Problem Set 5: Predictive Models

Instructions

Load the "Session10PimaDiabetesCleanRand.csv" file (available in the ProblemSet5/PS5 folder on GitHub). This is a cleaned up version of the data from session 5 with 725 individuals, 7 independent/predictor variables and Diabetes is a binary response variable (1=diabetes, 0=no diabetes).

Make the first two thirds of the samples (samples 1 to 483) into a training set and the last third into a test dataset (samples 484 to 725). For simplicity, do not worry about scaling the independent variables. Use two different predictive modeling methods to create a classifier with the training data and then determine how well they do in predicting diabetes in the test data set.

1. Use logistic regression + stepwise variable selection (`stepAIC` function in the MASS package) to **find the "best" model**.
2. Your choice: choose one method among decision trees, neural networks, or random forests.

For the logistic regression analysis, **report the independent variables included and excluded in the "best" model.** For each method/model you fit, **report the three most important predictive variables** (they may not be the same across methods).

Note: If you use neural networks with the `nnet` package, the `olden` function from the "NeuralNetTools" packages can be used to assess relative importance of the variables in the model. You can get variable importance from a decision tree with `mod$variable.importance` for the decision tree or from the summary function.

For the each method/model fit, **report the training data area under the curve (AUC) and the test data AUC.**

Answer the following questions

- For each method/model fit, is the training AUC better or worse than the test AUC?
- In comparing the predictive ability of the two models with each other, is better to compare the AUC from the training data or from the test data? Why?

- In general, is the test AUC expected to be larger or smaller than the test AUC? Why? Is this always the case?
- Which of the two methods/models fits/performs better?

Data Source

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988), ed. R. A. Greenes, pp. 261–265. Los Alamitos, CA: IEEE Computer Society Press.

The report

Develop a report (I recommend a Word (or other text editor) document) for your problem set that includes answers to all of the questions posed above, showing plots where appropriate.

Save your report as a PDF file and submit your report through the course 2GW site. Clean up your code and submit it as a supplementary file, along with your main report.

Due date

Day 7, Week 10