MATH7340 HW13

Chengbo Gu

Problem 1 (70 points) Analysis of the ALL data set.

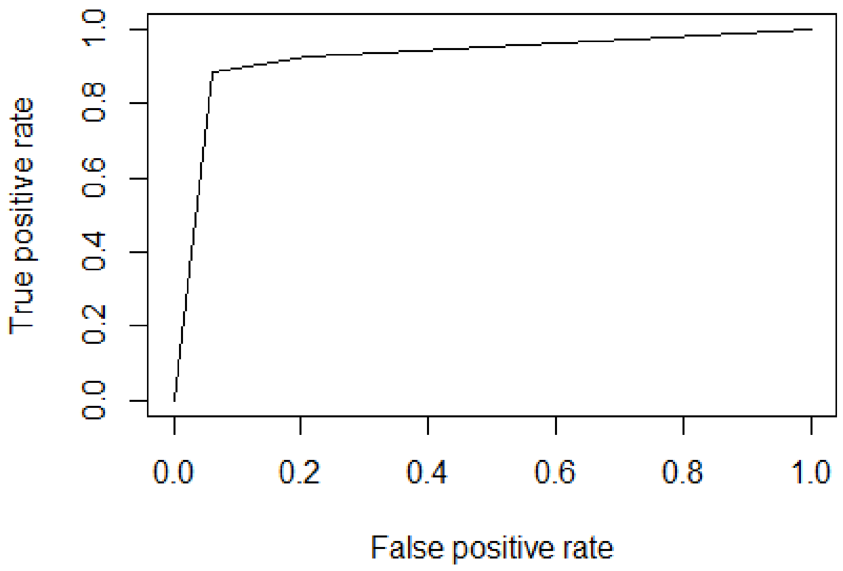
```
library(ALL)
library(ROCR)
library(e1071)
require(rpart)
require(caret)
```

(a) Define an indicator variable IsB such that IsB=TRUE for B-cell patients and IsB=FALSE for T-cell patients.

```
data(ALL)
IsB <- factor(ALL$BT %in% c("B", "B1", "B2", "B3", "B4"))
```

(b) Use two genes "39317_at" and "38018_g_at" to fit a classification tree for IsB. Print out the confusion matrix. Plot ROC curve for the tree.

```
probedat <- as.matrix(exprs(ALL[c("39317_at", "38018_g_at"),]))
c.tr <- rpart(IsB ~ ., data = data.frame(t(probedat)))
rpartpred <- predict(c.tr, type="class")
table(rpartpred, IsB)

##          IsB
## rpartpred FALSE TRUE
##     FALSE    31   11
##     TRUE      2   84

pred.prob <- predict(c.tr, type="prob")[,2]
pred <- prediction(pred.prob, IsB == "TRUE")
perf <- performance(pred, "tpr", "fpr")
plot(perf)
```



(c) Find its empirical misclassification rate (mcr), false negative rate (fnr) and specificity. Find the area under curve (AUC) for the ROC curve.

```
# mcr
mcr <- sum(rpartpred != IsB)/length(IsB)
mcr

## [1] 0.1015625

# fnr
fnr <- sum(rpartpred == "FALSE" & IsB == "TRUE")/sum(IsB == "TRUE")
fnr

## [1] 0.1157895

# specificity
fpr <- sum(rpartpred == "TRUE" & IsB == "FALSE")/sum(IsB == "FALSE")
spec <- 1-fpr
spec

## [1] 0.9393939

performance(pred, "auc")@y.values[[1]]

## [1] 0.922807
```

The empirical misclassification rate is 10.16%, false negative rate is 11.58%, specificity is 93.94%.

The area under curve is 0.922807.

(d) Use 10-fold cross-validation to estimate its real false negative rate (fnr). What is your estimated fnr?

```
data <- data.frame(IsB, t(probedat))
n <- dim(probedat)[2]
index <- 1:n
K <- 10
flds <- createFolds(index, k=K)
fnr.cv.raw <- rep(NA, K)
for (i in 1:K){
  testID <- flds[[i]]
  data.tr <- data[-testID,]
  data.test <- data[testID,]
  tree.cv <- rpart(IsB ~ ., data = data.tr)
  tree.cv.pred <- predict(tree.cv, newdata = data.test, type = "c")
  fnr.cv.raw[i] <- sum(tree.cv.pred == "FALSE" & data.test$IsB == "TRUE")/sum(data.test$I
sB == "TRUE")
}
fnr.cv <- mean(fnr.cv.raw)
fnr.cv

## [1] 0.1038131
```

The estimated fnr is 10.38%.

