HW #1 Due: 3/13/2018

1. It is known that the HIV test has only 0.1% of false positive and false negative, respectively. However, for a specific group of people, the prevalence of HIV positive rate is 0.01 %. If a person belongs to such a group and is found to be positive in the HIV test, find the probability that the person is really infected.
2. UC Irvine has a large repository for various kinds of data. In this problem, you are asked to use the iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris) to perform the experiments. Implement the k-NN classifier for the classification task. To begin one experiment, randomly draw 70 % of the instances for training and the rest for testing. Repeat the drawing and the *k*-NN classification 10 times and compute the average accuracy. Then, plot the curve of *k* versus accuracy for $k = 1, 3, \ldots, 15$. For simplicity, use the Euclidean distance in your computation.
3. Following problem 2, if you do not have the test dataset (i.e., you have only the 70 % of dataset), how do you determine the optimal value of *k*? Use your own approach to find such a value and compare the results you have in problem 2. Comment on your results.
4. In the class, we covered the naive Bayes classifier, but only with discrete-type features. Consult any paper to learn how to extend this approach to continuous-type features. Explain your finding as an algorithm.
5. Repeat problem 2 with your algorithm in problem 4. Compare the accuracy of naive Bayes classifier with the *k*-NN.
6.