Aprendizagem 2023

# Lab 9: Clustering

## Prof. Rui Henriques

## Practical exercises

**1.** Consider the following training data without and the cluster centres

| | $y_1$ | $y_2$ |
|---|---|---|
| $\mathbf{x}_1$ | 0 | 0 |
| $\mathbf{x}_2$ | 1 | 0 |
| $\mathbf{x}_3$ | 0 | 2 |
| $\mathbf{x}_4$ | 2 | 2 |

$$\boldsymbol{u}_1 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \qquad \boldsymbol{u}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

a) Compute the silhouette of observation $\mathbf{x}_1$, cluster $c_1$ and overall solution

Preserving the Euclidean distance assumption, let us compute the silhouette of $\boldsymbol{c_1}$:

$$s(\mathbf{x_1}) = 1 - \frac{a(\mathbf{x_1})}{b(\mathbf{x_1})} = 1 - \frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2}{\frac{1}{2}(\|\mathbf{x}_1 - \mathbf{x}_3\|_2 + \|\mathbf{x}_1 - \mathbf{x}_4\|_2)} = 1 - \frac{1}{2.4} = 0.58(3)$$

$$s(\mathbf{x_2}) = 1 - \frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_2}{\frac{1}{2}(\|\mathbf{x}_2 - \mathbf{x}_3\|_2 + \|\mathbf{x}_2 - \mathbf{x}_4\|_2)} = 1 - \frac{1}{\sqrt{5}} = 0.553$$

$$s(\mathbf{c_1}) = \frac{s(\mathbf{x_1}) + s(\mathbf{x_2})}{2} = \frac{0.58(3) + 0.553}{2} = 0.568$$

The silhouette of a solution is the average of cluster silhouettes.

$$s(\mathbf{x_3}) = 0.052, \qquad s(\mathbf{x_4}) = 0.225, \qquad s(\mathbf{c_2}) = 0.133$$

$$silhouette(C) = \frac{s(\mathbf{c_1}) + s(\mathbf{c_2})}{2} = \frac{0.568 + 0.133}{2} = 0.35$$

Silhouette level is moderate, thus there is some evidence for clusters to be cohesive and well-separated.

b) Knowing $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_4$ to be annotated as positive and $\mathbf{x}_3$ as negative (ground truth). Compute the purity of *k*-means against the given ground truth.

$$purity(C, L) = \frac{1}{n}\sum_{k=1}^{K} \max_j \left(|\boldsymbol{c}_k \cap \boldsymbol{l}_j|\right) = \frac{1}{4}(1 + 2) = \frac{3}{4}$$