(e) Do a logistic regression, using genes "39317_at" and "38018_g_at" to predict IsB. Find an 80% confidence interval for the coefficient of gene "39317_at".

```
fit.lgr <- glm(IsB~. , family=binomial(link='logit'), data = data)
summary(fit.lgr)
```

```
##
## Call:
## glm(formula = IsB ~ ., family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.94163  -0.11964   0.03874   0.24553   2.27493
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.0388     4.2492  -1.892  0.05851 .
## X39317_at    -0.9902     0.3184  -3.110  0.00187 **
## X38018_g_at   2.9625     0.7235   4.095 4.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 146.114  on 127  degrees of freedom
## Residual deviance:  57.885  on 125  degrees of freedom
## AIC: 63.885
##
## Number of Fisher Scoring iterations: 7
```

```
confint(fit.lgr, level=0.8)
```

```
## Waiting for profiling to be done...
```

```
##                   10 %       90 %
## (Intercept) -13.767525 -2.8118382
## X39317_at    -1.427390 -0.6047588
## X38018_g_at   2.120174  3.9861802
```

80% confidence interval for the coefficient of gene "39317_at" is (-1.427390, -0.6047588).

(f) Use n-fold cross-validation to estimate misclassification rate (mcr) of the logistic regression classifier. What is your estimated mcr?

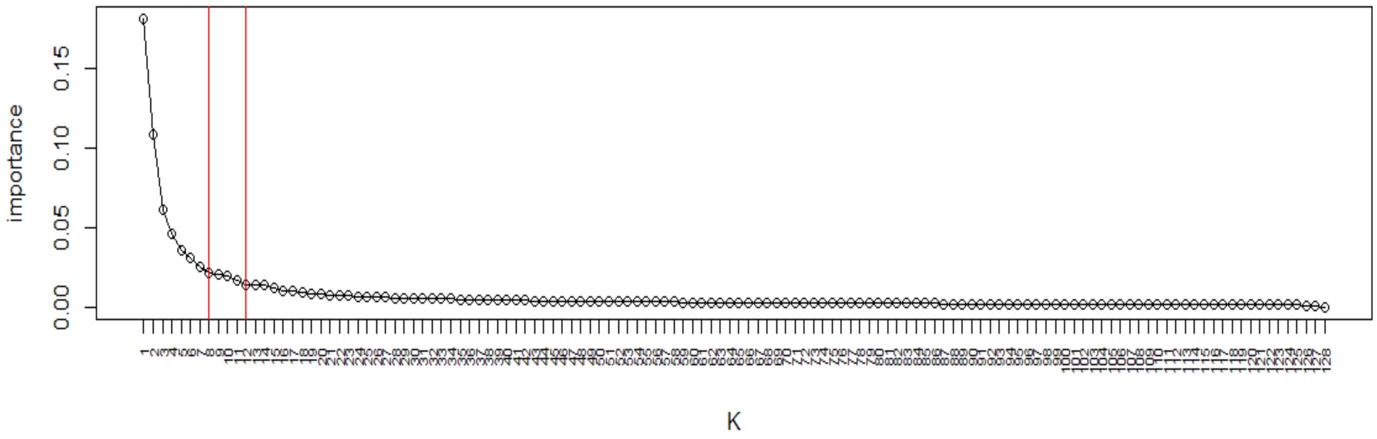
```
n<-dim(data)[1]
lgr.mcr.cv.raw <- rep(NA, n)
for (i in 1:n) {
  data.tr <- data[-i,]
  data.test <- data[i,]
  fit.lgr <- glm(IsB~., family = binomial(link='logit'), data = data.tr)
  pred.prob <- predict(fit.lgr, newdata = data.test, type ="response")
  pred.B <- (pred.prob > 0.5)
  lgr.mcr.cv.raw[i] <- pred.B!=data.test$IsB
}

lgr.mcr.cv <- mean(lgr.mcr.cv.raw)
lgr.mcr.cv
```

```
## [1] 0.09375
```

The estimated mcr is 9.38%.

