IT496: Introduction to Data Mining

Lecture 04 – 06

Statistics for Data Mining - III

[Measures of Proximity]

Arpit Rana

1st / 3rd / 4th August 2023

Measures of
Similarity and Dissimilarity

. . .how alike and unalike the data objects are in comparison to one-another. . .

Disclaimer: Most images incorporated within the presentation slides have been sourced from Introduction to Data Mining by Peng-Ning Tan..

Definitions

Measures of Proximity[1]
(between two objects[2])

Similarity

A numerical measure of the degree to which the two objects are *alike*.

- They are usually non-negative (>= 0) and are often between 0 (unalike) and 1 (alike).

Dissimilarity

A numerical measure of the degree to which the two objects are *different*.

- They usually fall in the interval [0, 1], but it is also common for them to range between [0, ∞).

1 - *For convenience, the term proximity is used to refer to either similarity or dissimilarity.*

2 - *The proximity between two objects is a function of the proximity between the corresponding attributes of the two objects,*

Motivation

They are used by a number of data mining techniques,

- such as clustering, nearest neighbor classification, and anomaly detection.
- some approaches transforms the data to a similarity (dissimilarity) space and then performs the analysis, e.g., *kernel methods*.

Facts

We will observe the following -

- *Jaccard* and *Cosine* similarities are used on *sparse* data, e.,g., documents, user ratings
- *Euclidean* distance and *Correlation* are used on *dense* data, e.g., time-series, multidimensional data
- *Correlation* captures the linear relationship while **mutual information** detects non-linear relationships between the two variables.

Transformations

Transformations are often applied -

- to convert a similarity to a dissimilarity, or vice versa, or
- to transform a proximity measure to fall within a particular range, such as [0,1].

Linear Transformations

- It preserves the relative distances between points.
  - *Min-max transformation*, $d' = \dfrac{d - d_{min}}{d_{\max} - d_{\min}}$

- If the similarity falls in the interval [0,1], then the dissimilarity can be defined as

$$d = 1 - s$$

- Other transformations (*any monotonically decreasing function*):

Linear Transformations

- A few examples of *monotonically decreasing function of the distance d that returns similarity within the range of [0, 1]*:

d	$s = \frac{1}{d+1}$	$s = e^{-d}$	$s = 1 - \frac{d - d_{\text{min}}}{d_{\text{max}} - d_{min}}$
0	1	1.00	1.00
1	0.5	0.37	0.99
10	0.09	0.00	0.90
100	0.01	0.00	0.00

Objects with Single Attribute

Proximity

We first discuss proximity between objects having a single attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$	$s = 1 - d = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}$
Ordinal	$d = \frac{\lvert x - y \rvert}{n - 1}$ (values mapped to integers 0 to n − 1, where n is the number of values)	$s = 1 - d$
Numeric (interval or ratio)	$d = \lvert x - y \rvert$	$s = \frac{1}{d+1}$; $s = 1 - \frac{d - d_{\text{min}}}{d_{\text{max}} - d_{min}}$ $s = e^{-d}$; $s = -d$

A Scenario on an Ordinal Attribute

Consider an attribute that measures the quality of a product, e.g., a candy bar, on the scale {*poor*, *fair*, OK, *good*, *wonderful* }.

- A product, x_1, which is rated ***wonderful***, would be closer to a product x_2, which is rated ***good***, than it would be to a product x_3, which is rated OK.

- To make this observation quantitative, the values of the ordinal attribute are often mapped to successive integers, beginning at 0 or 1, e.g., {*poor*=0, *fair*=1, OK=2, *good*=3, *wonderful*=4 }

  Then, $d(x_1, x_2) = 3 - 2 = 1$ or, if we want the dissimilarity to fall between 0 and 1,

  $d(x_1, x_2) = (3-2) / 4 = 0.25$

<u>Ques.</u> Is the difference between the values '*fair*' and '*good*' really the same as that between the values 'OK' and '*wonderful*' ?

Dissimilarities between Data Objects

. . . that involve multiple attributes. . .

