faculdade de tecnologia | departamento de engenharia elétrica
Universidade de Brasília

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 - Assignment 2 Prof. Daniel Guerreiro e Silva

Instructions

- You may use a text editor (Word, LibreOffice, LaTeX, etc.) to write your answers.
- The use of machine learning libraries that provide off-the-shelf algorithms is forbidden (e.g., scikit-learn, Keras, PyTorch, MATLAB/Octave toolboxes). You must implement the algorithms yourself. However, auxiliary tasks (e.g., performance metrics, statistical analysis, dataset splitting, preprocessing) may be performed using such libraries.
- The recommended programming languages are Python or MATLAB/Octave.
- AI-based tools (Gemini, ChatGPT, etc.) may be used for assistance. However, you will be considered the sole author of the entire document (both text and code) and will assume full responsibility for any cheating, plagiarism, or meaningless content.
- Submit all your answers in a single PDF file and your code in a separate ZIP file.

Questions

1) Consider the three-dimensional dataset in data_pca.csv. Apply PCA to study the data.
   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?
2) Implement the k-means clustering algorithm. Then, choose a 3D dataset to apply your tool and test it. Explain the steps of your solution and justify your choice of the number of clusters, using graphical or numerical evidence.
3) You are given a dataset of 3168 labeled instances for voice-based gender classification: data_gender_voice.csv. Each instance contains 19 acoustic features extracted from voice recordings in the 0-280Hz frequency band. The associated label is in the last column of the dataset, where ‘1’ indicates male and ‘0’ indicates female.
   a) Perform a basic feature analysis: plot the histograms of the features and compute their correlations.

b) Implement a logistic regression model to classify voice instances by gender. Use a 20% split for the test set, with the remaining instances for training. Keep in mind that the dataset is ordered, therefore, randomization is necessary before splitting. Consider whether preprocessing (e.g. normalization) is appropriate. Present and discuss the results on the test set, including:

- the ROC curve;
- the F1-score versus decision threshold curve.

See Chapter 20 of the textbook and this paper for more information about the ROC curve and F1-score.

c) Which decision threshold is the most appropriate? Why? Using this threshold, compute and plot the confusion matrix and the classifier accuracy on the test set. Discuss your results.