Assignment Two
ECE 4950

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 3/05/2017, 23.59.59 Eastern time**. The total points is 60.
- Submission rules are the same as previous assignments.

Problem 1 (20 points). Consider following data set:

Gender	Height(in)	Foot size
M	72	12
M	68	9
M	75	11
M	64	10.5
F	65	8
F	67	7.5
F	62	6
F	70	8.5
F	64	8

1. What are the class probabilities Pr(M) and Pr(F) according to empirical rule covered in the class?
2. Assuming Gaussian naive Bayes estimate following:
   means: $\mu_{H|M}$, $\mu_{S|M}$, $\mu_{H|F}$, $\mu_{S|F}$,
   variances: $\sigma^2_{H|M}$, $\sigma^2_{S|M}$, $\sigma^2_{H|F}$, $\sigma^2_{S|F}$, where (H=Height, S=Foot size).
3. Suppose you were given following example:

Height(in)	Foot size
68	9.5

   Using means and variances from above part compute $p(H|M)$, $p(S|M)$, $p(H|F)$, $p(S|F)$.

4. Using Gaussian naive Bayes assumption, what would you classify above sample, male or female?

Problem 2 (20 points). Recall the WDBC dataset, and the data-sets we used in the last assignment. In particular, we had two datasets, the first consisted of `X-trn-200.csv`,`Y-trn-200.csv`, `X-tst-200.csv`, and `Y-tst-200.csv`. The second dataset contained four files ending with `-400.csv`. This time, in addition, we are providing four files that end with `-50.csv` that has 50 training examples and 519 test examples obtained as before.

1. Write a code that implements Gaussian Naive Bayes over the new dataset, and the two previous datasets (three datasets in total `-50.csv, -200.csv, -400.csv`).
2. What is the accuracy of the the algorithm over the two test sets?
3. Use the code you wrote for decision tree in the first assignment on the dataset ending with `-50.csv`, again plotting the test and train accuracy as a function of the maximum depth. Which one has a better test error: Naive Bayes or the decision tree with the best depth?
4. Instead of using maximum depth to prevent overfitting, modify the code to obtain decision trees with minimum leaf size i. For $i = 2, \ldots, 15$, plot the test and training accuracy. Does this achieve a higher accuracy compared to the maximum depth criterion? Is it better than Naive Bayes?

Please submit all the codes, and plots.

Problem 3 (20 points). In this problem we will implement the Multinomial Naive Bayes (add-β) algorithm for the MNIST dataset. Each image is a 28x28 gray-scale image. Therefore each image is a 784 dimensional feature vector. There are 60,000 training examples, and 10,000 test examples. The training data is `mnist_train.csv`, that contains 785 entries in each row, where the first entry is the digit label, and the remaining 784 are the values of the 784 pixels. The test data is `mnist_test.csv` which has the same structure.

1. Implement a Multinomial Naive Bayes for smoothness parameter β= 0.0000001, 0.0001, 0.1, 1, 1000, 1000000, 1000000000, 1000000000000. Plot the error rate as a function of β.
2. What does the classifier do as $\beta \to \infty$? Please explain the error rate obtained for $\beta \to \infty$.

Notes regarding the programming assignment:

- It is perhaps easiest to implement these problems using Numpy, and Scipy, and in particular `sci-kit-learn`. Look for naive Bayes implementations in sci-kit-learn.