PubH 7475/8400 Homework 1 (Spring 2007)

Due on Jan 30, 2007

1. Consider a breast cancer data set available in R; see backpage for more details on the data. In R, use the following commands to download the data:

```
> library("MASS")
> ?biopsy
> data(biopsy)
> biopsy[1:5,]
       ID V1 V2 V3 V4 V5 V6 V7 V8 V9  class
1 1000025  5  1  1  1  2  1  3  1  1 benign
2 1002945  5  4  4  5  7 10  3  2  1 benign
3 1015425  3  1  1  1  2  2  3  1  1 benign
4 1016277  6  8  8  1  3  4  3  7  1 benign
5 1017023  4  1  1  3  2  1  3  1  1 benign
```

Because there are some missing values for V6, you can use the following to delete the observations with missing values:

```
> biopsy2<-biopsy[!is.na(biopsy$V6),]
```

Alternatively, you can download the data from UC-Irvine Machine Learning Databases.

(a) Randomly split the data into a training set and a test data containing about 2/3 and 1/3 of total observations respectively.

(1) Apply a linear regssion model to obtain its training, test and LOOCV (based on only the training set) error rates;

(2) Apply kNN and for various values of k, show their training, test and LOOCV error rates;

(3) Apply a logistic regression model to obtain its training, test and LOOCV (based on only the training set) error rates;

Which of the training error rate and LOOCV error rate approximates the test error rate better? (30 pts)

(b) Randomly split the data into a training set and a test data containing about 1/3 and 2/3 of total observations respectively. Repeat (1)-(3) in 1. How the performance depends on the size of the training dataset? (15 pts)

In (1) and (3), you can either use all the 9 predictors directly, or even better, use a variable selection scheme to select a model.

2. **(PubH 8400)** Read Breiman (2001) and Hand (2006) (downloadable from the course web page); summarize the main points of each paper and briefly explain your view(s). (20 pts)

Please attach your computer program and relevant output.