Homework 1   (50 marks)

Submission by 6 Oct 2022 (Thursday, 2359pm)

Project 1: Prepare a brief report using A4 papers (in PDF format, maximum 8 pages excluding appendix, include your name and Student ID in the first page of the report)

Obtain data from LearnUs web-site: `train.txt` and `test.txt`. These two data sets are in the format (x_1, x_2, y) where $\mathbf{x} = [x_1, x_2]^T \in \mathbb{R}^2$ is the feature vector and $y \in \{0, 1\}$ is the target vector. The file `train.txt` contains 1000 samples and the file `test.txt` contains 3000 samples. Use the entire 1000 samples from `train.txt` for training and use the entire 3000 samples from `test.txt` for testing.

Perform classification using the **Naive Bayes** (which assumes independent Gaussian distribution for each feature and using maximum likelihood for prediction), the **Linear regression**, and the k-**NN** classifier.

Include the following items (with observations and discussions) in your report:

(i) A plot of the training and test data and a brief description of the data. (5 marks)

(ii) A program (build your own Matlab function or Python subroutine) to compute/count the classification error rate. (Note: classification error is different from least squares error) (5 marks)

(iii) A program (build your own Matlab function or Python subroutine) to compute the Naive Bayes decision output. Provide both the training and the test classification accuracies apart from the subroutine codes. (5 marks)

(iv) A program (build your own Matlab function or Python subroutine) to compute the linear regression output. Provide both the training and the test classification accuracies apart from the subroutine codes. (5 marks)

(v) A brief description regarding settings and utilization of a build-in Matlab or Python subroutine for the k-NN classifier. In other words, you will be using such a build-in k-NN subroutine for classifying the given data subsequently. (5 marks)

(vi) Train the k-NN classifier using all the data in `train.txt` for $k \in \{1, 5, 10, 15, 20\}$ and then use test it using `test.txt`. Tabulate and graph both the training and the test error rates for these k settings.
(5 marks)

(vii) Plot the trained decision boundary for the Linear regression model. (5 marks)

(viii) Provide a brief discussion regarding your observations. (5 marks)

(ix) Include all codes in the appendix in small size fonts. (10 marks)

Note: plagiarism found in any form (e.g., copied pictures, results or texts) will be heavily penalized.