Syllabus
STAT 695
Competitve Predictive Modeling
Spring, 2016
(2 credits)

Description

Statistical machine learning has became the central tool for a large number of research and practical fields, such as business decision making, imaging processing, and detecting disease relevant factor, and particularly predictive modeling. A vast amount of statistical tools and models have been discussed in literature for predictive modeling from both theoretical and methodological perspectives. In this course, instead of focusing on the theoretical aspects, students will gain extensive practice in building and testing predictive models through directed participation in predictive modeling competitions, including Kaggle (`https://www.kaggle.com/`) and the Data Mining Cup (`http://www.data-mining-cup.de/en/`). Competitions will include predicting responses of mixed types. Practical utilization of statistical learning tools will be discussed along with the competition. In addition, students will gain extensive practice in using R software for data wrangling and modeling.

Prerequisites

Experience coding in R; at least one regression course such as STAT 511; permission of the instructors.

Instructors

Professor Brooke Anderson, `brooke.anderson@colostate.edu`
Office: 146 Environmental Health Building
Office hours: By appointment

Professor Wen Zhou, `riczw@colostate.edu`,
`http://www.stat.colostate.edu/~riczw/`
Office: 208 Statistics Building
Office hours: By appointment

Meetings

Tuesdays 4:00-4:45; Thursdays 3:50-4:05; Location: Stats 006

Textbooks

Required:

- Max Kuhn and Kjell Johnson, *Applied Predictive Modeling*, Springer (2013). (Available online through CSU library)
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer (2013). `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`

Recommended:

- Alan J. Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer (2008).

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer (2009) (The book is free online).
- Christopher M. Bishop, *Pattern Recognition and Machine Learning*, (2006).
- Ian Witten, Eibe Frank, and Mark Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, (2011).

Course Objectives

The primary aim of this course is to use three in-depth, practical examples of predictive-modeling challenges (two Kaggle competitions and the Data Mining Cup) to allow students to gain extensive practical experience in data processing, data matrix construction, variable selection, model fitting, model ensemble, and evaluating predictive models. With that aim, this course will introduce students to a variety of topics in machine learning to provide them with strategies and approaches to tackle these specific applied examples. Some of the challenges that will come up will be specific to these three problems, but this experience of completing three predictive modeling competitions will provide students with the conceptual understanding, programming tools, and strategies to tackle their own predictive modeling challenges. This course would also provide an excellent background for students aiming to take more theoretical courses on statistical learning in the future.

Specific objectives include:

(1) Students will work in multi-disciplinary teams to compete in the two Kaggle predictive modeling competitions.
(2) Students will learn how to fit and evaluate a variety of predictive models, including: classification and regression trees, support vector machines, logistic and linear regression models, tree ensemble models, Naive Bayes, k-nearest neighbors, and neural networks.
(3) Students will learn strategies in data wrangling and feature engineering to improve predictive models.
(4) Students will learn to use resampling methods to assess the performance of predictive models.
(5) Students will gain extensive additional experience working on complex modeling problems using R statistical software.
(6) Students will be recommend to form a team to participate the annual Data Mining Cup held from April to May. The final result will be released on July 2016, and the top 10 teams will be invited to Berlin, Germany, for the final ceremony; more details see `http://www.data-mining-cup.de/en/`, `http://magazine.amstat.org/blog/2013/10/01/iowa-state-dmc/`.

Course Topics

- Definition of machine learning.
- Classification models: K-nearest neighbors, naive Bayes, logistic regression, classification trees, bagging, random forests, boosting, support vector machines.
- Regression models: K-nearest neighbors, naive Bayes, linear regression, regression trees and tree ensembles, non-linear regression, support vector machines, ridge regression, lasso.
- Model fitting and tuning.
- Model evaluation, including with re-sampling techniques.
- Data pre-processing, feature selection, measuring variable importance, and visualizing data.
- Linear model selection and regularization and the challenges of high-dimensional data.
- Neural networks and deep learning.
- Unsupervised learning.

