

Lab 9

The “Pima Indian Diabetes dataset” (see attached information) contains information on various medical measurements on 768 individuals, and an indicator of whether they later developed diabetes.

I’ve put the data at

<http://www.stat.washington.edu/stephens/S423/pima.txt>

The columns of the data are separated by commas - if you don’t know how to read these kinds of data into R, try `?read.table`. Note also that there is no header to the file (a header is a row containing the column names)

Since the file does not contain a header, the columns are unnamed. You might find it useful to attach some names to the columns for easy reference, which you can do using (if your dataframe is called `pima`):

```
colnames(pima) <-  
c("preg", "pg", "dbp", "tricep", "insulin", "bmi", "pedig", "age", "diabetes")
```

You can then `attach(pima)` and refer to columns by name.

1. Some of the cases have “missing values”, which have unfortunately been coded as 0s. In some cases it is impossible to tell whether a value is really a zero, or missing: I suggest you assume these observations are really zero. In other cases it seems implausible that the real value is zero, so the value must be missing: remove these missing values by removing appropriate columns and/or rows from your dataframe. (Remember the “didFinal” example, for how to remove rows; removing columns is similar). Explain how you decided which rows or columns to remove.
2. Plot some plots (histograms, barplots, scatterplots) to try to determine which variables appear to be good predictors of diabetes onset. (Note: colored scatterplots can be helpful for showing the distribution of two variables for different classes; eg try `plot(pg,dbp,col=diabetes+1)`. You might also try `pairs(pima,col=diabetes+1)`. Note: we need the +1 so that the colors used are 1(black) and 2(red) for the groups 0 and 1. (Color 0 is white, and so doesn’t show up!)
3. Try picking a single variable that looks like it might predict well, and fit a logistic regression (Remember: `glm(y ~ x,family = ‘binomial’)`) using that variable to predict whether each individual will have diabetes. Can you interpret the sign of the coefficient you obtain from your regression? Is it plausible that the risk is monotone in that variable? Try plotting a plot of

the regression curve (use the fitted values in your model object). What is the equation of this curve? How many of the examples in the training set would your classifier predict correctly, if you threshold at probability=0.5? (the `table` command may be useful here).

4. Try fitting a logistic regression of diabetes on all the variables. Remove non-significant variables, and see if this increases the AIC score. Try finding a model that has a good AIC score by trial and error. How well does it predict on the whole data set (using the threshold of 0.5 as above)? Why must we be careful about results obtained on the data set used to train the classifier? Test its performance more carefully, using cross-validation, splitting your data into equal-sized test and training sets. Is the accuracy higher or lower than when you tested on the whole data? Is this what you would expect?
5. Remember that logistic regression is in some sense the optimal procedure when the x values in the two groups you are comparing are normally distributed with the same variance. Do any of the x variables in the model you chose using AIC look like they might not be very normal, but could be improved by a transformation? If so, try to transform each x so that it looks close(r!) to normal in each group, and then redo the cross-validation test to see if it improves accuracy.
6. What is the effect of changing the probability “threshold” for classification to values bigger or smaller than 0.5? (note the effect on “false positives” vs “false negatives”.) What might you do if false negatives were considered more important to avoid?