NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

PREDICT 412: Advanced Modeling Techniques
Term Project Topics

Authorship Identification (AUTHOR)

This project involves the development of classification models to match authors with Amazon book reviews. The data set is part of the Machine Learning Repository at the University of California-Irvine: http://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set .There are 30 observations for each of 50 authors and 10,000 attributes for each observation. Experiment with alternative classification methods, including committee and ensemble methods. Utilize various methods for evaluating classifier performance. Recommend a model and describe its predictive accuracy. Izenman (2008) and Duda, Hart, and Stork (2001) are good references for classification methods. Also relevant to the authorship identification problem is the classic work by Mosteller and Wallace (1984).

Investment Analysis (INVEST)

This project involves the development of a model to predict the price of stock tomorrow based upon past stock prices and economic indicators. Pick a public company. Obtain the stock price time series for this company and other companies in the same industrial sector. Obtain stock market and economic index time series for the same period. Develop models for predicting the price of the company's stock using traditional econometric and time series methods. For the same time period, gather macroeconomic measures relating to the United States economy. Determine the degree to which these macroeconomic measures may be used to improve your forecast of future stock prices.

Collaborative Filtering of Jokes (JESTER)

Collaborative filtering methods have been applied in recommender systems for books, articles, movies, music, matchmaking, and Web locations. These advanced modeling techniques are important to firms like Amazon.com, Netflix, Pandora, and Google. A related area is affinity or market basket analysis. This project involves the exploration of collaborative filtering algorithms in the analysis of the Jester data set available at http://eigentaste.berkeley.edu/dataset/ , which was described by Goldberg, et al. (2001). Use the first Jester data set, which is available as three Excel spreadsheet files. Responses were obtained using an online survey of more than seventy thousand volunteers, with each volunteer rating a subset of 100 jokes between April 1999 and May 2003. Ratings were obtained using a continuous sliding scale from -10.00 to +10.00. A useful R package for this project is recommenderlab (Hahsler, 2012).

Movie Map Magic (MOVIES)

This project involves text analytics, multivariate methods, and classification methods. This project is takes an approach to the problem of finding movies that we like, an approach that builds upon techniques like multidimensional scaling. No customer data are used in this analysis. Rather we first build a model from the movie scripts alone, the text data. We select twenty to thirty movie script cohort pairs as our text data corpus, divide the cohort pairs into movie training and test sets, extract features from the corpus, and use multivariate analysis (multidimensional scaling, perhaps, see Izenman, 2009) to create a map of the training set movies in multi-dimensional space (trying two dimensions initially, but using more dimensions as needed). We use a distance-based clustering procedure to identify clusters of movies in this multi-dimensional space. A classification model is then built on the training set using the text features as input and the movie cluster as the response/class. Finally, we test our model on the movie test set to see how many movies are correctly classified (that is, matched up with or close in space to their cohorts). To ensure consistent text data selection and coding in the initial movie scripts, we use a single public-domain data source for the movies. One such source is Drew's Script-o-Rama at www.script-o-rama.com. Another is the Internet Movie Database (IMDb).

NFL Sports Analytics (NFL)

Who will win the next game and by how many points? Many have tried to predict the outcome of sporting events. Many have failed. Some claim to have succeeded. Using data from the previous NFL seasons, develop a model for predicting which team will win the next game. Evaluate the predictive accuracy of the model and develop a plan for using the model to predict who will win scheduled games in this year's NFL season. Miller (2008 and 2014) suggests various modeling techniques and provides references to relevant literature in sports analytics. The NFL data set was obtained from NFLdata.com.

Social Media and Privacy (PRIVACY)

This project involves an investigation of the degree to which social media data can be used to infer personality characteristics, as measured by a Big Five personality questionnaire. The social media data consist Twitter postings. Paths to the source material for this modeling project may be found on kaggle:
http://www.kaggle.com/c/twitter-personality-prediction

R Recommender System (RECOMMEND)

This project involves a recommendation system for R packages, drawing upon a competition sponsored by kaggle. The competition presented 99,640 records of data for 52 R users. A description of the data for this competition is available at http://www.kaggle.com/c/R. Conway and White (2012) show how to approach this problem using a nearest neighbor algorithm. A useful R package for this project is recommenderlab (Hahsler, 2012).

R Package Structure (STRUCTURE)

This project involves using methods of network analysis to examine the package-to-package structure of the R system. Data are gathered from the R package documentation, which includes information on source code interdependencies across R packages ("depends" and "reverse depends"). The task is to define a network diagram for the R system or for a subset of the R system. Application of social network measures may provide additional insight into the structure of the system. A useful R package for this project is sna (Butts, 2013).

References

Butts, C. T. (2013). sna: Tools for social network analysis. Comprehensive R Archive Network. http://cran.r-project.org/web/packages/sna/sna.pdf

Conway D. & White, J. M. (2012). *Machine learning for hackers.* Sebastopol, Calif.: O'Reilly.

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business School Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed). New York: Wiley.

Goldberg, K, Roeder, T., Gupta, D. & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4*(2), 133-151.

Hahsler, M. (2012). recommenderlab: Lab for developing and testing recommender algorithms. R package retrieved from the World Wide Web at http://cran.r-project.org/web/packages/recommenderlab/index.html

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning.* New York: Springer.

Miller, T. W. (2008). *Without a tout: How to pick a winning team.* Madison, Wisc.: Research Publishers.

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson.

Mosteller, F. & Wallace, D. L. (1984). *Applied Bayesian and classical inference: The case of the Federalist papers.* New York: Springer.