Distances

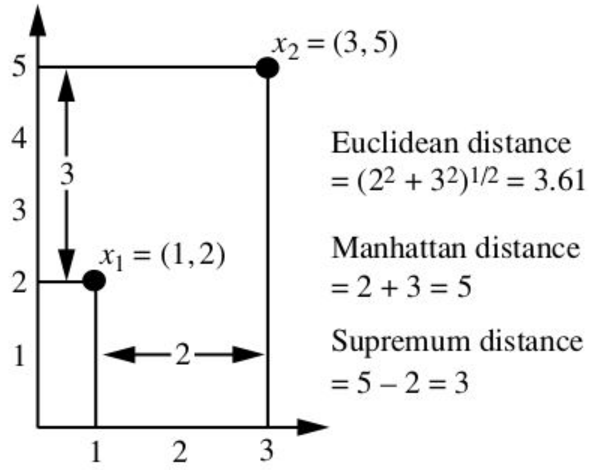
If $d(x, y)$ is the distance between two points (in our case, data objects), x and y,

$$d(x,y) = \left(\sum_{k=1}^{p} |x_k - y_k|^r \right)^{1/r}$$

This is **Minkowski** distance metric. Where, r is a *parameter*.

- $r = 1$, **Manhattan** distance or L_1 norm
  (e.g. *Hamming* distance for binary attributes)
- $r = 2$, **Euclidean** distance or L_2 norm
- $r = \infty$, **Supremum** distance or L_{max} or L_∞ norm

  (maximum difference between any attribute of the objects)



Distance as a Metric

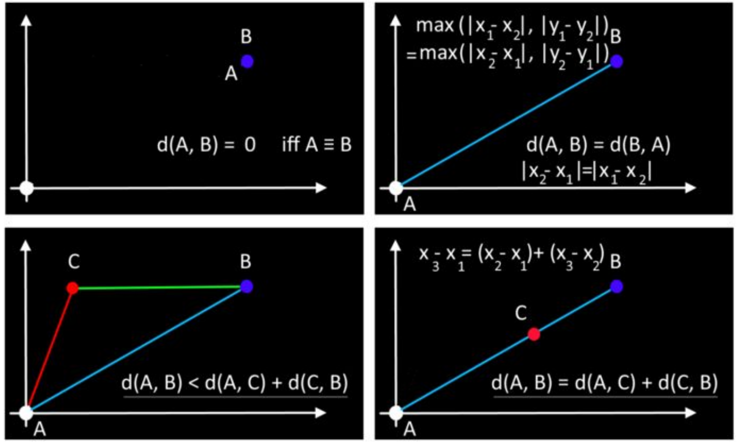
If $d(x, y)$ is the distance between two points (in our case, data objects), x and y, then the following properties hold.

1. **Positivity**
   (a) $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$,
   (b) $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$.
2. **Symmetry**
   $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$.
3. **Triangle Inequality**
   $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.

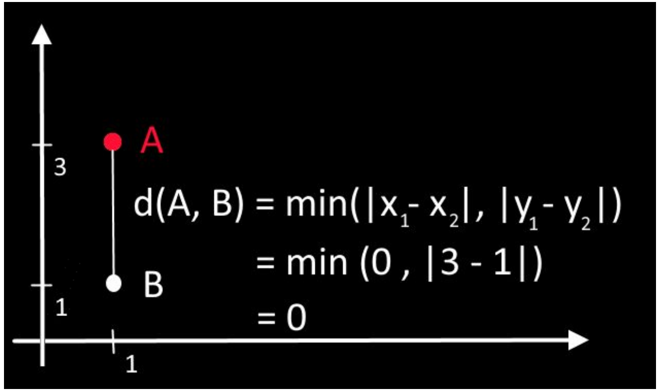
Measures that satisfy all three properties are known as ***metrics***.

Distance as a Metric: Examples

Is supreme distance a metric?



What about a *minimum difference* in any attribute of the two objects (i.e. $r = -\infty$) ?



A counterexample where positivity condition violates

<u>Ques</u>. If A and B are two sets, then which of the following is a metric?

- d (A, B) = | A – B |
- d (A, B) = | A ⊖ B |

Similarities between Data Objects

. . . that involve multiple attributes. . .

Similarities between Data Objects

For similarities, the triangle inequality (or the analogous property) typically does not hold, but symmetry and positivity typically do.

