Lecture 12: Clustering

Dr. Roman V Belavkin

BIS3226

Contents

1 Metric Spaces **1**

2 Data as Vectors in Metric Spaces **3**

3 The Clustering Problem **4**

1 Metric Spaces

Metric Spaces

Let X be a set. How can we compare the elements of X?

Definition 1 (Metric)**.** is a function $d : X \times X \to \mathbb{R}$ that is

1. Non-negative: $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$.
2. Symmetric: $d(x, y) = d(y, x)$.
3. Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$.

Definition 2 (Metric space)**.** is a pair (X, d) — a set X with a metric d.

Example 3 (Discrete space). Let d be defined as

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

Metrics in Vector Spaces

- An m-dimensional real vector space is denoted $\mathbb{R}^m$
- Elements of $\mathbb{R}^m$ are called *vectors*.
- Each vector $\mathbf{x} \in \mathbb{R}^m$ is a point in the space represented by its coordinates:

$$\mathbf{x} = (x_1, \ldots, x_m)$$

- Each coordinate is a real number $x_i \in \mathbb{R}$.
- Note that coordinates are relative to a chosen *basis*.

Example 4. In $\mathbb{R}^1$ (one-dimensional space or a line) points are represented by just one number, such as $\mathbf{x} = (2)$ or $\mathbf{y} = (-1)$.

Example 5. In $\mathbb{R}^3$ (three-dimensional space) points are represented by three coordinates x_1, x_2 and x_3, such as $\mathbf{x} = (2, -1, 3)$.

Metrics on Vectors

- How can we compute the distance between different vectors?
- For two vectors $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$, we can compute the differences of their coordinates:

$$x_1 - y_1\,, \quad x_2 - y_2\,, \quad \cdots \quad x_m - y_m$$

- Computation of metrics in vector spaces often uses absolute values of the differences:

$$|x_1 - y_1|\,, \quad |x_2 - y_2|\,, \quad \cdots \quad |x_m - y_m|$$

Example 6 (Taxicab (Manhattan) distance).

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_m - y_m|$$

Euclidean Distance

Definition 7 (Euclidean distance)**.**

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \cdots + |x_m - y_m|^2}$$

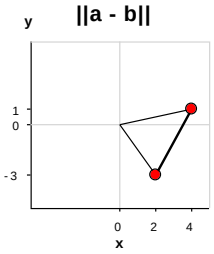
Remark 1 (Euclidean space)**.** *is vector space in which metric is given by Euclidean distance.*

Example 8. Let $\mathbf{x} = (2)$ and $\mathbf{y} = (-1)$ in $\mathbb{R}^1$. Then

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{(2+1)^2} = 3$$

Example 9. Let $\mathbf{x} = (2, -3)$ and $\mathbf{y} = (4, 1)$ in $\mathbb{R}^2$. Then

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\| &= \sqrt{|2-4|^2 + |-3-1|^2} \\ &= \sqrt{20} \approx 4.47 \end{aligned}$$



2 Data as Vectors in Metric Spaces

Data and Similarity

- A bank gathered information about its customers:

Case:	Age	Gender	M. Income (£ K)	M. Expenses (£ K)	Home owner	Credit score
1	21	0	2	1	0	3
2	18	1	1	2	0	1
3	50	1	6	2	1	5
4	23	0	3	1	1	4
5	40	1	3	2	0	2

- We may consider each variable (age, gender, income, etc) as a coordinate x_i and each case as a vector in an m-dimensional space.

- What does a distance between the vectors represent?
- How far should be similar cases from each other?
- Which of the cases 1, 2, 3 or 4 is the most similar to case 5?

