36-350: Data Mining

Lab 13
Date: November 21, 2003 **Due: end of lab**

Interspersed throughout this lab are questions that you will have to answer at check-off.

1. Download the files for this lab from the course web page to the desktop:

   `http://www.stat.cmu.edu/~minka/courses/36-350/lab/`

2. Open a Word or Notepad document to record your work.

Start R

3. `Start -> All Programs -> Class software -> R 1.7.0`

4. Load the special functions for this lab:

   `File -> Source R code...`

   Browse to the desktop and pick `lab12.r` (it may have been renamed to `lab12.r.txt` when you downloaded it). Another window will immediately pop up for you to pick the `mining.zip` file you downloaded.

The dataset

5. The dataset is the same as lab 12, with training data in `x.tr` and test data in `x.te`. Two other matrices, `nx.tr` and `nx.te`, have the predictors numerically coded, for use with logistic regression. You want to predict `Class`.

Misclassification rates

6. These are the proportions of good and bad loans in the test set:

```
  Bad  Good
0.316 0.684
```

   What is the misclassification rate of a classifier which always reports "Good"? What is the misclassification rate of a classifier which always reports "Bad"? The minimum of these two is the *baseline rate*. Any classifier which does worse than the baseline is essentially worthless.

7. As in the last lab, construct a pruned classification tree and a k-nearest-neighbor classifier with $k = 6$. *What are their misclassification rates on the test set?*

8. Construct a linear classifier on the all-numeric version of the training set. *What is its misclassification rate on the test set?*

9. The variable `fmla` contains a formula with the most important predictors. Use quadratic expansion on this formula to build a quadratic classifier. *What is its misclassification rate on the test set? Of the four classifiers, which do better than baseline?*

Misclassification costs

10. The variable `costs` gives the cost to the bank of different types of misclassification:

```
      predicted
truth  Bad Good
  Bad    0    5
  Good   1    0
```

    What is the average cost (total cost divided by the test set size) of a classifier which always reports "Good"? What is the average cost of a classifier which always reports "Bad"? The minimum of these two is the *baseline cost*.

11. To minimize cost, the bank should say "Good" only when the probability of "Good" exceeds 5/6. For each of the four classifiers, compute the average cost of this policy on the test set. *Which are below baseline?*

12. You can now get checked off. Save all of your results for the homework.

Logistic regression The `logistic` function is similar to `tree`:

```
fit = logistic(<formula>,<data>)
summary(fit)
```

Deviance and misclassification rate are also the same:

```
deviance(fit,<data>,rate=T)
misclass(fit,<data>,rate=T)
```

Confusion matrix A confusion matrix cross-classifies the predictions of a model and the true responses. The model says "Yes" when the probability of "Yes" exceeds `p` (which is 0.5 by default).

```
confusion(<model>,<data>,p)
```

If `costs` is a corresponding table of costs, this will compute the total cost of the classifier on the data set:

```
sum(confusion(<model>,<data>,p)*costs)
```

Quadratic expansion To create a formula with quadratic terms added, use one of

```
expand.quadratic(fit)
expand.quadratic(<data>)
expand.quadratic(<formula>)
```

(Just like `expand.cross` from lab 10.) `expand.quadratic` assumes all predictors are numeric. The resulting formula may be very big and cause R to run slowly.