STAT 598Z: Homework 5

Due: 3rd April 2012

1. This homework will contribute 10 points towards your final score.
2. Attempt as many problems as possible.
3. Only neatly handwritten solutions will be accepted. Alternatively you may use LaTeX to typeset your solutions.
4. Hand in your HW (including print outs of your source code) at the beginning of the class on 3rd April 2012. Additionally source code (if any) should be emailed to `stat598z@gmail.com` **before** the assignments are submitted in the class. No late submissions will be accepted!
5. Program files should be named after the problem (e.g. solution to problem 1 should be problem1.py etc).
6. Remember to seed your random number generators!

Problem 1 (10 pt) Download `a9a` (training) and `a9a.t` (test) dataset from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#a9a`.

- Write a k-nearest neighbor classifier which uses data points in training data to classify data points in test data.
- Vary the value of k and find the optimal value of k which will maximize the accuracy in test dataset. Write a short report (2 pages maximum) of your findings.
- Using the $O(\cdot)$ notation, describe the time complexity of classifying a point in the test dataset using your code written above. You may want to use the following notation: number of training points (resp. test points) is n_1 (resp. n_2), dimension of data is d, and number of nearest neighbors considered is k. If you define additional notations then explain it clearly for full credit.