Assignment 1

(*Note:* 1. Each problem needs to be coded both in R and python.)

(*Note:* 2. You are not supposed to use any library function for least squares, linear regression, or k-Nearest neighbors. You need to code from scratch.)

(*Note:* 3. The hint given at the end is to help you think of an approach. You can use other methods if you find them comfortable.)

1. Construct a data set with 200 points using the following steps:

   (a) Generate 10 two-dimensional vectors, $m_0, m_1, \ldots, m_9 \in \mathbb{R}^2$, such that they are i.i.d. $\sim \mathcal{N}\left([1,0], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$. Generate 10 more two-dimensional vectors, $m'_0, m'_1, \ldots, m'_9 \in \mathbb{R}^2$, such that they are i.i.d. $\sim \mathcal{N}\left([0,1], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$.

   (b) Generate 10 data points $(X_{i*10+1}, X_{i*10+2}, \ldots, X_{i*10+10}) \in \mathbb{R}^2$ such that they are i.i.d. $\sim \mathcal{N}\left(m_i, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)$, for all $i = 0, 1, \ldots, 9$. Thus a 100 feature vectors are generated. Label them all as $(+1)$, i.e., $Y_1 = Y_2 = \ldots = Y_{100} = 1$.

   (c) Generate 10 data points $(X_{i*10+101}, X_{i*10+102}, \ldots, X_{i*10+110}) \in \mathbb{R}^2$ such that they are i.i.d. $\sim \mathcal{N}\left(m'_i, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)$, for all $i = 0, 1, \ldots, 9$. Thus another 100 feature vectors are generated. Label them all as (-1), i.e., $Y_{101} = Y_{102} = \ldots = Y_{200} = -1$.

   Plot the generated features in the form of a scatterplot. Represent the features having different labels with different colors.

2. Compute the classifier using the linear model. Plot the classifying line along with the scatter plot of the generated features, showing features above and below the line. Print the training error (i.e., the average of misclassified data) for the classifier computed.

3. Compute the classifier using k-nearest neighbor method, for $k = 15$. Plot the classifying curve and print the training error as mentioned in Q2.

4. Repeat Q3 with $k = 1$.

5. Generate 10000 test vectors as follows: Generate 5000 feature vectors as given in Q1(b) (i.e., 500 vectors for each m_i), and label them all as $(+1)$.

Similarly, generate 5000 more feature vectors as mentioned in Q1(c) (i.e., 500 vectors per m_i'), and label them all as (-1). Assume that the labels correspond to the true responses. Compute and print the test error based on the classifiers derived using (i) linear model, (ii) 15-NN, and (iii) 1-NN.

(*Hint:* For Q3 and Q4, the region of classification can be computed as follows: (i) Quantize both axes to form a grid, (ii) Apply the k-NN rule in each of those points, (iii) Join the common regions using curves.)