HW2

CISC648010 - Spring 2022

Due Date: Feb 25 at 11 PM

1 Problem 1 (LDA and decision boundary) 15 pts

Consider the following dataset: $D = [(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)]$, where,

$$x_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, x_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, x_4 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \tag{1}$$

$$y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2 \tag{2}$$

Using the LDA classifier, find the decision boundary. (15 pts)
Hint the decision boundary should be a line in the Cartesian plane.

2 Problem 2 (LDA Spam filtering) 15 pts

Download **problem2.py** from Canvas. This code loads spambase dataset.

In this code x is $n \times d$ where $n = 4601$ and $d = 57$. The different features correspond to different properties of an email, such as frequency with which certain characters appear y is a vector of labels indicating spam (label 1) not spam (label 0). For the detailed description of the dataset, visit the UCI Machine Learning Repository or Google Spambase. The provided Code split the data into training and test dataset.

In this problem, you should train an LDA classifier for the problem of spam detection, using the training data. Use the test data to report the test error. Upload your code on Canvas as a single file named **prob2_<lastname>.py**. Please also report the test error.

Notes: The base codes have been written using python 3. If you get an error when you run problem2.py or problem1.py, you may be using a wrong version of python.

If you do not know which environment you should use for writing a python code, you can start with google colab. If you need help with google colab, please check here: https://www.youtube.com/watch?v=i-HnvsehuSw