TÉCNICO LISBOA

Aprendizagem 2025

Homework IV

Deadline 23/10/2025 (Wednesday) 23:59 via Fenix as PDF

- Submit Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx_report.pdf with your report and Gxxx_notebook.ipynb with your notebook demo according to the suggested templates
- It is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is kept
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic or manual clearance, homework is nullified and IST guidelines apply for content sharers and consumers, irrespectively of the underlying intent
- Please consult the FAQ before posting questions to your faculty hosts

I. Pen-and-paper [9 v]

Part A - Clustering [4.5v]

Consider the observations:

x	y_1	y_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

Suppose that the initial seeds (centers of each cluster) are x_1, x_4 and x_7.

a. Draw the initial points and the cluster centers. [0.5v]
b. Perform one epoch of the k-means algorithm. Use Euclidean distance. [1.5v]
c. Draw the new centers of the clusters, draw the points and identify which cluster they belong to (for this, you can circle the points and cluster center). [1v]
d. Discuss the impact of different centroid initializations on the convergence and performance of the k-means algorithm. [1.5v]

Part B - PCA [4.5v]

We have three 3D points:

D	y_1	y_2	y_3	y_4
x_1	5	0	1	+
x_2	0	5	0	+
x_3	1	0	-1	-

a. Compute the covariance matrix. [1v]
b. Determine the projection plane (2-dim. space) that minimizes the projection error. [2.5v]
c. Does this plane help to discriminate between the 2 classes? Justify your answer by plotting the points into the projection plane. [1v]

TÉCNICO LISBOA

Aprendizagem 2025

Homework IV

Deadline 23/10/2025 (Wednesday) 23:59 via Fenix as PDF

II. Programming and critical analysis [11 v]

Part A: Clustering [5.5 v]

Consider the `diabetes.csv` dataset (available on the course website), which aims to predict whether a patient has diabetes based on various health-related attributes. Normalize data using **MinMaxScaler**. **Always** use `random_state = 42`.

1. Using sklearn, apply K-Means clustering on the normalized data with k = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11}, max_iter = 500. Plot the SSE (sum of squared errors) values for each number of clusters. [1.5 v]
2. Using k=6, assign each observation to a cluster and classify them based on the majority class of the cluster they belong to (Example: If cluster A has 100 observations of class 0 and 99 observations of class 1, classify all 199 observations as the majority class: 0). Compute the confusion matrix, accuracy, precision, recall, and F1-score for this classification model. Is this a good classification model? Justify your answer based on your experience with classification models, and discuss the limitations of using clustering for classification tasks. [2 v]
3. Print the class distribution and the cluster centers for this model. Choose the 3 most discriminative clusters and interpret the results. What can you conclude about the patients in those clusters? [2 v]

Part B: PCA [5.5 v]

Using the same dataset with the same normalization.

1. Apply Principal Component Analysis (PCA) and plot the cumulative explained variance by the PCA components. How many principal components should be retained to explain at least 80% of the total variance? [1.5 v]
2. Plot the class distribution along the first principal component and comment. [1 v]
3. Apply Linear Discriminant Analysis (LDA) and plot the class distribution along the LDA component. Comment on the results. [1.5 v]
4. Which of the two methods (PCA or LDA) would be more appropriate for building a discriminant rule? Justify your answer based on the results obtained and explain why that method is more suitable. [1.5 v]

END