Similarity measures and standardization for gene expression data clustering

Giorgio Valentini

e-mail: valentini@dsi.unimi.it

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano

Metrics and distances

A metric, d, satisfies the following 5 properties:

(i) Non negativity $d(x, y) \geq 0$

(ii) symmetry $d(x, y) = d(y, x)$

(iii) identity $d(x, x) = 0$

(iv) definiteness $d(x, y) = 0$ if and only if $x = y$

(v) triangle inequality $d(x, y) + d(y, z) \geq d(x, z)$.

We can also consider pairwise distances, which are functions that are required to satisfy the first three properties only.

We will refer to distances which include metrics and only mention metrics when the behavior of interest is specic to them.

Similarity functions

A similarity function S is more loosely defined and satisfies the three following properties:

(i) Non negativity $S(x, y) \geq 0$;

(ii) symmetry $S(x, y) = S(y, x)$;

(iii) The more similar the objects a and b, the greater $S(x, y)$.

Measuring Similarity

- Similarity function *sim* or dissimilarity/distance function *dist*
- Similarity function *sim*(x, y):
  Large value: x and *y* are more similar, small value: x and y are less similar
- Often $0 \leq \text{sim}(x, y) \leq 1$
- Dissimialrity/Distance function *dist*(x, y):
  Small value: x and *y* are more similar, large value: x and y are less similar
- Dual properties for a similarity function
- Definition of a similarity or dissimilarity function is application dependent

Minkowski distances

- One group of popular distance measures for interval-scaled variables are **Minkowski distances**

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

where $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ are two p-dimensional data objects (e.g. vectors of gene expression data), and q is a positive integer

Example: Manhattan and Euclidean distances

- If $q = 1$, the distance measure is **Manhattan (**or **city block) distance**

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

- If $q = 2$, the distance measure is **Euclidean distance**

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

Minkowski distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = (\sum_{k=1}^{n} | x_k - y_k |^r)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the kth attributes (components) of data objects x and y (e.g. gene expression levels corresponding to two different patients)

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors.
- $r = 2$. Euclidean distance.
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_{∞} norm) distance.
  - This is the maximum difference between any component of the vectors.
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

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

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	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

Distance Matrix

Cosine Similarity

•If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where • indicates vector dot product and $\| d \|$ is the length of vector d.

•Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\|$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

$\|d_2\|$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

cos(d_1, d_2) = .3150

Correlation

- Correlation measure the linear relationship between objects.
- To compute correlation, we standardize data objects, x and y, and then take the dot product.