Course Expectations & Grading

Grades will be based on attendance, participation, regular submission of model results to each of the two Kaggle competitions, written reports on each challenge (one per group), in-class presentations on final

models for each Kaggle challenge and the DMC, and student presentations on topics from Kuhn and Johnson (2013).

Assignments & Readings

- **Week 1:** What is machine learning? Classification models: K-nearest neighbors and naive Bayes.
  Reading: James et al.: Chs. 2, 4.1-2, 4.4. Kuhn and Johnson Chs. 1, 2, 13.5-6.
  Competition: Kaggle: Surviving the Titanic
- **Week 2:** Metrics of performance of classification models. Classification models: logistic regression models and classification trees.
  Reading: James et al.: Chs. 2.2, 4.3, 4.5, and 8.1. Kuhn and Johnson: Chs. 11, 12.1-3, 14.1-2.
  Tuesday assignment: Fit some different Naive Bayes and k-Nearest Neighbors models. For a few of your models, try measuring some other metrics of performance, like sensitivity, specificity, AUC, Youden's J Index, positive predictive value, etc.
  Thursday assignment: Try fitting the following types of models: LDA (see week 1 reading), logistic regression, classification tree. Can you get any of your models to beat the Sex-only benchmark model? Can you get any to beat the to-date class best (Casey's 0.8)?
  Competition: Kaggle: Surviving the Titanic
- **Week 3:** Using resampling to measure performance of classification models. Classification models: ensemble models (bagging, random forest, boosting).
  Reading: James et al.: Chs. 5, 8.2. Kuhn and Johnson: Ch. 14.3-7.
  Tuesday assignment: Try using cross-validation, either to tune certain paramters in a model (e.g., k in k-NN), or to compare the performance of different models
  Thursday assignment: Try fitting the following types of ensemble models: bagging, boosting, random forests. Can you get any to beat the to-date class best (0.8)?
  Competition: Kaggle: Surviving the Titanic
- **Week 4:** Classification models: support vector machines.
  Reading: James et al.: Ch. 9
  Tuesday assignment: Try to fit a SVM model to the Titanic data
  Thursday assignment: Present 5 slides on your work on the Titanic prediction as well as a short write-up describing your work on the challenge.
  Competition: Kaggle: Surviving the Titanic
  Graded products: Students present final predictive model for Titanic competition.
- **Week 5:** Student presentations based on Kuhn and Johnson 2013: Data pre-processing; Over-fitting and model tuning; Remedies for severe class imbalance; Feature selection
  Reading: Kuhn and Johnson: Chs. 3, 4, 16, 19
  Tuesday: Presentations from Casey, Veronica, and Wande
  Thursday: Presentations from Julia and Ahmed. Discussion of all presentations
  Graded products: Students presentations on material from chapters in Kuhn and Johnson 2013. Presentations are (all from Kuhn and Johnson): Ch. 3.1-3.3: Casey; Ch. 3.4-3.7: Veronica; Ch. 4.6-4.8: Wande; Ch. 16.1-16.7: Julia; 19.1-3: Ahmed

- **Week 6:** Assessing performance of regression models. Regression models: Linear regression models. Over-fitting and model tuning re-visited.
  Reading: James et al.: Ch. 3. Kuhn and Johnson: Chs. 4 (re-visited), 5, 6.1-2.
  Tuesday: Explore the data for the new competition. Try using a linear model to predict the new competition data. What is the best RMSLE you can achieve in the training dataset? In the testing dataset?
  Be prepared to discuss: Why is this competition evaluated with RMSLE rather than RMSE? What pre-processing do you think might be important, based on EDA of the dataset? Can you think of any ways you might want to use the predictors to engineer new features?

