{\text{Pearson}}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} \tag{9}$$

Where $\boldsymbol{x} = \left(x_1 - \overline{\boldsymbol{x}}, ..., x_d - \overline{\boldsymbol{x}}\right)^T$ and $\boldsymbol{y} = \left(y_1 - \overline{\boldsymbol{y}}, ..., y_d - \overline{\boldsymbol{y}}\right)^T$

$$\overline{\boldsymbol{x}} = \frac{1}{d}\sum_{i=1}^{d} x_i \text{ and } \overline{\boldsymbol{y}} = \frac{1}{d}\sum_{i=1}^{d} y_i \tag{10}$$

- The difference between r_{cosine} and r_{Pearson} does not depend directly on $\boldsymbol{x}$ and $\boldsymbol{y}$ but on their corresponding difference vectors.

Similarity Measures

Pearson's correlation coefficient

$$r_{\text{Pearson}}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} \tag{9}$$

Where $\boldsymbol{x} = \left(x_1 - \overline{\boldsymbol{x}}, ..., x_d - \overline{\boldsymbol{x}}\right)^T$ and $\boldsymbol{y} = \left(y_1 - \overline{\boldsymbol{y}}, ..., y_d - \overline{\boldsymbol{y}}\right)^T$

$$\overline{\boldsymbol{x}} = \frac{1}{d}\sum_{i=1}^{d} x_i \text{ and } \overline{\boldsymbol{y}} = \frac{1}{d}\sum_{i=1}^{d} y_i \tag{10}$$

Properties

- It is clear that $r_{\text{Pearson}}\left(\boldsymbol{x}, \boldsymbol{y}\right)$ takes values between -1 and +1.
- The difference between s_{inner} and r_{Pearson} does not depend directly on x and y but on their corresponding difference vectors.

Similarity Measures

Pearson's correlation coefficient

$$r_{\text{Pearson}}\left(\boldsymbol{x},\boldsymbol{y}\right)=\frac{\boldsymbol{x}^{T}\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \tag{9}$$

Where $\boldsymbol{x}=\left(x_1-\overline{\boldsymbol{x}},\ldots,x_d-\overline{\boldsymbol{x}}\right)^{T}$ and $\boldsymbol{y}=\left(y_1-\overline{\boldsymbol{y}},\ldots,y_d-\overline{\boldsymbol{y}}\right)^{T}$

$$\overline{\boldsymbol{x}}=\frac{1}{d}\sum_{i=1}^{d}x_i \text{ and } \overline{\boldsymbol{y}}=\frac{1}{d}\sum_{i=1}^{d}y_i \tag{10}$$

Properties

- It is clear that $r_{\text{Pearson}}\left(\boldsymbol{x},\boldsymbol{y}\right)$ takes values between -1 and +1.
- The difference between s_{inner} and r_{Pearson} does not depend directly on $\boldsymbol{x}$ and $\boldsymbol{y}$ but on their corresponding difference vectors.

Similarity Measures

The Tanimoto measure or distance

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}} \tag{11}$$

Something Notable

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + \frac{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})}{\boldsymbol{x}^T \boldsymbol{y}}} \tag{12}$$

Meaning

The Tanimoto measure is inversely proportional to the squared Euclidean distance.

Similarity Measures

The Tanimoto measure or distance

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T \boldsymbol{y}} \tag{11}$$

Something Notable

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + \frac{(\boldsymbol{x}-\boldsymbol{y})^T (\boldsymbol{x}-\boldsymbol{y})}{\boldsymbol{x}^T \boldsymbol{y}}} \tag{12}$$

Meaning

The Tanimoto measure is inversely proportional to the squared Euclidean distance.

Similarity Measures

The Tanimoto measure or distance

$$s_T\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\left\|\boldsymbol{x}\right\|^2 + \left\|\boldsymbol{y}\right\|^2 - \boldsymbol{x}^T \boldsymbol{y}} \tag{11}$$

Something Notable

$$s_T\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{1}{1 + \frac{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})}{\boldsymbol{x}^T \boldsymbol{y}}} \tag{12}$$

Meaning

The Tanimoto measure is inversely proportional to the squared Euclidean distance.

Outline

We consider now vectors whose coordinates belong to a finite set

Given $F = \{0, 1, 2, ..., k-1\}$ where k is a positive integer

There are exactly k^d vectors $\boldsymbol{x} \in F^d$

Now consider $\boldsymbol{x}, \boldsymbol{y} \in F^d$

With $A(\boldsymbol{x}, \boldsymbol{y}) = [a_{ij}]$, $i, j = 0, 1, ..., k-1$ be a $k \times k$ matrix

Where

The element a_{ij} is the number of places where the first vector has the i symbol and the corresponding element of the second vector has the j symbol.

We consider now vectors whose coordinates belong to a finite set

Given $F = \{0, 1, 2, ..., k-1\}$ where k is a positive integer

There are exactly k^d vectors $\boldsymbol{x} \in F^d$

Now consider $\boldsymbol{x}, \boldsymbol{y} \in F^d$

With $A(\boldsymbol{x}, \boldsymbol{y}) = [a_{ij}]$, $i, j = 0, 1, ..., k-1$ be a $k \times k$ matrix

Where

The element a_{ij} is the number of places where the first vector has the i symbol and the corresponding element of the second vector has the j symbol.

We consider now vectors whose coordinates belong to a finite set

Given $F = \{0, 1, 2, ..., k-1\}$ where k is a positive integer

There are exactly k^d vectors $\boldsymbol{x} \in F^d$

Now consider $\boldsymbol{x}, \boldsymbol{y} \in F^d$

With $A(\boldsymbol{x}, \boldsymbol{y}) = [a_{ij}]$, $i, j = 0, 1, ..., k-1$ be a $k \times k$ matrix

Where

The element a_{ij} is the number of places where the first vector has the i symbol and the corresponding element of the second vector has the j symbol.

