PubH 7475/8475/Stat 8056 Homework 1 (Spring 2019)

Due on Feb 4, 2019

1. Consider a breast cancer data set available in R. In R, use the following commands to learn and 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 set 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 a set of the 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 set 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. Read Guha et al (2012) about the main idea of "divide and recombine" (DR) (downloadable from the course web page). Suppose we have data $D = \{X_1, X_2, ..., X_n\}$, n identically and independently distributed (iid) (scalar) observations from a normal distribution $N(\mu, \sigma^2)$.

   (a) Show that, based on all n observations in D, the maximum likelihood estimates μ and σ^2 are $\hat{\mu}(D) = \sum_{i=1}^{n} X_i/n$ and $\hat{\sigma}^2(D) = \sum_{i=1}^{n}[X_i - \hat{\mu}(D)]^2/n$. (10 pts)

   (b) Now suppose that n is an even number; we divide the sample into two equally sized subsamples $D_1 = \{X_1, X_2, ..., X_{n/2}\}$ and $D_2 = \{X_{n/2+1}, X_{n/2+2}, ..., X_n\}$. If we apply the MLEs to the two subsamples, then combine them to obtains the DR estimate for μ as the following

   $$\tilde{\mu}(D) = [\hat{\mu}(D_1) + \hat{\mu}(D_2)]/2.$$

   How is the DR estimate compared to the original MLE $\hat{\mu}(D)$ (i.e. the same, worse or better)? why? (10 pts)

   (c) **(8000)** If we apply the MLEs to the two subsamples, then combine them to obtains the DR estimate for σ^2 as the following

   $$\tilde{\sigma}^2(D) = [\hat{\sigma}^2(D_1) + \hat{\sigma}^2(D_2)]/2.$$

   How is the DR estimate compared to the original MLE $\hat{\sigma}^2(D)$ (i.e. the same, worse or better)? why? (20 pts)

3. **(8000)** Read Breiman (2001), Hand (2006) and Donoho (2015) (downloadable from the course web page); summarize the main points of each paper and briefly explain your view(s). (30 pts)

Please attach your computer program and relevant output.