Thursday: Try fitting one of the types of models that requires tuning (e.g., k-NN, random forest) using functions from the caret package. What is the best RMSLE you can achieve in the training dataset? In the testing dataset?
Be prepared to discuss: How can you adapt the caret functions to use RMSLE rather than RMSE when tuning? Are you running into any computational problems given the size of the dataset? If so, how are you handling them?
Competition: Kaggle: Bike sharing demand: `https://www.kaggle.com/c/bike-sharing-demand`

- **Week 7:** Linear model selection and regularization. Shrinkage methods and dimension reduction methods.
Reading: James et al.: Ch. 6.
Tuesday: Try using some of the variable selection methods from the reading (best subset selection, forward / backward stepwise selection) to improve on your best model so far for the bike share competition.
Be prepared to discuss: Why is variable selection an extra step you might want to take for regression, but not for tree-based models? Do you think that best subset selection is a realistic method with this dataset? Did you try to add new "engineered" features before you performed subset selection?
Thursday: Try fitting a regression model using one or more of the following: ridge regression, the lasso, PC regression, partial least squares.
Be prepared to discuss: Which of these methods did you think, before you tried them, would be most helpful in the context of this challenge? Why (or did you not think there would be much difference)? Which method proved to help the most? Were there any challenges in applying these methods to this dataset?
Competition: Kaggle: Bike sharing demand: `https://www.kaggle.com/c/bike-sharing-demand`
- **Week 8:** Regression models: Non-linear regression models, regression trees. Measuring predictor importance.
Reading: James et al.: Ch. 7. Kuhn and Johnson: Chs. 7 and 8.
Tuesday: Try fitting a non-linear regression with splines in a GLM and / or a GAM.
Be prepared to discuss: Did non-linear terms for some predictors help improve your model? Which predictors do you think might have a non-linear relationship with the number of bicycles used?
Thursday: Try fitting a regression tree and an ensemble of trees (e.g., random forest, bagging, or boosting).
Be prepared to discuss: Do you think that the tree you fit was big enough, given that the many-interactions GLM does so well? What are the R defaults for when to stop growing a tree, and is there a way to change them? From the ensemble model, which variables were the most important? Was boosting better than a random forest or bagging? How long did it take to fit these ensemble models?
Competition: Kaggle: Bike sharing demand: `https://www.kaggle.com/c/bike-sharing-demand`
- **Week 9:** Regression models: Regression trees. Feature selection / engineering re-visisted.
Reading: James et al.: Ch. 8 (re-visited). Kuhn and Johnson: Chs. 3 and 19 (re-visited).
Tuesday: Try fitting additional regression trees and / or ensembles of trees (e.g., random forest, bagging, or boosting). Assess the importance of different predictors for the ensemble models.
Be prepared to discuss: Which variables seem most important? Are these consistently picked up by all the different types of tree methods (e.g., tree, random forest, bagging, ...)
Thursday: Go back and pick a few of the best models you've fit so far. See if you can engineer new features to add to the models to improve performance.
Be prepared to discuss: Which new features did you create and test? Did they help improve any of your top models? Why do you think the new features did or did not help?
Competition: Kaggle: Bike sharing demand: `https://www.kaggle.com/c/bike-sharing-demand`
- **Week 10:** Student team presentations on final model for current Kaggle competition.
Reading: Background information on this year's Data Mining Cup

Competition: Kaggle: Bike sharing demand: `https://www.kaggle.com/c/bike-sharing-demand`
Graded products: Students present final predictive model for current Kaggle competition.

- **Week 11:** Factors that can affect model performance. Case study: Grant application models.
  Reading: Kuhn and Johnson: Chs. 15, 20.
  Competition: This year's Data Mining Cup competition.
- **Week 12:** High-dimensional data (re-visited). Case study: Concrete Mixture Strength models.
  Reading: James et al. : Ch. 6 (re-visited). Kuhn and Johnson: Ch. 10.
  Competition: This year's Data Mining Cup competition.
- **Week 13:** Neural networks. Deep learning methods.
  Reading: Kuhn and Johnson: Ch. 13.2.
  Competition: This year's Data Mining Cup competition.
- **Week 14:** Visualization. Unsupervised learning.
  Reading: James et al.: Ch 10.
  Competition: This year's Data Mining Cup competition.
- **Week 15:** Student team presentations on final model for this year's DMC competition.
  Competition: This year's Data Mining Cup competition.
  Graded products: Students present final predictive model for DMC.