Using Indicator functions

We can think of each element a_{ij}

$$a_{ij} = \sum_{r=1}^{d} I\left[(i,j) == (x_r, y_r)\right] \quad (13)$$

Thanks to Ricardo Llamas and Gibran Felix!!! Class Summer 2015.

Contingency Matrix

Thus

This matrix is called a **contingency table.**

For example if $d = 6$ and $k = 3$

With $\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T$ and $\boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$

Then

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{14}$$

Contingency Matrix

Thus

This matrix is called a **contingency table.**

For example if $d = 6$ and $k = 3$

With $\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T$ and $\boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$

Then

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{14}$$

Contingency Matrix

Thus

This matrix is called a **contingency table.**

For example if $d = 6$ and $k = 3$

With $\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T$ and $\boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$

Then

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{14}$$

Thus, we have that

Easy to verify that

$$\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} a_{ij} = d \tag{15}$$

Something Notable

Most of the proximity measures between two discrete-valued vectors may be expressed as combinations of elements of matrix $A(x, y)$.

Thus, we have that

Easy to verify that

$$\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} a_{ij} = d \tag{15}$$

Something Notable

Most of the proximity measures between two discrete-valued vectors may be expressed as combinations of elements of matrix $A(\boldsymbol{x}, \boldsymbol{y})$.

Outline

Dissimilarity Measures

The Hamming distance

It is defined as the number of places where two vectors differ:

$$d\left(\boldsymbol{x},\boldsymbol{y}\right)=\sum_{i=0}^{k-1}\sum_{j=0,j\neq i}^{k-1}a_{ij} \tag{16}$$

The summation of all the off-diagonal elements of A, which indicate the positions where $\boldsymbol{x}$ and $\boldsymbol{y}$ differ.

Special case $k=2$, thus vectors are binary valued

The vectors $\boldsymbol{x}\in F^{d}$ are binary valued and the Hamming distance becomes

$$\begin{aligned}d_H\left(\boldsymbol{x},\boldsymbol{y}\right)&=\sum_{i=1}^{d}\left(x_i+y_i-2x_iy_i\right)\\&=\sum_{i=1}^{d}\left(x_i-y_i\right)^2\end{aligned}$$

Dissimilarity Measures

The Hamming distance

It is defined as the number of places where two vectors differ:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{k-1} \sum_{j=0, j \neq i}^{k-1} a_{ij} \quad (16)$$

The summation of all the off-diagonal elements of A, which indicate the positions where $\boldsymbol{x}$ and $\boldsymbol{y}$ differ.

Special case $k = 2$, thus vectors are binary valued

The vectors $\boldsymbol{x} \in F^d$ are binary valued and the Hamming distance becomes

$$\begin{aligned} d_H(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i=1}^{d} (x_i + y_i - 2x_i y_i) \\ &= \sum_{i=1}^{d} (x_i - y_i)^2 \end{aligned}$$

Dissimilarity Measures

The l_1 distance

It is defined as in the case of the continuous-valued vectors,

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} w_i |x_i - y_i| \tag{17}$$

Outline

Similarity Measures

The Tanimoto measure

Given two sets X and Y, we have that

$$s_T(X, Y) = \frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}} \tag{18}$$

where $n_X = |X|$, $n_Y = |Y|$, $n_{X \cap Y} = |X \cap Y|$.

Similarity Measures

The Tanimoto measure

Given two sets X and Y, we have that

$$s_T(X, Y) = \frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}} \tag{18}$$

where $n_X = |X|$, $n_Y = |Y|$, $n_{X \cap Y} = |X \cap Y|$.

Thus, if x, y are discrete-valued vectors

The measure takes into account all pairs of corresponding coordinates, except those whose corresponding coordinates (x_i, y_i) are both 0.

Similarity Measures

The Tanimoto measure

Given two sets X and Y, we have that

$$s_T(X,Y) = \frac{n_{X\cap Y}}{n_X + n_Y - n_{X\cap Y}} \tag{18}$$

where $n_X = |X|$, $n_Y = |Y|$, $n_{X\cap Y} = |X \cap Y|$.

Thus, if $\boldsymbol{x}, \boldsymbol{y}$ are discrete-valued vectors

The measure takes into account all pairs of corresponding coordinates, except those whose corresponding coordinates (x_i, y_i) are both 0.

Thus

We now define

$n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$ and $n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$

In other words

n_x and n_y denotes the number of the nonzero coordinate of x and y.

Thus, we have

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ii}}{n_X + n_Y - \sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ij}} \quad (19)$$

Thus

We now define

$n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$ and $n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$

In other words

n_x and n_y denotes the number of the nonzero coordinate of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Thus, we have

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ii}}{n_X + n_Y - \sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ij}} \tag{19}$$

Thus

We now define

$n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$ and $n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$

In other words

n_x and n_y denotes the number of the nonzero coordinate of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Thus, we have

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ii}}{n_X + n_Y - \sum_{i=1}^{k-1} \sum_{j=i}^{k-1} a_{ij}} \tag{19}$$

Outline

Between Sets

Jaccard Similarity

Given two sets X and Y, we have that

Between Sets

Jaccard Similarity

Given two sets X and Y, we have that

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{20}$$

Although there are more stuff to look for

Please Refer to

1. Theodoridis' Book Chapter 11.
2. Rui Xu and Don Wunsch. 2009. *Clustering*. Wiley-IEEE Press.