$$x'_k = (x_k - mean(x_k)) / std(x)$$

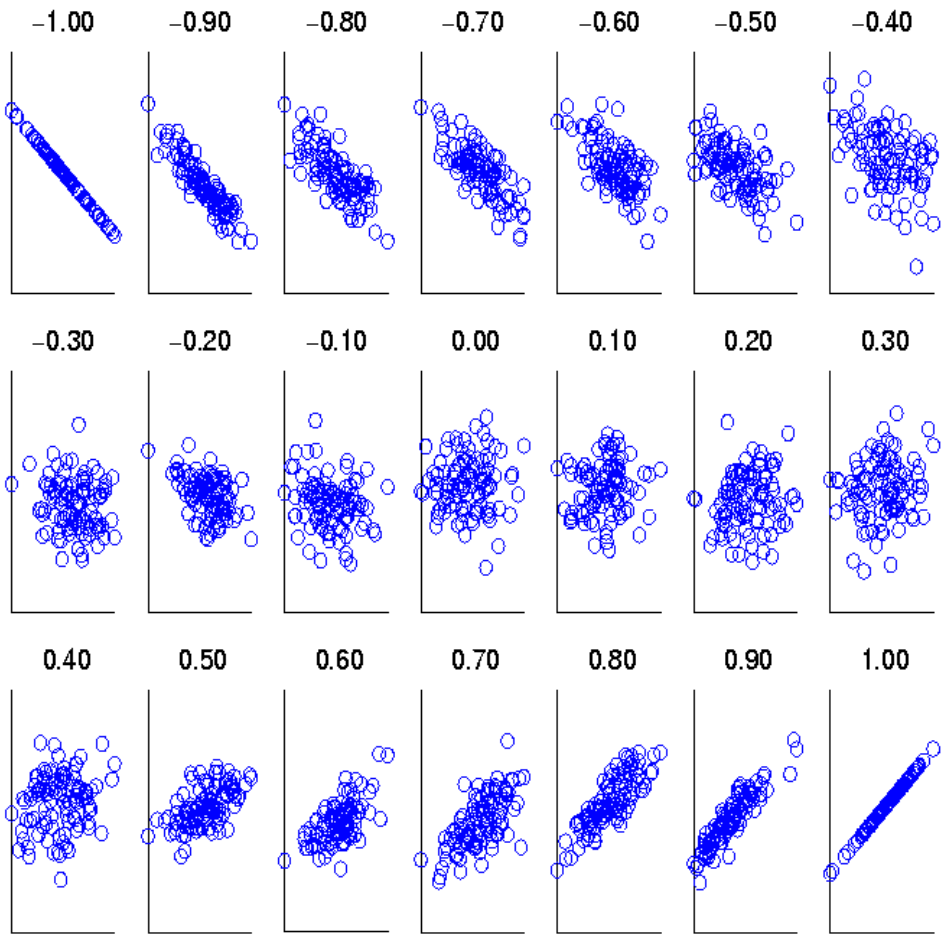
$$y'_k = (y_k - mean(y_k)) / std(y)$$

$$correlation(x, y) = r(x, y) = x' \bullet y' \qquad -1 \le r(x, y) \le +1$$

- A distance (dissimilarity measure) can be easily obtained from the correlation:

$$d(x,y) = (1 - r(x,y))/2$$

Visually evaluating correlation

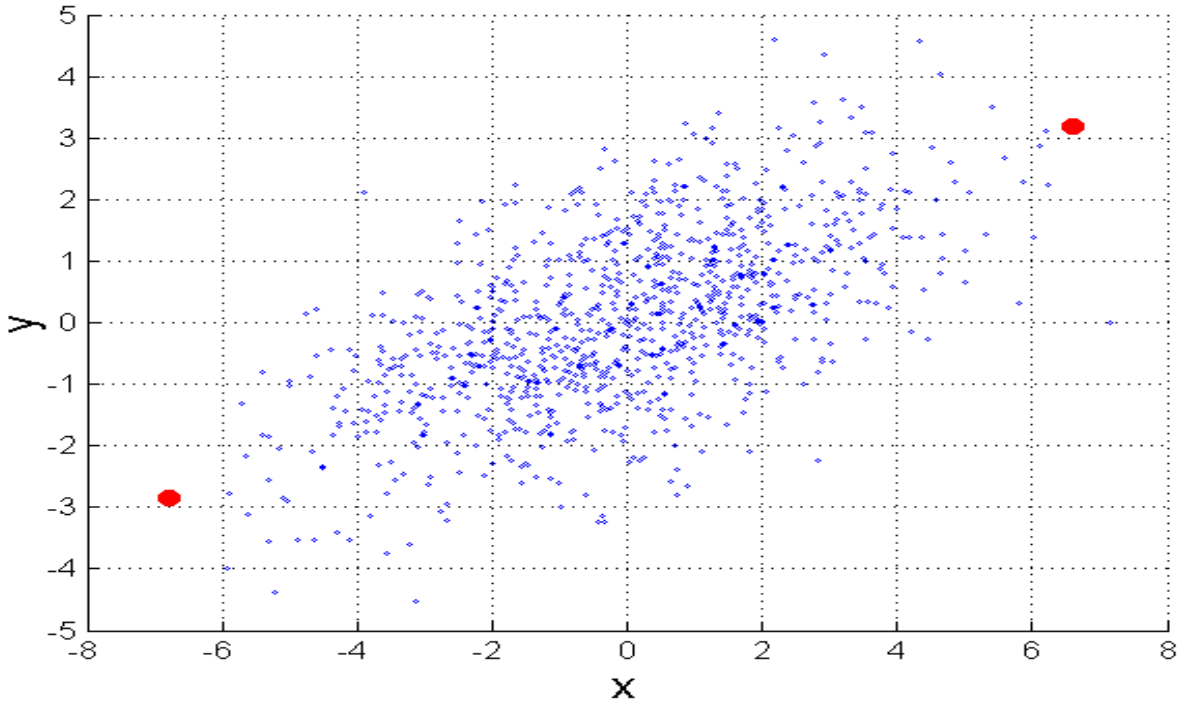


Scatter plots showing the similarity from –1 to 1.

Mahalanobis Distance

$$mahalanobis(x, y) = \sqrt{(x - y)\Sigma^{-1}(x - y)^T}$$

Σ is the covariance matrix. If Σ is the identity matrix Mahalonobis distance is reduced to the Euclidean distance.



For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Metrics and distances.