If $s(x, y)$ is the similarity between two points (in our case, data objects), x and y, then the following properties hold.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ only if $\mathbf{x} = \mathbf{y}$. $(0 \leq s \leq 1)$
2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

Simple Matching Coefficient (SMC)

Let **x** and **y** be two objects that consist of **n** binary attributes. The comparison of two such objects, i.e., two binary vectors, leads to the following four quantities (frequencies).

f_{00}, the number of attributes where **x** = 0 and **y** = 0
f_{01}, the number of attributes where **x** = 0 and **y** = 1
f_{10}, the number of attributes where **x** = 1 and **y** = 0
f_{11}, the number of attributes where **x** = 1 and **y** = 1

$$SMC(x,y) = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

In case of transaction data or documents (i.e. where binary representation is sparse), SMC *is not a correct measure.*

Jaccard Coefficient

Let **x** and **y** be two objects that consist of **n** binary attributes. The Jaccard coefficient only captures the asymmetric binary attributes.

$$J(x, y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Jaccard Coefficient

$$\mathbf{x} = (1,0,0,0,0,0,0,0,0,0)$$
$$\mathbf{y} = (0,0,0,0,0,0,1,0,0,1)$$

$f_{01} = 2$	the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
$f_{10} = 1$	the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
$f_{00} = 7$	the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
$f_{11} = 0$	the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$$

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0$$

In case of documents (i.e. non-binary sparse representation), Jaccard is not the right choice.

Cosine Similarity

Let x and y be two objects (e.g. documents) that consist of n non-binary attributes. The cosine similarity between the two vectors is defined as below.

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \, \|\mathbf{y}\|} = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|},$$

Where,

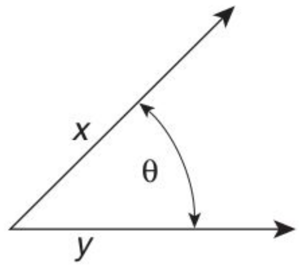
$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^{n} x_k y_k = \mathbf{x}'\mathbf{y},$$

$$\|\mathbf{x}\| = \sqrt{\textstyle\sum_{k=1}^{n} x_k^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}'\mathbf{x}}.$$

Cosine Similarity

It is a measure of the (cosine of the) angle between x and y.

$$\cos(\mathbf{x}, \mathbf{y}) = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle = \langle \mathbf{x}', \mathbf{y}' \rangle,$$



Dividing x and y by their lengths normalizes them to have a length of 1.

- Thus, if the cosine similarity is 1, the angle between x and y is 0°, and x and y are the same except for length.
- If the cosine similarity is 0, then the angle between x and y is 90°, and they do not share any terms (words).

Cosine Similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle = \langle \mathbf{x}', \mathbf{y}' \rangle,$$

$$\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$
$$\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = 3 \times 1 + 2 \times 0 + 0 \times 0 + 5 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 2 = 5$$
$$\|\mathbf{x}\| = \sqrt{3 \times 3 + 2 \times 2 + 0 \times 0 + 5 \times 5 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.48$$
$$\|\mathbf{y}\| = \sqrt{1 \times 1 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 2 \times 2} = 2.45$$
$$\mathbf{cos(x, y) = 0.31}$$

The inner product depends only on components that are non-zero in both vectors (i.e. asymmetric attributes).

Correlation

It measures the strength and direction of a ***linear and monotonic relationship*** between two *sets of values (or attributes)* that are *observed together (i.e., paired values)*.

Pearson correlation between two variables x and y is defined as below.

$$\operatorname{corr}(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{covariance}(\mathbf{x}, \mathbf{y})}{\operatorname{standard\_deviation}(\mathbf{x}) \times \operatorname{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \ s_y},$$

$$\operatorname{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$\operatorname{standard\_deviation}(\mathbf{x}) \quad = \quad s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\operatorname{standard\_deviation}(\mathbf{y}) \quad = \quad s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

Correlation

Correlation is always in the range −1 to 1.

- A correlation of 1 (−1) means that x and y have a perfect positive (negative) linear relationship;

  i.e., $x = ay + b$, where a and b are constants.