(g) Conduct a PCA on the scaled variables of the whole ALL data set (NOT just the two genes used above). We do this to reduce the dimension in term of genes (so this PCA should be done on the transpose of the matrix of expression values). To simply our future analysis, we use only the first K principal components (PC) to represent the data. How many PCs should be used? Explain how you arrived at your conclusion. Provide graphs or other R outputs to support your choice.

```r
PCA<-prcomp(t(exprs(ALL)), scale=TRUE)
importance <- summary(PCA)$importance[2,]
K = (1:128)
plot(K, importance, type='o', xaxt='n')
axis(1, at = K, las=2, cex.axis=0.65)
abline(v=8, col="red")
abline(v=12, col="red")
```



Seems that k in range 8-12 is appropriate here.

We can draw the conclusion from the plot (importance vs K). After k=12, the line becomes flat.

(h) Do a SVM classifier of IsB using only the first five PCs. (The number K=5 is fixed so that we all use the same classifier. You do not need to choose this number in the previous part (g).) What is the sensitivity of this classifier?

```r
data.pca <- PCA$x[,1:5]
svm <- svm(data.pca, IsB, type = "C-classification", kernel = "linear")
svm.pred <- predict(svm , data.pca)

# tpr
sum(svm.pred == "TRUE" & IsB == "TRUE")/sum(IsB == "TRUE")
```

```
## [1] 0.9894737
```

The sensitivity of this classifier is 98.95%.

(i) Use leave-one-out cross-validation to estimate misclassification rate (mcr) of the SVM classifier. Report your estimate.

```r
n <- dim(data.pca)[1]
svm.mcr.cv.raw <- rep(NA, n)

for (i in 1:n){
  svm.cv <- svm(data.pca[-i,], IsB[-i], type = "C-classification", kernel="linear")
  svm.cv.pred <- predict(svm.cv, t(data.pca[i,]))
  svm.mcr.cv.raw[i]<-svm.cv.pred!=IsB[i]
}
svm.mcr.cv <- mean(svm.mcr.cv.raw)
svm.mcr.cv

## [1] 0.0390625
```

The estimated mcr of SVM classifier is 3.91%.

(j) If you had to choose between classifiers in part (e) and in part (h), which one would you choose? Why?

I would like to choose SVM. SVM has lower cross-validation mcr (3.91%) compared with the one (9.38%) in logistic regression.

2. (30 points) Choosing Classifiers and Number of Principal Components for PCA reduced iris data set.

In the last example of this module, we compared three classifiers on the iris data by working on the first three principal components. We choose the best classifiers based on cross-validated misclassification rate. We can also choose the number of principal components to use by cross-validation, instead of fixing it at K=3.

