ISyE 619 HW 2 Solution

Spring 2014

Due: 3/24 on class

Problem 1. (3 points) Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. Show both the training and test error for each choice. The zipcode data are available from the course website. (**Please provide your code**)

Data Description: for the training and test data sets, the first column stands for the response (Y) and the other columns stand for the independent variables (X_i's). In particular, consider only the class label $Y = 2$ and $Y = 3$, and k = 1, 3, 5, 7 and 15. (Hint: When using the linear regression for classification, you may consider a classifier as follows: if the fitted value of y is larger than 2.5, classify the label as 3; otherwise, classify the label as 2.). The following R code may be useful (the desired training data is "ziptrain23" and the corresponding test data is "ziptest23"):

Hint for code:

```
zip.train <- read.table(file="…/zip.train.csv", sep = ",");
ziptrain23 <- subset(zip.train, zip.train[,1]==2 | zip.train[,1]==3);
zip.test <- read.table(file="…/zip.test.csv", sep = ",");
ziptest23 <- subset(zip.test, zip.test[,1]==2 | zip.test[,1]==3);
# linear Regression
mod1 <- lm( V1 ~ . , data= ziptrain23);
pred <- predict.lm(mod1, ziptest23);
#KNN
library(class)
knn();
```

Problem 2. (6 points) In previous questions, we applied kNN to the zipcode data using class label $Y = 2$ and $Y = 3$. Now apply the following methods to the zipcode data. Compare the training and testing errors. (**Please provide your code**)

(1) LDA,

(2) QDA,

(3) Naive Bayes,

(4) logistic regression

(5) SVM

(6) CART

Problem 3. (6 points) Consider the extension of the previous problem to a multi-class classification problem. Apply kNN (k = 1, 3, 5, 7, 15), LDA, QDA, Naive Bayes and logistic regression to the zipcode data using class label $Y = 2$, $Y = 3$ and $Y = 5$. Compare the training and testing errors. (**Please provide your code**)

(1) LDA,

(2) QDA,

(3) Naive Bayes,

(4) logistic regression

(5) SVM

(6) CART

Problem 4. (3 points) Suppose we have features $\boldsymbol{x} \in R^p$, a two-class response, with class sizes N_1, N_2.

Show that the LDA rule classifies to class 2 if

$$\mathbf{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{u}}_2 - \widehat{\boldsymbol{u}}_1) > \frac{1}{2}\widehat{\boldsymbol{u}}_2^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{u}}_2 - \frac{1}{2}\widehat{\boldsymbol{u}}_1^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{u}}_1 + \log\left(\frac{N_1}{N}\right) - \log\left(\frac{N_2}{N}\right)$$

and class 1 otherwise.