hugomcp@di.ubi.pt, 2025/26

Machine Learning

Practical Sheet 8: Unsupervised Learning

1. Consider the following dataset (2-D points):

   A = (1,1), B = (1,2), C = (2,1), D = (5,5), E = (6,5), F = (5,6)

Use **c = 2** clusters. Initial cluster centers:

- $v_1^{(0)} = (1,1)$
- $v_2^{(0)} = (5,5)$

Perform **one iteration** of C-means (assignment then re-compute centroids), according to three distance metrics:

- Euclidean $d_2(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$
- Manhattan $d_1(x,y) = | x_1 - y_1 | + | x_2 - y_2 |$
- Chebyshev $d_\infty(x,y) = \max(| x_1 - y_1 |, | x_2 - y_2 |)$

For each metric do these steps:

a) Make a 6×2 table of distances $d(x_i, v_1^{(0)})$and $d(x_i, v_2^{(0)})$.

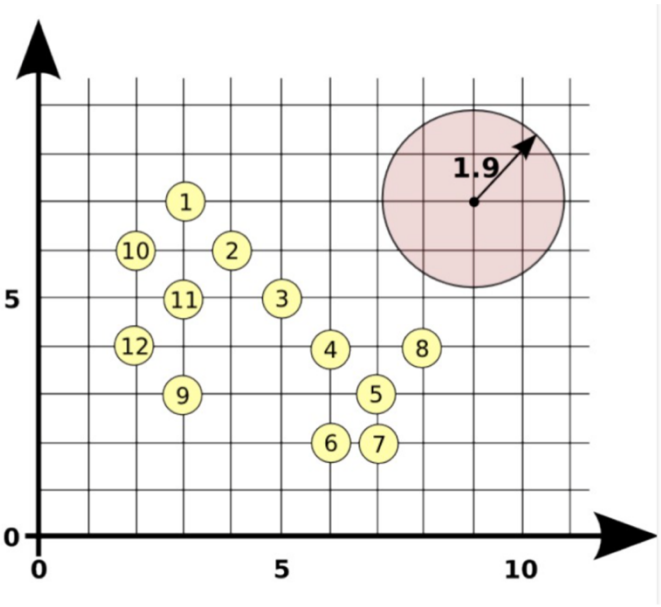
b) For each point assign it to the **closest** center (if tie, assign to cluster 1).

c) Compute new centers $v_j^{(1)}$as the arithmetic mean of points assigned to cluster j:

$$v_j^{(1)} = \frac{1}{n_j} \sum_{x_i \in C_j} x_i \, .$$

d) Report the assignments and $v_1^{(1)}, v_2^{(1)}$. Write a one-line comment: did the metric change the assignment or the updated centers?

2. Consider the dataset below. Apply the DBSCAN algorithm with $\varepsilon = 1.9$ and minPoints=3. Start by discriminate between “Core”, “Border” and “Noise” points. Finally, indicate the clusters obtained.



3. Consider the "AR.zip" dataset, available at the course web page. It contains 3,315 images, of 136 subjects, each one represented in the RGB color space and having dimensions 576 (rows) x 768 (columns). Create a "Python" that:

    a) Load the set of images;

    b) Obtains the PCA representation of this data, keeping 95% of the amount of information (variance)

    c) Cluster the data, using the standard Euclidean distance metric

    d) Repeat the clustering step, but this time using a metric that is appropriate for the PCA space (remember, that eigenvectors are sortered according to the amount of information of the original space they carry).

    e) Obtain a **SOM** manifold;

    f) Find the **SOM** topology that is more appropriate for distinguishing between...

        1. Identities

        2. Gender