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 1
Prof. Daniel Guerreiro e Silva

Instructions

- This assignment is divided into two parts. In the first part, you must write the answers by hand. In the second part, you may use a text editor (Word, LibreOffice, LaTeX, etc.) to write your answers.
- The use of machine learning libraries with off-the-shelf algorithms is forbidden (e.g., scikit-learn, Keras, PyTorch, MATLAB toolboxes, etc.). You must implement the algorithms yourself. 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.

PART 1

1) X is a continuous-valued random variable with uniform density in $(-1,+1)$.
   a) Draw its probability density function. What is the area under the curve? Justify your answer.
   b) Draw its cumulative distribution function.
   c) Calculate the probability of the event $X \in (-0.2, 0.2)$.
   d) Calculate the expected value $E[X]$, the second $E[X^2]$ and the fourth moment $E[X^4]$ of the random variable. Calculate its variance $Var[X]$, as well.
2) X is a discrete-valued random variable with uniform distribution over the set $\{-2,-1,0,1,2\}$. Draw its probability mass function and calculate $E[X]$ and $Var[X]$.
3) Consider the normal random variables $X_1 \sim N(-2,2)$, $X_2 \sim N(1,4)$, with $Cov(X_1, X_2) = -0.8$. Calculate the joint pdf of $X = (X_1, X_2)^T$.
4) Assume you want to design a classifier which receives univariate data as input and must choose between class +1 or class -1. The data associated with class “+1” follow a gaussian density with mean 2 and unit variance. On the other hand, the data associated with class “-1” follow a uniform density between -2 and 2. The prior probabilities are $P(C_{+1}) = 0.6$ and $P(C_{-1}) = 0.4$.

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

a) Apply Bayes' methodology to obtain the optimal classification model and show, using a diagram, the model's architecture (with clear indication of the discriminant function(s)).
b) Calculate the model's decision boundary. Analytically calculate the classifier error probability (the integrals can be numerically solved).
c) Recalculate the model's decision boundary if the Maximum-Likelihood criterion is adopted, this time. Analytically calculate the classifier error probability (the integrals can be numerically solved).
d) Compare the performances of both methodologies.

PART 2

5) Consider the dataset of 3000 bivariate, labeled samples in data.csv file. Firstly, split the dataset into training (70%) and test set (30%).
   a) Assume the samples labeled with "+1" are drawn by a bivariate gaussian density with parameters μ_{+1}, Σ_{+1}, while the samples labeled with "-1" are drawn by another gaussian density with parameters μ_{-1}, Σ_{-1}. The prior probabilities $P(C_{+1})$ e $P(C_{-1})$ are unknown as well. Estimate the missing parameters (with the training set), present the discriminant functions of the Bayes classifier (MAP criterion) and evaluate its performance (accuracy, precision, recall) over the test set. Comment all your results.
   b) Calculate the decision boundary and plot it on a graph with the samples of the test set. Comment your results.
6) Do your own implementation and train a multivariate linear regression model for a given problem (suggestions: Kaggle, UCI Machine Learning Repository). You may use a validation set to train models with different subsets of features and select the best one. Then, use a test set to report and comment the final model results.