**BENG 420/520 - Homework #1**

Due on 9/23/2019, 11:59pm

**K-nearest neighbor (K-NN) classifier:**

1. Create a new Matlab file called h1\_plot.m and program in this file only. Load the data (h1\_data.mat) in Matlab. The variable “**features**” contains values of two features. Each row in “**features**” represents one data point. The variable “**classlabels**” contains the corresponding class labels of all the data points being either 1 or 2.
2. *Generate a 2-dimensional plot of the feature values of all data points (Figure 1).* The x-axis is of the value of the first feature (column #1 in variable “**features**”) and the y-axis is of the value of the second feature (column #1 in variable “**features**”). Plot the two classes of data points as indicated by “**classlabels**” by two different symbols of different colors. *Describe your observation of the data.*
3. Use the given data to better understand the K-NN classification algorithm. Implement the K-NN classifier. Use Euclidean distance as the distance metric. Do not use any existing Matlab functions.
4. Use the first **n\_train (500 or 2500)** data points as the training set and the data points between rows 4001 and 4500 for validation (i.e. to fine tune parameter “K”). Iterate all K values between 1 and 80. *Plot the classification error percentage* of the validation set as a function of **K**, which is defined as the number of data points being misclassified divided by the total number of data points evaluated. *Explain your observation on the relationship between K and the classification error. Is there an optimal value of K? Was the relationship expected? Discuss the influence of K on classification error, as shown by this example.* Finallytest the performance of your classifier using the optimal **K** value you determined from validation data set on a separate testing set (data points between rows 5001 and 5500) and *report the error you observe*. *Compare the classification error percentages of the optimal* ***K*** *on the testing set and the validation set.*
5. Run your classifier by setting **n\_train** as 500 and 2500 respectively (Figure 2 and Figure 3) and answer the above questions for each case. *What are the strength(s) and weakness(s) of K-NN classifier demonstrated by your results? How does the size of the training set affect the classification accuracy?*
6. Submit your answers in a document and your Matlab code in a single file h1\_plot.m. Make sure it produces the three figures noted above. Please comment your code which will be helpful for the grader to give partial credits.