Math3349 – Data Mining: Assignment 2

October 16, 2012

Marks will be given for layout and for clarity of exposition. Please enter your results into a pdf file, and submit by 5pm on Mon November 5 to `john.maindonald@anu.edu.au`.

This assignment will work with the data in the `ticdata` dataset from the *kernlab* package. This contains product usage and socio-demographic data derived from zip area codes. The challenge is to use the training data to build a model that, applied to the test data (or to other data from the same population), will identify the subset of 800 who are most likely to buy an insurance. Data can alternatively be obtained, separated into separate "training" and "test" sets, from the UCI KDD Archive at `http://kdd.ics.uci.edu`.

Start by running the code:

```
library(kernlab)
data(ticdata)
## Use first 5822 observstions for training
tic0 <- ticdata[1:5822, ]
tic1 <- ticdata[-(1:5822), ]
```

Exercises

1. Write notes on the following:

    a) Comment on the distinction between source population and target population. Under what circumstances does the distinction have major implications for the interpretation and use of results. Give an example, perhaps relating to a dataset used in the course.

    b) Distinguish between training set accuracy, cross-validation accuracy, OOB accuracy and test set accuracy. There is however a measure of accuracy that, if available, is definitive for the way that results of the analysis are commonly used. What is that measure?

c) Under what circumstances is training set accuracy likely to be seriously biased? Under what circumstances can one expect cross-validation accuracy to be a good measure? Under what circumstances is OOB accuracy a good measure?

d) Comment on issues that arise for the use of the leave-one-out cross-validation measure of accuracy. Are there cautions that need to be observed?

e) Suppose that we have a dataset of 5000 observations in which the first 5000 are from the year 2010, while the final 5000 are from the year 2011. Suggest three ways in which the data might be split into training and test sets. What are the advantages and disadvantages of each? Is there, with data of this type, a better strategy than a single split into training and test set?

2. Here are alternative model fits:

```
ticm1.rf <- randomForest(CARAVAN ~ ., data=tic0[,-1])
ticm2.rf <- randomForest(CARAVAN ~ ., sampsize=c(348,348),
                         data=tic0[,-1])
```

a) Compare the confusion matrix estimates and the overall accuracy between these two model fits.

b) The results of the analysis will be used to select 1000 individuals for targeting for an advertising campaign. A reasonable way to select such a group is to choose the 1000 individuals who, based on the model, seem most likely to buy policies. Use the list element `votes` for this purpose.

```
nr <- (1:nrow(tic0))[order(ticm1.rf$votes[,2],
                           decreasing=TRUE)[1:1000]]
## Check the number of caravan insurance purchasers
table(tic0[nr, 86])
```

Check how the number of (caravan) insurance purchasers that are in the sample of 1000 varies with the number sampled from those who had not purchased insurance. What choice of the `sampsize` argument is for this purpose close to optimal? You may find the following function useful:

```
bestsize <- function(n0=696, mtry=9, nselect=800,
                     form=CARAVAN ~ ., data=tic0[,-1])
```

```
{
    ticm.rf <- randomForest(form, sampsize=c(n0,348),
                            mtry=mtry, data=data)
    nr <- (1:nrow(tic0))[order(ticm.rf$votes[,2],
                           decreasing=T)[1:nselect]]
    sum(tic0[nr, 86]=="insurance")
}
```

[The setting `mtry=9` is the default for calling `randomForest()` with the present data. In the next question, we will want the option to vary it.]

3. The model can be further tuned by varying the parameter `mtry`.

    a) Investigate whether such tuning can be used to increase the estimated number of insurance purchasers in a sample of 1000?

    b) The function `tuneRF()` can be used to tune the choice of `mtry` to give maximum predictive accuracy. Investigate whether, and if so how, the optimum choice of `mtry` varies with the `sampsize` setting. (NB: One can pass `sampsize` as an argument to `tuneRF()`.)

    What pattern of difference, if any, do you observe between (a) and (b) in the value(s) of `mtry` that is (are) optimal?

4. Now check whether the variable `STYPE`, thus far omitted from consideration because `randomForest()` cannot handle a factor with 37 levels, may be a useful predictor. The idea is to check which of the 37 levels commonly appear together at the same terminal node. The following code is for illustration, and will need to be adapted to what you have learned from working through earlier questions:

```
## Substitute more optimal choices of sampsize and mtry.
## Choose a more managable subsample
subrow <- sample(1:nrow(tic0), 2000)
tic9 <- tic0[subrow, ]
tab9 <- table(tic9[,86])
ssize <- as.vector(rep(tab9[2], 2))
ticm9.rf <- randomForest(CARAVAN ~ ., sampsize=ssize, mtry=7,
                         data=tic9[,-1], proximity=TRUE)
pts <- cmdscale(1-ticm9.rf$proximity)
```

Now plot the points, identifying the levels of `STYPE`:

```
xyplot(gpts[,3]~gpts[,2], panel=function(x,y,...)
                panel=panel.text(x,y,labels=paste(1:37)))
```

Use the plot as a basis for forming a new factor that groups existing levels into perhaps 5 or 6 levels of the new factor. Does including this new factor in the random forest model improve predictive accuracy?

5. Use the random forest model that you finally identify to make predictions for the `tic1` data, then identifying the 1000 observations that are from individuals who are predicted as most likely to be insurance holders. How many of these do you find to hold insurance?

6. This question will investigate use of the function `lda()`. Break the `tic0` data in two parts – `tic00` consisting of the first 4000 observations, and `tic01` consisting of observations 4001 to 5822. Fit the model to the `tic0` dataset, and determine: (a) the training set accuracy; (b) the leave-one-out cross-validation accuracy; and (c) the test set accuracy, using `tic01` as the test data. Do these agree? If not, how can the differences be explained?

Note: Include any additional R code that you have used in an appendix to your assignment. Be careful to identify the question to which the code relates.

R code from assignment:
`http://www.maths.anu.edu.au/~johnm/courses/mathdm/2012/assignment2.R`