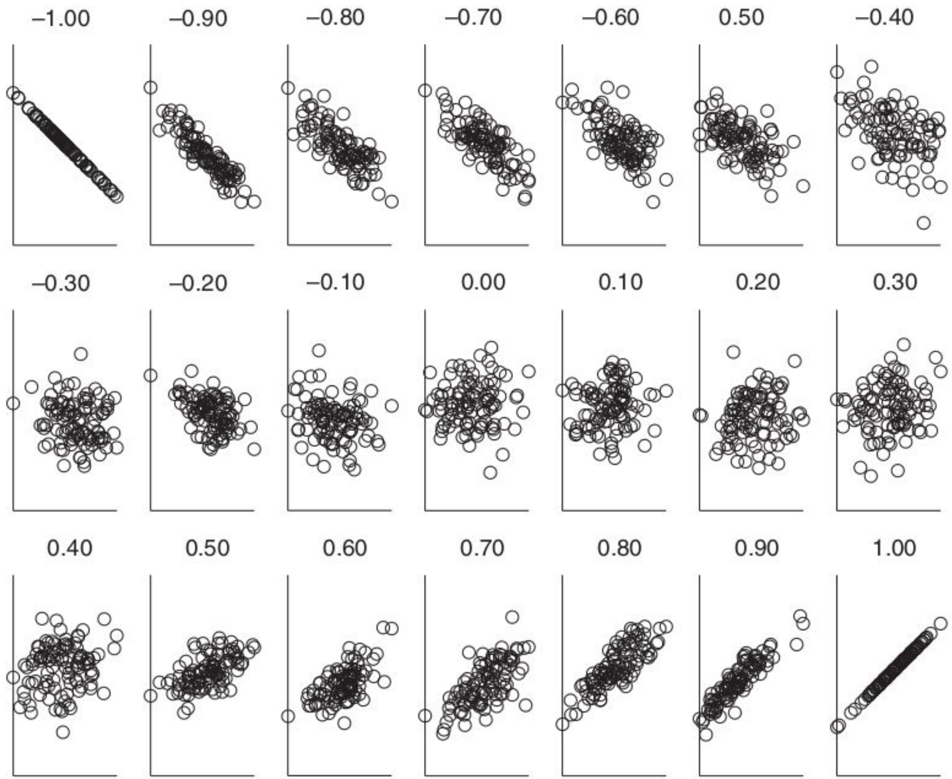
$$\mathbf{x} = (-3,\ 6,\ 0,\ 3, -6)$$
$$\mathbf{y} = (\ 1, -2,\ 0, -1,\ 2)$$
$$corr(\mathbf{x}, \mathbf{y}) = -1 \qquad x_k = -3y_k$$

$$\mathbf{x} = (3, 6, 0, 3, 6)$$
$$\mathbf{y} = (1, 2, 0, 1, 2)$$
$$corr(\mathbf{x}, \mathbf{y}) = 1 \qquad x_k = 3y_k$$

- If the correlation is 0, then there is *no linear relationship* between the two sets of values. However, nonlinear relationships can still exist.
  - For example, $y = x^2$, but their correlation is 0.

$$\mathbf{x} = (-3, -2, -1,\ 0,\ 1,\ 2,\ 3)$$
$$\mathbf{y} = (\ 9,\ 4,\ 1,\ 0,\ 1,\ 4,\ 9)$$

Correlation



Correlation vs. Covariance

Covariance signifies the *direction of the linear relationship* between the two variables. i.e.,

- If the variables are *directly proportional* or *inversely proportional* to each other.

  (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

Correlation explains the change in one variable leads the amount of proportion change in the second variable.

- It measures both the strength and the direction of the relationship between two variables.

Invariance to Transformation

Invariance to Transformations

A proximity measure is considered to be *invariant* to a data transformation if its value remains unchanged even after performing the transformation.

