HW #1 Due: 3/21/2019

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. The textbook claims that the set of rectangles in $R^2$ has a VC dimension of 4. Why is it the case? You may draw plots to show it.
3. UC Irvine has a large repository for various kinds of data. In this problem, you are asked to use the dataset of "Breast Cancer Wisconsin (Original) Data Set" (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) to perform the experiments. To simplify the problem, you just need to classify whether the subject is benign or malignant (attribute 11). Implement the k-NN classifier for the classification task. To begin one experiment, randomly draw 70 % of the instances from each class for training and the rest are 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* = 3, …, 15. For simplicity, use the Euclidean distance in your computation.
4. Following problem 3, compute the covariance matrix of the dataset. The matrix is of size 9 × 9 (attribute 2 – 10). Do you see strong correlation between any two attributes?
5. Consult any paper to learn how to extend the *k*-NN approach to perform regression. Based on your findings, implement the program and test it on "Computer Hardware Data Set," which is available at https://archive.ics.uci.edu/ml/datasets/Computer+Hardware for downloading.