Use the leave-one-out cross-validation to choose the number of principal components together with the classifier. Please report the empirical misclassification rates (on whole data set) and the leave-one-out cross-validation misclassification rates for each value of K=1, 2, 3, 4 principal components and for each of the three classifiers: logistic regression, support vector machine and classification tree. Based on those rates, what is your choice?

```
rm(list=ls())
library(VGAM)
data(iris)

lgr.classification <- function(iris2) {
  iris2.lgr <- vglm(Species~. , family = multinomial, data = iris2)
  pred.prob <- predict(iris2.lgr, as.data.frame(iris2[,-1]), type = "response")
  pred.lgr <- apply(pred.prob, 1, which.max)
  pred.lgr <- factor(pred.lgr, levels=c("1", "2", "3"), labels = levels(iris2$Species))

  mcr.lgr <- mean(pred.lgr!=iris2$Species)
  cat("empirical mcr for logistic regression:", mcr.lgr, "\n")

  mcr.cv.raw<-rep(NA, n)
  for (i in 1:n) {
    lgr.fit <- vglm(Species~., family=multinomial, data=iris2[-i,])
    pred.prob <- predict(lgr.fit, iris2[i,], type="response")
    pred <- apply(pred.prob, 1, which.max)
    pred <- factor(pred, levels=c("1","2","3"), labels=levels(iris2$Species))
    mcr.cv.raw[i]<- mean(pred!=Species[i])
  }
  mcr.cv<-mean(mcr.cv.raw)
  cat("cross-validation mcr for logistic regression:", mcr.cv, "\n\n")
}
```

```
svm.classification <- function(iris2) {
  iris2.svm <- svm(Species~., type = "C-classification", kernel = "linear", data = iris2)

  svmpred <- predict(iris2.svm , data.pca)
  mcr.svm<- mean(svmpred!=Species)
  cat("empirical mcr for SVM:", mcr.svm, "\n")
  mcr.cv.raw<-rep(NA, n)
  for (i in 1:n) {
    svmest <- svm(Species~., type = "C-classification", kernel = "linear", data = iris2[-
i,])
    svmpred <- predict(svmest, iris2[i,])
    mcr.cv.raw[i]<- mean(svmpred!=iris2$Species[i])
  }
  mcr.cv<-mean(mcr.cv.raw)
  cat("cross-validation mcr for SVM:", mcr.cv, "\n\n")
}

tree.classification <- function(iris2) {
  fit <- rpart(Species ~ ., data = iris2, method = "class")
  pred.tr<-predict(fit, iris2, type = "class")
  mcr.tr <- mean(pred.tr!=Species)
  cat("empirical mcr for classification tree :", mcr.tr, "\n")

  mcr.cv.raw<-rep(NA, n)
  for (i in 1:n) {
    fit.tr <- rpart(Species ~ ., data = iris2[-i,], method = "class")
    pred <- predict(fit.tr, iris2[i,], type = "class")
    mcr.cv.raw[i]<- mean(pred!=Species[i])
  }
  mcr.cv<-mean(mcr.cv.raw)
  cat("cross-validation mcr for classification tree:", mcr.cv, "\n\n\n")
}

pca.iris<-prcomp(iris[,1:4], scale=TRUE)
Species <- iris$Species
n <- length(Species)
for (k in 1:4){
  cat("K = ", k, "\n")
  data.pca <- as.matrix(pca.iris$x[, 1:k])
  iris2 <- data.frame(Species, data.pca)
  lgr.classification(iris2)
  svm.classification(iris2)
  tree.classification(iris2)
}
```

The summary table of mcrs is on the last page.

```
## K =  1
## empirical mcr for logistic regression: 0.07333333
## cross-validation mcr for logistic regression: 0.07333333
##
## empirical mcr for SVM: 0.07333333
## cross-validation mcr for SVM: 0.08
##
## empirical mcr for classification tree : 0.06666667
## cross-validation mcr for classification tree: 0.1066667
```

```
##
##
## K =  2
## empirical mcr for logistic regression: 0.08
## cross-validation mcr for logistic regression: 0.08
##
## empirical mcr for SVM: 0.08666667
## cross-validation mcr for SVM: 0.08666667
##
## empirical mcr for classification tree : 0.06666667
## cross-validation mcr for classification tree: 0.1066667
##
##
## K =  3
## empirical mcr for logistic regression: 0.01333333
## cross-validation mcr for logistic regression: 0.02666667
##
## empirical mcr for SVM: 0.02666667
## cross-validation mcr for SVM: 0.04666667
##
## empirical mcr for classification tree : 0.06666667
## cross-validation mcr for classification tree: 0.14
##
##
## K =  4
## empirical mcr for logistic regression: 0.01333333
## cross-validation mcr for logistic regression: 0.02
##
## empirical mcr for SVM: 0.02
## cross-validation mcr for SVM: 0.02666667
##
## empirical mcr for classification tree : 0.06666667
## cross-validation mcr for classification tree: 0.14
```

I would like to choose logistic regression with K=4 because it has the lowest cross-validation error among all the models.

empirical	CV	K=1		K=2		K=3		K=4	
Logistic regression		7.33%	7.33%	8%	8%	1.33%	2.67%	1.33%	2%
SVM		7.33%	8%	8.67%	8.67%	2.67%	4.67%	2%	2.67%
Classification tree		6.67%	10.67%	6.67%	10.67%	6.67%	14%	6.67%	14%