Statistical Learning and Data Mining, Midterm Exam 2017/11/20

1. Consider the "auto" data set. Let $y =$ mpg(miles per gallon) be the response variable and $x_1 =$ displacement, $x_2 =$ horsepower, $x_3 =$ weight, $x_4 =$ acceleration, $x_5 =$ year be the predictors.

   (a) Fit a linear regression model of y on the five predictors $x_1, \ldots, x_5$ using the training data. Write down the fitted model, the training and the test mean square errors. Rebuild the model using only those predictor(s) which is(are) statistically significant (use 5% significance level). Write down the fitted model, the training and test mean square errors for the new model.

   (b) Fit the ridge regression model using the tuning parameter λ which minimizes the 10-fold cross validation error. Write down the fitted ridge regression model, the training and test mean square errors.

   (c) Fit the lasso regression model using the tuning parameter λ which minimizes the 10-fold cross validation error. Write down the fitted lasso regression model, the training and test mean square errors.

   (d) Based on the results of (a)–(c), which model will you suggest to use for future prediction? Justify your answer.

   (e) Use the training data set and the bootstrap method with 1,000 replication to estimate the standard errors (s.e.) of the regression coefficient estimates for the three fitted models in (a)-(c), that is

   - Multiple linear regression with the five predictors.
   - Ridge regression model with λ chosen in (b).
   - Lasso regression model with λ chosen in (c).

   For each coefficient estimate identify the model which has the smallest s.e.. Note that there are five coefficient estimates and three models. Explain the result.

2. Consider the digits recognition data "digits". In the data set, the response $y = label$ represents the label of digits (from 0 to 9) and the predictors $\boldsymbol{x} = (pixel_1, \ldots, pixel_{784})$ are the corresponding grayscale value(灰階數值) of the 28×28 (image sizes) pixels.

   (a) Perform PCA on the training set. Find the smallest k, denoted by k^*, such that the variance of the first k principal components exceed 90% of the total variance. Let $\boldsymbol{x}^* = (pixel_1^*, \ldots, pixel_{k^*}^*)$ denote the first k^* principal component scores of $\boldsymbol{x} = (pixel_1, \ldots, pixel_{784})$. Use the training loadings to find the first k^* principal component scores of the training data set and the test data set.

   (b) Treat $\boldsymbol{x}^*$ as the new predictors and use the LDA method to predict y. Find the training and test classification errors.

(c) Treat $\boldsymbol{x}^*$ as the new predictors and use the QDA method to predict y. Find the training and test classification errors.

(d) Use the new predictors $\boldsymbol{x}^*$ to perform KNN on the training data. Provide a table of the training and test classification errors for $1 \leq K \leq 10$. Which value of K attains the smallest test classification error? Plot the training and test classification errors versus K ($1 \leq K \leq 10$) in a single figure. Make sure to label the training and test error curves.

(e) Fit lasso multiple logistic models using $\boldsymbol{x}^*$ to predict y where the tuning parameter λ is chosen by minimizing the 10-fold cross validation classification error. What are the training and test errors? (Hint : Use `cv.glmnet()` with `family="multinomial"` and predict the class which has the maximum probability)

(f) Based on the results of (a)–(e), which model will you suggest to use for future prediction? Justify your answer.

3. Consider the "heart" data set. Fit a logistic regression model on the data to predict the coronary heart disease(chd)(冠狀動脈心臟疾病). In the data set, the response y=chd", where $y = 1$ indicates the case of coronary heart disease. Denote the predictors x_1 =sbp, x_2 =tobacco, x_3 =ldl, x_4 =adiposity, x_5 =famhist(Present:0, Absent:1), x_6 =typea, x_7 =obesity, x_8 =alcohol, x_9 =age.

(a) Find the accuracies of the fitted logistic regression model for the 90 threshold values $\{0.01h \ : \ 1 \leq h \leq 90, \ h \in \mathbb{N}\}$. Plot the accuracy v.s. the thresholds. Find the threshold that has the maximum accuracy and weite down its confusion matrix. (Hint : Use glm with `family="binomial"`)

(b) Find the sensitivity and specificity for the thresholds $\{0.01h \ : \ 1 \leq h \leq 90, \ h \in \mathbb{N}\}$ and plot the ROC curve. Find the threshold that has the maximum (sensitivity+specificity) and write down its confusion matrix.

(c) Based on the results of (a)–(b), which model will you suggest? Why?