Property	Cosine	Correlation	Minkowski Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

Invariance to Transformations

Consider the following two vectors $\mathbf{x}$ and $\mathbf{y}$ with seven numeric attributes.

$$\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$$
$$\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$$

$$\mathbf{y_s} = 2 \times \mathbf{y} = (2, 4, 6, 8, 0, 0, 0)$$
$$\mathbf{y_t} = \mathbf{y} + 5 = (6, 7, 8, 9, 5, 5, 5)$$

Measure	$(\mathbf{x}, \mathbf{y})$	$(\mathbf{x}, \mathbf{y_s})$	$(\mathbf{x}, \mathbf{y_t})$
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Invariance to Transformations

Consider the document space -

- $\mathbf{x}$ and $\mathbf{y}$ both are document vectors representing term frequencies
- $\mathbf{y}_s$ denotes the scaled version of $\mathbf{y}$ with the same term distribution; i.e., just a larger document
- $\mathbf{y}_t$ denotes a different document with large number of words with non-zero frequency that do not occur in $\mathbf{y}$.

So, which similarity measure will be the ideal choice?

Invariance to Transformations

Consider that

- **x** represents a location's temperature measured on the Celsius scale for seven days.

- Let **y**, $\mathbf{y}_s$, and $\mathbf{y}_t$ be the temperatures measured on those days at a different location, but using three different measurement scales.

So, which similarity measure will be the ideal choice?

Hint: Different units of temperature have different offsets (e.g., Celsius and Kelvin) and different scaling factors (e.g., Celsius and Fahrenheit).

Invariance to Transformations

Consider a scenario where

- $\mathbf{x}$ represents the amount of precipitation (in *cm*) measured at seven locations.
- Let $\mathbf{y}$, $\mathbf{y}_s$, and $\mathbf{y}_t$ be estimates of the precipitation at these locations, which are predicted using three different models.
- We would like to choose a model that accurately reconstructs the measurements in $\mathbf{x}$ without making any error.

So, which proximity measure will be the ideal choice?

Mutual Information (MI)

Given that the values come in pairs, MI shows how much information one set of values provides about another.

It is used when a *nonlinear relationship* is suspected between the pairs of values.

- If the two sets of values are independent, then their MI is 0.
- If the two sets of values are completely dependent, then they have maximum MI.
- MI does not have a maximum value, but can be normalized to [0, 1].

Mutual Information (MI)

Let X can take m distinct values, $u_1, u_2, \ldots, u_m$ and Y can take n distinct values, $v_1, v_2, \ldots, v_n$.

Then their individual and joint entropy can be defined in terms of the probabilities of each value and pair of values as follows:

$$H(X) = -\sum_{j=1}^{m} P(X = u_j) \log_2 P(X = u_j)$$

$$H(Y) = -\sum_{k=1}^{n} P(Y = v_k) \log_2 P(Y = v_k)$$

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} P(X = u_j, Y = v_k) \log_2 P(X = u_j, Y = v_k)$$

The mutual information of X and Y can now be defined straightforwardly:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

One way to normalize MI is to divide it by $log_2(min(m, n))$.

Mutual Information (MI): Example

Suppose $y = x^2$, and their correlation is 0.

$$\mathbf{x} = (-3, -2, -1, \ 0, \ 1, \ 2, \ 3)$$
$$\mathbf{y} = (\ 9, \ \ 4, \ \ 1, \ 0, \ 1, \ 4, \ 9)$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

= (2.8074 + 1.9502) - 2.8074

= 1.9502

On normalizing the value,,

$= 1.9502 / \log_2(4)) = 0.9751$

x_j	$P(\mathbf{x} = x_j)$	$-P(\mathbf{x} = x_j) \log_2 P(\mathbf{x} = x_j)$
-3	1/7	0.4011
-2	1/7	0.4011
-1	1/7	0.4011
0	1/7	0.4011
1	1/7	0.4011
2	1/7	0.4011
3	1/7	0.4011
$H(\mathbf{x})$		2.8074

y_k	$P(\mathbf{y} = y_k)$	$-P(\mathbf{y} = y_k) \log_2(P(\mathbf{y} = y_k)$
9	2/7	0.5164
4	2/7	0.5164
1	2/7	0.5164
0	1/7	0.4011
$H(\mathbf{y})$		1.9502

x_j	y_k	$P(\mathbf{x} = x_j, \mathbf{y} = x_k)$	$-P(\mathbf{x} = x_j, \mathbf{y} = x_k) \log_2 P(\mathbf{x} = x_j, \mathbf{y} = x_k)$
-3	9	1/7	0.4011
-2	4	1/7	0.4011
-1	1	1/7	0.4011
0	0	1/7	0.4011
1	1	1/7	0.4011
2	4	1/7	0.4011
3	9	1/7	0.4011
$H(\mathbf{x}, \mathbf{y})$			2.8074

