HW #1 Due: 3/27/2020

1. In the lecture, we mentioned that different algorithms have different problems. Use your own words to explain the shortcomings of each of the following methods:
   - Neural networks (particularly CNN)
   - C4.5 decision tree
   - Adaboost
2. 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. (problem in ppt file)
3. 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 Naïve Bayes classifier for the classification task. To begin one trial, randomly draw 70 % of the instances for training and the rest for testing. Repeat the trials 10 times and compute the average accuracy. As the features are continuous variables, you may want to use the Gaussian model in probability computation.
4. We can also use the iris dataset for regression work. Consult the Internet to learn the equations of linear regression, and use the first three features (sepal length, sepal width, petal length) in each sample as input to predict the fourth feature (petal width). To conduct one trial, again you need to divide the dataset into a training set (70%) and a test set (30%). Furthermore, you need to build the model for each iris class. Repeat 10 trials and report the average MSE for each class. Is linear regression an acceptable regression model in this problem? Why?
5. Write a program to reproduce the likelihood function on pp. 7 of the "Parametric estimation" PPT file. Use 201 points in the plot (i.e., take 201 different θ values from 0 to 1). Is your $\hat{\theta} = 0.6$?