Title:	Proximity Measures
Course:	Machine Learning
Instructor:	Claudio Sartori
Date:	
Master:	Data Science and Business Analytics
Academic Year:	2022/2023

BOLOGNA BUSINESS SCHOOL

Alma Mater Studiorum Università di Bologna

Similarity and dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity and Dissimilarity by Attribute type

p and q are the values of an attribute for two data objects

Attribute type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 \text{ if } p = q \\ 1 \text{ if } p \neq q \end{cases}$	$s = \begin{cases} 1 \text{ if } p = q \\ 0 \text{ if } p \neq q \end{cases}$
Ordinal Values mapped to integers 0 to V-1	$d = \frac{\|p-q\|}{V-1}$	$s = 1 - \frac{\|p-q\|}{V-1}$
Interval or Ratio	$d = \|p - q\|$	$s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min(d)}{\max(d) - \min(d)}$

Euclidean distance – L_2

$$\text{dist} = \sqrt{\sum_{d=1}^{D} (p_d - q_d)^2}$$

- Where D is the number of dimensions (attributes) and p_d and q_d are, respectively, the d-th attributes (components) of data objects p and q
- Standardization/Rescaling is necessary if scales differ

Minkowski distance – L_r

$$\text{dist} = \left(\sum_{d=1}^{D} |p_d - q_d|^r \right)^{\frac{1}{r}}$$

- Where D is the number of dimensions (attributes) and p_d and q_d are, respectively, the d-th attributes (components) of data objects p and q
- Standardization/Rescaling is necessary if scales differ
- r is a *parameter* which is chosen depending on the data set and the application

Minkowski distance – Cases

$r = 1$ also named *city block*, *Manhattan*, L_1 norm

- it is the best way to discriminate between zero distance and *near zero* distance
- a ϵ change on any coordinate causes a ϵ change in the distance
- works better than euclidean in very high dimensional spaces

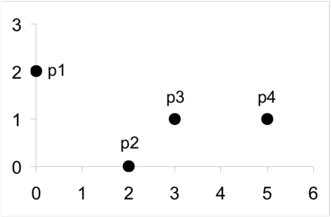
$r = 2$ euclidean, L_2 norm

$r = \infty$ also named Chebyshev, *supremum*, L_{max} norm, L_∞ norm

- considers only the dimension where the difference is maximum
- provides a simplified evaluation, disregarding the dimensions with lower differences

$$\mathrm{dist}_\infty = \lim_{r \to \infty} \left(\sum_{d=1}^{D} |p_d - q_d|^r \right)^{\frac{1}{r}} = \max_d |p_d - q_d|$$

Minkowski distances – Example



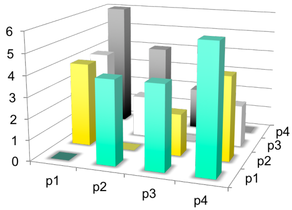
point	*x*	*y*
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L_1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

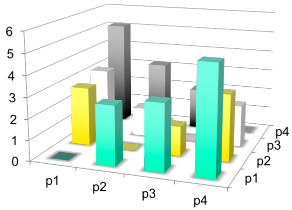
L_2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

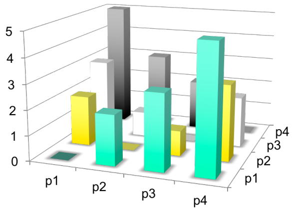
Comparison



L_1



L_2



L_∞

Mahalanobis Distance

Optional

- Considers data distribution
- The Mahlanobis distance between two points p and q decreases if, keeping the same euclidean distance, the segment connecting the points is stretched along a direction of greater variation of data
- The distribution is described by the covariance matrix of the data set

$$\Sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$\mathrm{dist}_m = \sqrt{(p-q)\Sigma^{-1}(p-q)^T}$$

Mahalanobis Distance – Example

OPTIONAL

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

$$A = (0.5, 0.5)$$

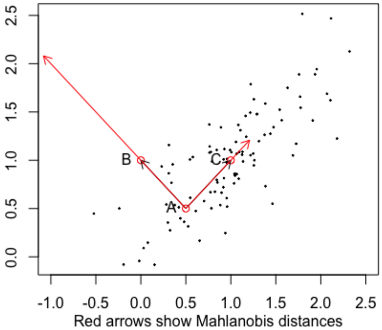
$$B = (0, 1)$$

$$C = (1, 1)$$

The euclidean distances AB and AC are the same

$$\mathrm{dist}_m(A, B) = 2.236068$$

$$\mathrm{dist}_m(A, C) = 1$$



Covariance matrix

OPTIONAL

- Variation of pairs of random variables
- The summation is over all the observations
- The main diagonal contains the variances
- The values are positive if the two variables grow together
- If the matrix is diagonal the variables are non–correlated
- If the variables are standardised the diagonal contains “one”
- If the variables are standardised and non correlated, the matrix is the identity and the Mahalanobis distance is the same as the euclidean

Common properties of a distance

1. Positive definiteness: $\text{Dist}(p, q) \geqslant 0 \; \forall p, q$
   and $\text{Dist}(p, q) = 0$ if and only if $p = q$
2. Symmetry: $\text{Dist}(p, q) = \text{Dist}(q, p)$
3. Triangle inequality: $\text{Dist}(p, q) \leqslant \text{Dist}(p, r) + \text{Dist}(r, q) \forall p, q, r$

A distance function satisfying all the properties above is called a metric

Common properties of a Similarity

1. $\text{Sim}(p, q) = 1$ only if $p = q$
2. $\text{Sim}(p, q) = \text{Sim}(q, p)$

Similarity between binary vectors

- Consider the counts below

  M_{00} the number of attributes where p is 0 and q is 0
  M_{01} the number of attributes where p is 0 and q is 1
  M_{10} the number of attributes where p is 1 and q is 0
  M_{11} the number of attributes where p is 1 and q is 1

- Simple Matching Coefficient

$$\mathrm{SMC} = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard Coefficient

$$\mathrm{JC} = \frac{\text{number of 11 matches}}{\text{number of non–both–zero attributes}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Cosine similarity

- It is the cosine of the angle between two vectors

$$\cos(p, q) = \frac{p \cdot q}{\|p\| \|q\|}$$

- Example

$$
\begin{aligned}
p &= 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0 \\
q &= 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2 \\
p \cdot q &= 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5 \\
\|p\| &= \sqrt{3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0} = 6.481 \\
\|q\| &= \sqrt{1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2} = 2.245 \\
\cos(p, q) &= .3150
\end{aligned}
$$

Extended Jaccard Coefficient (Tanimoto) OPTIONAL

- Variation of Jaccard for continuous or count attributes
  - reduces to Jaccard for binary attributes

$$\mathrm{T}(p, q) = \frac{p \cdot q}{\|p\|^2 + \|q\|^2 - p \cdot q}$$

Choose the right proximity measure

It depends on data

- Dense, continuous
  - a metric measure, such as the euclidean distance
- Sparse, asymmetric data
  - cosine, jaccard, extended jaccard