Issues related to Proximity Measures

There are a few important issues in proximity calculation:

- how to handle the case in which attributes have different scales and/or are correlated,
- how to calculate proximity between objects that are composed of different types of attributes, e.g., quantitative and qualitative, and
- how to handle proximity calculations when attributes have different weights; i.e., when not all attributes contribute equally to the proximity of objects.

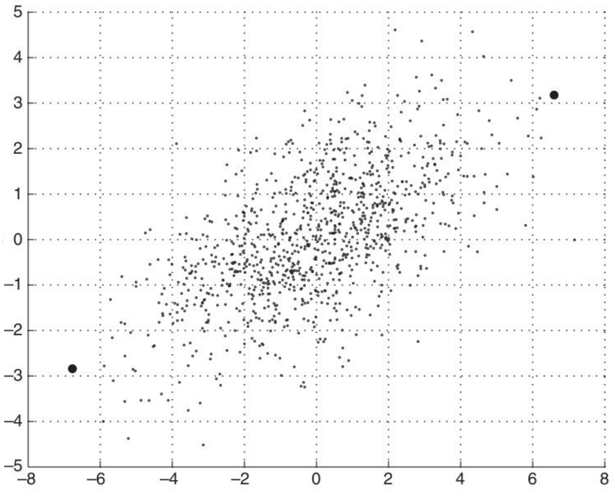
Issues related to Proximity Measures

- *how to handle the case in which attributes have different scales and/or are correlated,*
    - If the attributes are relatively uncorrelated, but have different ranges, then standardizing the variables is sufficient.
    - If the attributes are correlated and have different ranges of values, the *Mahalanobis distance* is useful.

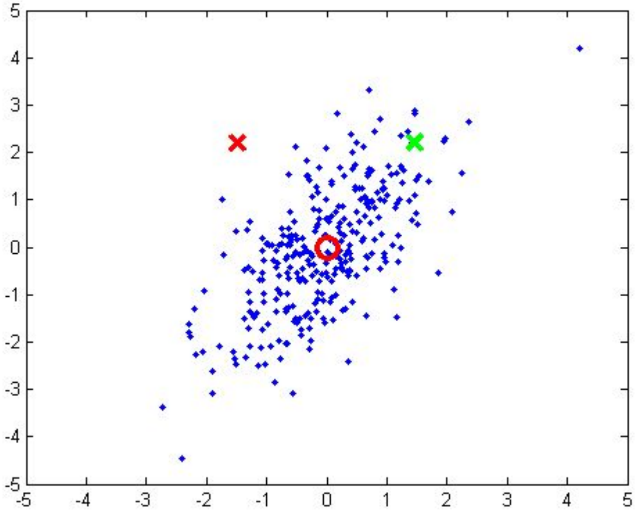
$$\text{Mahalanobis}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})},$$

Here, Σ^{-1} is the inverse of the covariance matrix of the data.

Issues related to Proximity Measures



$d_{eucl} = 14.7; \quad d_{mahl} = 6$



$d^r_{eucl} = d^g_{eucl}; \quad d^r_{mahl} = 4.12, \quad d^g_{mahl} = 2$

We can also use PCA to remove correlation among attributes. It transforms the data into orthogonal principal components. We will see it while studying dimensionality reduction.

Issues related to Proximity Measures

- *how to calculate proximity between objects that are composed of different types of attributes, e.g., quantitative and qualitative,*

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range $[0, 1]$.

2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is an asymmetric attribute and} \\ & \text{both objects have a value of 0, or if one of the objects} \\ & \text{has a missing value for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3: Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Issues related to Proximity Measures

- *how to handle proximity calculations when attributes have different weights; i.e., when not all attributes contribute equally to the proximity of objects.*

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k-1}^{n} w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} w_k \delta_k}.$$

Next lecture

Data Preprocessing

8th August 2023