GENIOMHE

Data-mining and Machine Learning

by Samuel Ortion

Prof.: Farida Zerhaoui

Fall 2023

Contents

1 Unsupervised Learning

Definition 1: Precision Medicine

Design of treatment for a given patient, based on genomic data.

Definition 2: Hierarchical clustering

Gene expression time series: look for genes with similar expression footprint.

Representation of data

- Tables;
- Trees / Graphs;
- Time series...

1.1 Distances and Similarities

Property 1 (Distance)**.** ***non-negativity*** $d(i, j) \geq 0$

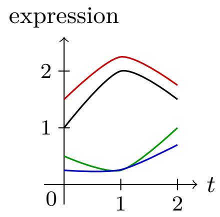

Figure 1.1 Example of gene expression time series

isolation $d(i,i) = 0$

symmetry $d(i,j) = d(j,i)$

triangular inequality $d(i,j) \leq d(i,h) + d(h,j)$

π **Definition 3:** Dissimilarity

Distance without triangular inequality.

π **Definition 4:** Similarity

Function s from $X \times X$ to $\mathbb{R}_+$ such that:

1. s is symmetric: $(x,y) \in X \times X; s(x,y) = s(y,x)$
2. $(x,y) \in X \times X; s(x,x) = s(y,y) > s(x,y)$.

Exercise 1:
Let $d(x,y)$ be the distance, $d(x,y) \in [0, +\infty[$.
What should be the similarity measure $S(x,y) = f(d(x,y))$ that satisfies the following property:

$$(x,y) \in X \times X \mid S(x,y) > S(x,y)$$

having $S(x,y) \leq M$, $S(x,y) \in]0,M]$. $\quad d(x,y) \geq 0 \, \forall (x,y)$

$$S(x,y) = \frac{M}{d(x,y)+1} \tag{1.1}$$

In eq. (1.1), $S(x,y)$ ranges from 0 to M.

$$\lim_{n\to\infty} \frac{M}{n+1} = 0 \qquad \lim_{n\to 0} \frac{M}{n+1} = M \tag{1.2}$$

1.2 Data Representation

Data matrix

Distance matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \ddots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

π **Definition 5:** Minkowski distance

$$L_p(x,y) = \left(|x_1 - y_1|^p + |x_2 - y_2|^p + \ldots + |x_d - y_d|^p\right)^{1/p} = \left(\sum_{i=1}^{d} (x_i - y_i)^p\right)^{1/p}$$

	s_1	s_2
p_1	0	1
p_2	1	0
p_3	3	2

Table 1.1 Example data matrix: 2 symptoms for 3 patients.

Definition 5 continued

where p is a positive integer.

Definition 6: Manhattan distance

$$L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$$

Definition 7: Euclidian distance

Let A and B be two points, with (x_A, y_A) and (x_B, y_B) their respective coordinates,

If $p = 2$, L_2 is the Euclidian distance:

Definition 8: Euclidian distance

$$d(x,y) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2| + \ldots + |x_d - y_d|^2}$$

We can add weights

1.2.1 K-means

The cost function is minimized:

$$Cost(C) \sum_{i=1}^{k} \ldots$$

Algorithm 1 K-means algorithm

```
Choose the number of clusters $k$.
Choose randomly $k$ means.
For each point, compute the distance between the point and each means. We allocate the point to the cluster represented by the clostest center.
We set each means to the center of the cluster, and reiterate.
```

Exercise 2:

We have six genes:

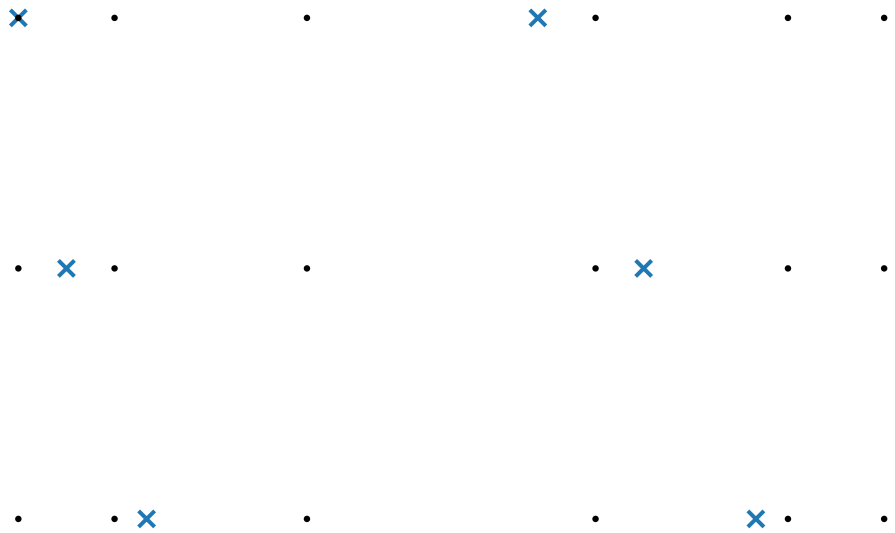

Figure 1.2 k-means states at each of the 3 steps

	g_1	g_2	g_3	g_4	g_5	g_6
$\times 10^{-2}$	10	12	9	15	17	18

Table 1.2 Sample values for six gene expressions.

With $k = 2$ and $m_1 = 10 \cdot 10^{-2}$ and $m_2 = 9 \cdot 10^{-2}$ the two initial randomly chosen means, run the k-means algorithm.