STAT 598Z: Homework 5

Due: 2nd April 2013

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 2nd April 2013. 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) Generate 7,000 samples from two-dimensional standard normal distribution with mean $(0, 0)$ and covariance matrix $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Let labels of these samples be 1. Generate 1,000 samples from the normal distribution with mean $(1, 0)$ and covariance matrix $\Sigma = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}$. Let labels of these new samples be 2. Generate 2,000 samples from the normal distribution with mean $(1, 1)$ and covariance matrix $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. Let labels of these new samples be 3.

- Plot these sample points, using blue color for class 1, red for class 2, and green for class 3.
- Implement a nearest neighbor classifier in Python. Generate 100 additional samples from each of the three classes and classify them using the nearest neighbor classifier you wrote. What is the accuracy of your classifier?
- Now, implement a k-nearest neighbor classifier, and change the value of k (for example, from 1 to 5) to find the k which maximizes the accuracy. Write a short report (2 pages maximum) on 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.

- Test your algorithm on any dataset of your choice from the LibSVM multiclass data repository `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html`.