Lab 04B: Week 12

Contents

The **specific aims** of this lab are:

- estimate logistic regression models
- identify which variables are significant and perform model selection
- interpret the coefficients from a logistic regression model
- use decision trees and random forests for classification
- evaluate out-of-sample performance using cross validation
- use estimated prediction/classification models to predict the outcomes for new observations

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO7 Perform statistical machine learning using a given classifier, and create a cross-validation scheme to calculate the prediction accuracy.

1 Questions

1.1 Who has diabetes?

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases (Johannes 1988). The objective is to predict whether or not a patient has diabetes based on certain clinical measurements. The larger database of patients has been subset such that all observations here are females at least 21 years old of Pima Indian heritage.

The data sets consists of several medical predictor variables and one target variable, `y` which equals 1 if an individual is diabetic and 0 otherwise. Predictor variables includes the number of pregnancies the patient has had (`npreg`), their BMI, insulin level (`serum`), age, triceps skin fold thickness `skin`, diastolic blood pressure (`bp`), plasma glucose concentration (`glu`) and diabetes pedigree function (`ped`).

It is available from various places, including Kaggle and the **reglogit** R package and I've uploaded a copy to the GitHub server.

```
library(tidyverse)
pima = readr::read_csv("https://raw.githubusercontent.com/DATA2002/data/master/pima.csv")
glimpse(pima)
```

```
Rows: 768
Columns: 9
$ npreg <dbl> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, 5, 7, 0, …
$ glu   <dbl> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125, 110, 16…
$ bp    <dbl> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92, 74, 80, 60,…
$ skin  <dbl> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0, 0, 23, 19, 0…
$ serum <dbl> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0, 0, 846, 175,…
$ bmi   <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.3, 30.5, …
$ ped   <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.248, 0.134…
$ age   <dbl> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 34, 57, 59…
$ y     <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, …
```

1. Visualise the data and look for any obvious relationships or patterns in the data. Perform any data cleaning as appropriate.

1.1.1 Logistic regression

2. Fit a logistic regression to the data and perform backward stepwise model selection using the AIC.
3. In the stepwise model, increases in which variables lead to higher odds of diabetes?
4. Write down the fitted stepwise model.
5. Predict the log-odds and the probability of a 35 year old woman who has been pregnant twice, with a BMI of 30, blood pressure of 72, glucose of 122 and diabetes pedigree function of 1. Compare it to a 50 year old woman with the same measurements, except a BMI of 40.
6. Generate a **confusion matrix** to assess the in-sample accuracy of the predictions from the stepwise model noting that the positive class is the presence of diabetes.
7. What is the sensitivity and specificity of using our stepwise model as a diagnostic tool to predict diabetes?
8. Perform 5 fold cross validation to get an idea of out of sample accuracy for the stepwise model.

1.1.2 Random forest

9. Fit and visualise a decision tree to this data.
10. Predict the outcomes for the two women described earlier using your estimated tree.
11. Evaluate in-sample performance using a confusion matrix (e.g. using the `confusionMatrix()` function from the **caret** package).
12. Evaluate out-of-sample performance using 5 fold cross validation
13. Fit a random forest and assess out of bag performance.

1.1.3 Comparison

17. Compare the out of sample accuracies for the different methods using 5 fold CV with 10 repeats.
18. Which model do you prefer?
19. Are our results generalisable to other populations?

1.2 Rock wallabies

Macropods defaecate randomly as they forage and scat (faecal pellet). Surveys are a reliable method for detecting the presence of rock-wallabies and other macropods. Scats are used as an indication of spatial foraging patterns of rock-wallabies and sympatric macropods.

Tuft et al. (2011) investigate a rock-wallaby colony in the Warrambungles National Park (NSW). They sampled $n = 200$ sites and recorded the presence or absence of scats as 1 (present) or 0 (absent).

Scats deposited while foraging were not confused with scats deposited while resting because the daytime refuge areas of rock-wallabies were known in detail for each colony and no samples were taken from those areas. Each of the 200 sites were examined separately to account for the different levels of predation risk and the abundance of rock-wallabies.

We will consider five main effects:

- `edible`: Percentage cover of edible vegetation
- `inedible`: Percentage cover of inedible vegetation
- `canopy`: Percentage canopy cover
- `distance`: Distance from diurnal refuge
- `shelter`: Whether or not a plot occurred within a shelter point (large rock or boulder pile)

As well as three interaction terms: - `edible * distance` - `edible * shelter` - `distance * shelter`

The data can be found in the **mplot** package:

```
# install.packages('mplot')
data("wallabies", package = "mplot")
glimpse(wallabies)
```

```
Rows: 200
Columns: 8
$ rw       <int> 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, …
$ edible   <int> 1, 0, 10, 0, 0, 10, 20, 15, 35, 25, 25, 10, 0, 20, …
$ inedible <int> 15, 0, 0, 50, 10, 50, 15, 50, 25, 30, 60, 100, 25, …
$ canopy   <int> 0, 0, 20, 40, 50, 0, 0, 10, 0, 0, 0, 0, 0, 0, 0, 20…
$ distance <int> 128, 131, 137, 136, 138, 140, 141, 141, 139, 138, 1…
$ shelter  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ lat      <dbl> -31.25447, -31.25456, -31.25461, -31.25468, -31.254…
$ long     <dbl> 148.9408, 148.9408, 148.9409, 148.9409, 148.9409, 1…
```

1. Visualise the data.
2. Fit the full logistic regression model including the five main effects and three interaction terms.
3. Perform stepwise selection using the AIC from the full model to identify a simpler model.
4. Write down the fitted stepwise model.
5. If you were to use a backward selection p-value approach to drop another variable from the stepwise model selected using AIC, which would you drop?
6. Use the **caret** package to compare the out of sample performance of the full model and the stepwise model with a simple model that only uses `edible` as a predictor. Use repeated 10-fold cross validation with 20 repeats.

2 Additional resources

For more details on decision trees see Hastie, Tibshirani, and Friedman (2009, sec. 9.2) and James et al. (2017, chap. 8).

As additional practice questions, I recommend these two DataCamp chapters:

- </> Logistic Regression
- </> Decision trees

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. New York, NY, USA: Springer. https://web.stanford.edu/~hastie/ElemStatLearn/.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2017. *An Introduction to Statistical Learning: With Applications in r*. New York: Springer. https://www-bcf.usc.edu/~gareth/ISL/.

Jed Wing, Max Kuhn. Contributions from, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2018. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Johannes, R S. 1988. "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus." *Johns Hopkins APL Technical Digest* 10: 262–66.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Therneau, Terry, and Beth Atkinson. 2018. *Rpart: Recursive Partitioning and Regression Trees*. https://CRAN.R-project.org/package=rpart.

Tuft, K. D., M. S. Crowther, K. Connell, S. Müller, and C. McArthur. 2011. "Predation Risk and Competitive Interactions Affect Foraging of an Endangered Refuge-Dependent Herbivore." *Animal Conservation* 14 (4): 447–57. https://doi.org/10.1111/j.1469-1795.2011.00446.x.