Departamento de Matemática

Multivariate Analysis

Master in Eng. and Data Science & Master in Mathematics and Applications

2^{nd} Test - Part I
Duration: 45 minutes

1^{st} Semester – 2020/2021
04/02/2021 – 16:45

Please justify conveniently your answers

If the second letter of your first name is between "A" and "L" solve **Group I - Version A**, otherwise solve **Group I - Version B**.

Any wrong choice of Group I Version will not be classified.

Group I - Version A — 10.0 points

1. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t \in \mathbb{R}^p$, n_k the number of observations belonging to the k-th cluster, C_k, and $\bar{\boldsymbol{x}}_k = \sum_{\boldsymbol{x}_h \in C_k} \boldsymbol{x}_h / n_k$ be the centroid of the k-th cluster.

   (a) Prove that (2.5)

   $$\frac{1}{n_k} \sum_{\boldsymbol{x}_h, \boldsymbol{x}_{h'} \in C_k} \sum_{j=1}^{p} (x_{jh} - x_{h'j})^2 = 2 \sum_{\boldsymbol{x}_h \in C_k} \sum_{j=1}^{p} (x_{hj} - \bar{x}_j)^2.$$

   (b) What does the objective function of the K-means clustering algorithm (using Euclidean distance) intends to optimize and what is the relevance of the previous equality? (2.0)

2. Consider the following data set:

	x_{i1}	x_{i2}
$\boldsymbol{x}_1$	-2	-1
$\boldsymbol{x}_2$	-3	0
$\boldsymbol{x}_3$	-2	2
$\boldsymbol{x}_4$	-2	4
$\boldsymbol{x}_5$	1	2

   (a) Consider as an initial partition $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5\}$ and $C_2 = \{\boldsymbol{x}_2, \boldsymbol{x}_4\}$. Compute the centroid of each cluster. (1.0)

   (b) Obtain the first two steps of the K-means clustering algorithm, using Euclidean distance. (3.0)

   (c) Compute the $\boldsymbol{x}_1$ average silhouette, based on the initial partition. Give an interpretation to the obtained result. (1.5)

If the second letter of your first name is between “A” and “L” solve **Group I - Version A**, otherwise solve **Group I - Version B**.

Any wrong choice of Group I Version will not be classified.

Group I - Version B **10.0 points**

1. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t \in I\!R^p$, n_k the number of observations belonging to the k-th cluster, C_k, and $\bar{\boldsymbol{x}}_k = \sum_{\boldsymbol{x}_h \in C_k} \boldsymbol{x}_h / n_k$ be the centroid of the k-th cluster.

   (a) Prove that (2.5)

$$\frac{1}{n_k} \sum_{\boldsymbol{x}_h, \boldsymbol{x}_{h'} \in C_k} \sum_{j=1}^{p} (x_{jh} - x_{h'j})^2 = 2 \sum_{\boldsymbol{x}_h \in C_k} \sum_{j=1}^{p} (x_{hj} - \bar{x}_j)^2.$$

   (b) What does the objective function of the K-means clustering algorithm (using Euclidean distance) intends to optimize and what is the relevance of the previous equality? (2.0)

2. Consider the following data set:

	x_{i1}	x_{i2}
$\boldsymbol{x}_1$	5	-3
$\boldsymbol{x}_2$	-2	-4
$\boldsymbol{x}_3$	-2	2
$\boldsymbol{x}_4$	-2	4
$\boldsymbol{x}_5$	1	2

   (a) Consider as an initial partition $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5\}$ and $C_2 = \{\boldsymbol{x}_2, \boldsymbol{x}_4\}$. Compute the centroid of each cluster. (1.0)

   (b) Obtain the first two steps of the K-means clustering algorithm, using Euclidean distance, using as initial partition the one defined in Question 2a. (3.0)

   (c) Compute the $\boldsymbol{x}_1$ average silhouette, based on the initial partition. Give an interpretation to the obtained result. (1.5)