Machine Learning in Advertising Technology

Vangie Shue

December 12, 2014

Abstract

This project will explore the application of Machine Learning in Technology. Specifically, we will seek to develop a reliable process for developing a classification algorithm on given data.

Introduction

Background on Advertising Technology use cases. Background on Cafemom and the specific problem: classify hispanic users

Classifying Rare Events

(1, pg141) Rare events are more statistically informative that zeros is seen in the variance matrix (1, p142) When sampling, we must be careful not to select on X differently for the two samples. (2) The problem is that maximum likelihood estimation of the logistic method is well-known to suffer from small-sample bias. The penalized likelihood or Firth method are the general approach to reducing small-sample bias.

Classification Algorithms

Logistic Regression

Generalized Linear Model

(4) glmnet fits a generalized linear model via penalized maximum likelihood. (6) it can deal with all shapes of data, including very large sparse data matrices.

Generalized Boosted Model

(7) Advantages over logistic regression: robust to outliers, can make predictions on missing data, handles unequal class sizes and unbalanced predictor variables, tend to have greater predictive ability (7) Drawbacks: trees can overfit especially if the number of ending nodes is small or number of trees is too large, definitely want to use cross validation (7) Use prediction rate as a measure of goodness of fit

Methods

Data Collection

Demdex has TraitsSegments, uuid and traitssegments collected Table generating enes language to determine hispanicnon-hispanic, 1 for es, 0 for en

- segment_hispanic: 80324851 total uuids
- segment_hispanic: 1698878 hispanic (2 percent)
- segment_hispanic2: 9733751 total uuids
- segment_hispanic2: 1868091 hispanic (19 percent)

Data Processing

alldata: 7193606 x 2105480 dataset: 1438721 x 44336 traindata: 1294849 x 44336 testdata: 143872 x 44336

Logistic Regression

We used `glmnet` to generate logistic regression model. Then `predict.glm` is used to predict on the validation sample. (3) We then determine the how low of a predicted probability is needed to accurately classify the hispanic division.

Generalized Boosted Regression Model

(7) we use distribution = bernoulli since this is a binary classification (use gaussian for adaboost), for n.trees we use a large number of trees that we can prune back later. The shrinkage is the step size, which we chose to be 0.1 since the a smaller step would take longer to model although it would yield better performance. Use cross-validation to determine interaction depth. Decreasing n.minobsinnode increases in-sample fit but risks overfitting. nTrain is used so that you can select the number of trees at the end.

Random Forest or Ferns

Support Vector Machines

```
> data(example)
```

The above is a snippet of code used.

Results

Compare time for execution & accuracy between modelsmethods. ex.UnweightedNon-filteredLogistic Regression

The below is a sample graph of data.

Something like a plot centered.

Conclusion

We demonstrated the application of Machine Learning in Advertising Technology, in particular for rare events.

Cannot accept comparisons without consideration to the implementation. Some may provide more "tuning" than other algorithms and therefore appear more accurate. However, we demonstrate which algorithm will work best for our usage.

(8) need to look into using spark with R for much larger data computation.

Acknowledgements

Patrick McCann & Cafemom, Professor Mohri

References

We used `RStudio Sweave` to build this LaTeX document

- 1 http://gking.harvard.edu/files/gking/files/0s.pdf (rare events)
- 2 http://www.statisticalhorizons.com/logistic-regression-for-rare-events (rare events)
- 4 http://cran.r-project.org/web/packages/glmnet/glmnet.pdf
- 3 http://stats.stackexchange.com/questions/25389/obtaining-predicted-values-y-1-or-0-from-a-logistic-regression-model-fit
- 5 http://machinelearningmastery.com/an-introduction-to-feature-selection/ (feature selection tips)
- best algorithms: http://www.researchgate.net/post/What_is_the_best_algorithm_for_classification_task
- http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/SVM (SVM)
- http://www.jstatsoft.org/v61/i10/paper (random ferns)
- http://www.statmethods.net/advstats/cart.html (rpart and random forest)
- 6 http://www.inside-r.org/packages/cran/glmnet/docs/glmnet
- 7 http://vimeo.com/71992876 (gbm)
- 8 http://datasaucer.blogspot.com/ (R for big data)