Faculdade de Tecnologia, Universidade de Brasília, Campus Universitário Darcy Ribeiro, Gleba A, Avenida L3 Norte, CEP 70.910-900, Brasília - DF
Caixa postal 4386, fone +55 61 3307 2300, fone/fax +55 61 3273 8893, ftd@unb.br, www.ft.unb.br

PPGEE2249 – Aprendizado de Máquina
Prof. Daniel Guerreiro e Silva

Assignment 3

1) (4 points) Consider the dataset of 3000 bivariate, labeled samples in data_bayesian_question.csv file. **It is forbidden to use Bayesian Classification libraries/toolboxes.**
   a) If the samples labeled with "+1" are drawn by a bivariate gaussian density with $\mu_{+1} = [6 - 4]^T, \Sigma_{+1} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$, the samples labeled with "-1" are drawn by another gaussian density with $\mu_{-1} = [0.52.5]^T, \Sigma_{-1} = \begin{bmatrix} 4 & -2.4 \\ -2.4 & 9 \end{bmatrix}$ and the prior probabilities are $P(C_{+1}) = 0.7$ e $P(C_{-1}) = 0.3$, present the discriminant functions of the Bayes classifier and evaluate its performance via the error rate over the dataset.
   b) Calculate the decision boundary and plot it on a graph with the dataset samples. Comment your results.
2) (3 points) Consider the three-dimensional dataset in data_pca_question.csv. Apply PCA to study the data, **writing your own code**.
   a) What is the sample mean of the data? Then, create a new dataset with null sample mean and unit variance.
   b) Calculate the sample covariance matrix of the dataset. Calculate its eigenvalues and eigenvectors.
   c) Based on the results of (a) and (b), analyze if it is possible to reduce the dataset to (i) one or (ii) two dimensions. Use numerical measures to justify your analysis.
   d) Plot, in the same 3D graph: (i) the original data, (ii) the **reconstructed** data from the 1D projection and (iii) the **reconstructed** data from the 2D projection. Comment the results.
   e) Based on the previous results, is PCA a useful tool for this dataset?
3) (3 points) Choose a 2D or 3D dataset to perform k-means clustering. Third-party libraries / toolboxes are allowed. Explain the steps of your solution and provide a justified decision on the number of clusters (graphically or numerically).