Metric as a Measure of Dissimilarity

- The database corresponds to the following set of vectors:

$$\begin{aligned} \mathbf{v} &= (21, 0, 2, 1, 0, 3) \\ \mathbf{w} &= (18, 1, 1, 2, 0, 1) \\ \mathbf{x} &= (50, 1, 6, 2, 1, 5) \\ \mathbf{y} &= (23, 0, 3, 1, 1, 4) \\ \mathbf{z} &= (40, 1, 3, 2, 0, 2) \end{aligned}$$

- If there is a metric d, then we can find the distances from $\mathbf{z}$:

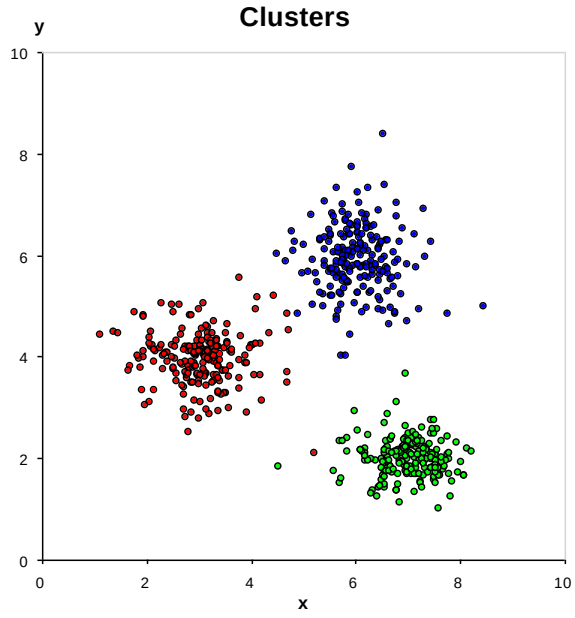
$$d(\mathbf{z}, \mathbf{v}), \ d(\mathbf{z}, \mathbf{w}), \ d(\mathbf{z}, \mathbf{x}), \ d(\mathbf{z}, \mathbf{y})$$

- The most similar to $\mathbf{z}$ is the 'closest' vector.

Remark 2. *The choice of metric is important, because generally different metrics will produce different results.*

3 The Clustering Problem

Clusters



- **Clusters** are groups of points.
- The groups can be based on similarity, such as closeness.
- 'Similar' data would correspond to points that are close to each other and would belong to the same group (cluster).

Definition 10 (Centroid)**.** is the centre of gravity of a cluster X_i, computed

as the average vector:

$$\mu_i = \bar{X}_i = \frac{\mathbf{x} + \cdots + \mathbf{z}}{n} = \Big(\frac{x_1 + \cdots + z_1}{n}, \ldots, \frac{x_m + \cdots + z_m}{n}\Big)$$

k-Means Clustering Algorithm

- Let X be a set of vectors in $\mathbb{R}^m$ (i.e. data)
- The goal of the k-means algorithm is to partition X into k clusters $X_1, \ldots, X_k$, represented by k centroids (means):

$$\mu_1\,,\ \mu_2\,,\quad \ldots\quad ,\ \mu_k$$

- The following is an outline of the k-means algorithm:
  1. Select the number of clusters k
  2. Define metric d on X
  3. Choose at random k vectors $\mu_1, \mu_2, \ldots, \mu_k$ in $\mathbb{R}^m$
  4. Repeat
     (a) Group $\mathbf{x} \in X$ into clusters $X_1, X_2, \ldots, X_k$ by computing $d(\mu_i, \mathbf{x})$.
     (b) Compute new $\mu_1, \mu_2, \ldots, \mu_k$ as centroids:
     $$\mu_1 = \bar{X}_1\,,\ \mu_2 = \bar{X}_2\,,\ \ldots\ ,\ \mu_k = \bar{X}_k$$
  5. Until finished.

Output of k-Means

- The values (coordinates) of k centroids:

$$\begin{aligned} \mu_1 &= (x_{11}, \ldots, x_{1m}) \\ \mu_2 &= (x_{21}, \ldots, x_{2m}) \\ &\cdots \\ \mu_k &= (x_{k1}, \ldots, x_{km}) \end{aligned}$$

- The partition of X: an assignment of each vector in X (data) to one of k clusters:

3 3 3 1 2 1 2 1 1 2 1

- Other information about the clusters (e.g. number of points, diameter).