Name	Formula
Euclidean metric	$d_E(\mathbf{x}_i, \mathbf{x}_j) = \{\sum_g w_g (x_{gi} - x_{gj})^2\}^{1/2}$
Unstandardized	$w_g = 1$
Standardized by s.d.	$w_g = 1/s_g^2$.
(Karl Pearson distance)	
Standardized by range	$w_g = 1/R_g^2$.
Mahalanobis metric	$d_{Ml}(\mathbf{x}_i, \mathbf{x}_j) = \{(\mathbf{x_i} - \mathbf{x_j}) S^{-1} (\mathbf{x_i} - \mathbf{x_j})'\}^{1/2}$ $= \{\sum_g \sum_{g'} s^{-1}_{gg'} (x_{gi} - x_{gj})(x_{g'i} - x_{g'j})\}^{1/2}$ where $S = (s_{gg'})$ is any $G \times G$ positive definite matrix, usually the sample covariance matrix of the variables. When the matrix is the identity, this reduces to the unstandardized Euclidean distance.
Manhattan metric	$d_{Mn}(\mathbf{x}_i, \mathbf{x}_j) = \sum_g w_g \lvert x_{gi} - x_{gj} \rvert$
Minkowski metric	$d_{Mk}(\mathbf{x}_i, \mathbf{x}_j) = \{\sum_g w_g \lvert x_{gi} - x_{gj} \rvert^{\lambda}\}^{1/\lambda}$, $\lambda \geq 1$. $\lambda = 1$: Manhattan distance $\lambda = 2$: Euclidean distance
Canberra metric	$d_C(\mathbf{x}_i, \mathbf{x}_j) = \sum_g \frac{\lvert x_{gi} - x_{gj} \rvert}{(x_{gi} + x_{gj})}$
One minus Pearson correlation	$d_{corr}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\sum_g (x_{gi} - \bar{x}_{.i})(x_{gj} - \bar{x}_{.j})}{\{\sum_g (x_{gi} - \bar{x}_{.i})^2\}^{1/2} \{\sum_g (x_{gj} - \bar{x}_{.j})^2\}^{1/2}}$

The formulae refer to distances between observations (arrays).

Standardization (1)

- Standardization of the features is an important issue when considering distances between objects.
- Samples or genes are assigned to classes on the basis of their distance from other objects.
- The distance or similarity function that is used generally has a large effect on the performance of the classification or clustering procedure.
- The distance function and its behavior are intimately related to the scale on which measurements are made.
- There are no objective methods for dealing with this problem.
- The solution is generally problem specic.

Standardization (2)

A common type of data transformation for continuous measurements is standardization.

For microarray data both genes and/or observations (arrays) can be standardized. Which of the two should be carried out is dependent upon whether samples or genes are being clustered or classied.

Standardizing genes: $$x_{ga} \leftarrow (x_{ga} - \bar{x}_{g.}) / \sigma_{g.}$$

so that each gene has mean zero and unit variance across arrays.

Standardizing arrays: $$x_{ga} \leftarrow (x_{ga} - \bar{x}_{.a}) / \sigma_{.a}$$

so that each array has mean zero and unit variance across genes.

Standardizing genes

Gene standardization in some sense puts all genes on an equal footing and weights them equally in the classication or clustering. Common standardization procedures are:

1. $$x_{ga} = \frac{(x_{ga} - \bar{x}_{g.})}{\sigma_{g.}}$$

where $\bar{x}_{g.}$ and $\bar{\sigma}_{g.}$ denote respectively the average and standard deviation of gene *g*'s expression levels across the *n* arrays.

2. $$x_{ga} = \frac{(x_{ga} - m_{g.})}{mad_{g.}}$$

where $m_{g.}$ and $mad_{g.}$ denote respectively the median and the median absolute deviation (MAD) of gene *g*'s expression levels across the *n* arrays.

3. $$x_{ga} = \frac{(x_{ga} - x_{g(1)})}{x_{g(n)} - x_{g(1)}}$$

where $x_{g(j)}$ denote the ordered expression levels for gene g,

Standardizing arrays

Standardization of arrays can be viewed as part of the normalization step.

Analogously to standardizing genes, we can normalize w.r.t to:

1. The average and standard deviation
2. The median and median absolute deviation
3. The fist ranked and the difference between the last and first ranked

but the normalization is performed across genes of the same array.

It is consistent with the common practice of using the correlation between the gene expression profiles of two mRNA samples to measure their similarity.

In practice, it is better to apply adaptive and robust normalization methods which correct for intensity, spatial, and other types of bias using robust local regression (see previous lecture on pre-processing and normalization).