Math3346, October 3, 2006.

Assignment 4, due October 20, 2006.

This assigment is based on Laboratory Exercises 9. Please forward an attached pdf file that has your assignment.

1 Different measures of accuracy

Explain the difference between training set (or resubstitution) accuracy, cross-validation accuracy, and test set accuracy? Under what circumstances is test set accuracy a better measure than cross-validation accuracy? Under what circumstances is cross-validation accuracy preferable? Why do we ignore `rpart`'s"relative error" when deciding what pruning is necessary? How does one ensure that model tuning and/or variable selection do not bias cross-validation assessments?
[2 marks]

2 `rpart` models

Fit an rpart model to the `Pima.tr` data, from the *MASS* package:

```
library(rpart)
Pima.rpart <- rpart(type ~ ., data=Pima.tr, method="class")
plotcp(Pima.rpart)
```

Now set the parameter `cp` so that it is possible to see where the cross-validation accuracy reaches a minimum.

2a Repeat the above several times. Create a table that shows how the average of the cross-validated error rate over the several runs varies with the number of splits. Why does the cross-validation error rate vary somewhat from run to run? Choose the optimum size of tree based on the minimum cross-validated error rate.
[1 mark]

2b What is the argument for using the one SE rule when determining the optimum size of tree? It is not obvious how to get a SE for the average over all runs, and the choice will be to an extent arbitrary. Suggest an ad hoc way to obtain a "standard error" that you might use. What are the potential consequences (i) if the "SE" is larger than is optimal; (ii) if it is smaller than is "optimal"? Apply your ad hoc rule, and choose the size of tree accordingly.
[2 marks]

2c Prune the model back to give the optimum tree, as determined by your use of the one SE rule. Show the confusion matrix. This is most easily done using the function `xpred()`. Plot a graph that shows the variation in expected error rate, for this model, with the proportion that has `type = Yes`.
[1 mark]

3 Analysis Using *svm*

Determine cross-validation (for `Pima.tr`) and test set accuracies when the `svm()` function from the *e1071* package is used, with the default choice of parameters. Why is the comparison of this result with `rpart()` perhaps unfair to `svm()`?
[2 marks]

4 Analysis Using *randomForest*

4a Determine cross-validation (for `Pima.tr`) and test set accuracies when the `randomForest()` function from the *randomForest* package is used. Show both the OOB confusion matrix and the confusion matrix that is calculated from the test data.
[1 mark]

4b Write a function that takes the combined `Pima.tr` and `Pima.te` data, and

- randomly splits it into two parts such that the training data have 132 out of the total of 355 of `type = No`, and 68 out of the total of 177 of `type = Yes`, as in the split into `Pima.tr` and `Pima.te`
- determines cross-validation and test set accuracies as before.

[A function `sampleFun()` is given below that should help get you started. If you use it, first run a test that demonstrates that you understand what the function does.]
[2 marks]

4c Run the function that was created in 4b ten times, and compare the OOB accuracy with the test set accuracy? Plot the test set accuracy against the OOB accuracy. Is there a consistent pattern?
[1 mark]

5 Discrimination with Multiple Groups

The package *mclust* has the data set `diabetes`. It does not automatically become available when the package is attached; instead, it is necessary to bring it into the workspace by typing `data(diabetes)`

5a Use `randomForest()` to create a model that uses variables 2 – 4 to predict `class`. Obtain the proximities. Then use the function `cmdscale()` to obtain a two-dimensional representation of the points.
[1 mark]

5b Now use `randomForest()` to predict `class`, using the ordinates from the use of `cmdscale()`. Compare the accuracy with the accuracy from the use of `randomForest()` with the columns of `diabetes`. What does this tell you about the adequacy of the proximity-based measure as a distance between points?

Be sure to run the whole sequence of calculations several times.

What point(s) stand out in the graph? In which cell(s) of the confusion matrix do these appear?
[2 marks]

Supplied Function

```
"sampleFun" <-
  function(df1=Pima.tr, df2=Pima.te){
    tab1 <- table(df1$type)
    df <- rbind(df1, df2)
    rownum <- seq(along=df[,1])
    Ry <- rownum[df$type=="Yes"]
    Rn <- rownum[df$type=="No"]
    Ry1 <- sample(Ry, tab1[2])
    Rn1 <- sample(Rn, tab1[1])
    rowset1 <- c(Ry1, Rn1)
    newdf1 <- df[rowset1,]
    newdf2 <- df[-rowset1,]
    list(tr=newdf1, te=newdf2)
  }
```