Machine Learning

Final exam. January 19, 2026

For each of the following tasks you must provide your code (in a text file submitted to Aula Global) and your answer (on a separate file, with the format that you prefer).

Do not compress your files!!

1. Download the dataset ML-26.rda, which contains information about the analysis of genuine and forged banknote-like specimens. Matrix trainData contains five numerical variables which summarise specific features of these notes; their authenticity is identified with labels 0 (genuine) and 1 (forged) in vector trainClass. An additional testing set and the corresponding class of the observations is provided in testData and testClass, respectively.

   (a) Run PCA on trainData and state which amount of variability is explained by each feature. Reduce the dimesionality of this data set to the number of features required to explain at least a 85% of that. State one of the (two) mathematical problems seen in class that yield the Principal Component Analysis as a solution.

   (b) Explain one linear method and one non-linear method of your choice for classification. Apply the selected technique to the dataset you obtained after reducing its dimensionality in part (a).

   (c) For each of the methods in part (b) above, predict the class of each observation in testData and compare it to the true class. Provide the misclassification rate.

   (d) Merge the training and the testing datasets. Run k-means on this (entire) dataset to cluster the observations into 2 clusters, including all the features. Find the number of observations that have been incorrectly clustered. Explain what is different (and particularly cumbersome) in the process of identifying this number.

   (e) Is there any improvement in the clustering results if you use the spectral clustering based on the mutual k-nearest neighbour graph and the random walk normalised Laplacian?

2. Explain briefly the Bayesian approach to the regression problem (with the standard Gaussian assumptions) and how this is used to provide the maximum posterior estimator of the parameters of the model.