STAT/IST 557: Data Mining Spring 2014

Class: Tuesday, Thursday 1:00 PM - 2:15 PM, Willard 273

Instructor: Le Bao lebao@psu.edu
Office Hours: Tuesday 2:15pm-2:45pm Willard 273
Thursday 2:15pm-3:30pm Wartik Lab 514C
or by appointment

TA: WenYu Hua wxh182@psu.edu
Office Hours: Wednesday 4:00pm~5:00pm Thomas 331B

Textbooks:

- An Introduction to Statistical Learning with Applications in R By Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
  The supplementary materials (R-code etc) can be found at http://www-bcf.usc.edu/~gareth/ISL/book.html

Recommended Textbooks:

- The Elements of Statistical Learning By Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome
- Statistical Learning from a Regression Perspective By Richard A. Berk
- Modern Multivariate Statistical Techniques By Alan Julian Izenman
- Pattern Recognition and Machine Learning by C. M. Bishop
- Pattern Recognition and Neural Networks by B. Ripley

Course Objective:

This course covers methodology, major software tools and applications in data mining. By introducing principal ideas in statistical learning, the course will help students to understand conceptual underpinnings of methods in data mining. It focuses more on usage of existing software packages (mainly in R) than developing the algorithms by the students. Students will be required to work on projects to practice applying existing software. The topics include:

- Introduction
- Exploratory data analysis
- Linear Regression
• Linear Regularization: ridge regression, lasso regression
• Model selection: LRT, AIC, BIC, Cross-validation
• Dimension reduction
• Regression methods for classification: regression on indicators, Logistic regression and generalized linear models
• Nonparametric methods: K-nearest neighbors
• Discriminant analysis: Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Regularized discriminant analysis (RDA), and Reduced-rank LDA.
• Generalized additive models and Classification and regression trees (CART)
• Bagging and Random Forest
• Support Vector Machine (SVM) and Boosting
• Cluster analysis: k-means, hierarchical clustering, model-based clustering, etc.
• Ensemble methods and model averaging, etc.

Website:
Class announcements and materials will be regularly posted on ANGEL, so it is recommended that you check the site frequently.
The lecture notes are also available at https://onlinecourses.science.psu.edu/stat857/

Prerequisites:
STAT 501, 511 or similar course that cover analysis of research data through simple and multiple regression and correlation; polynomial models; indicator variables; step-wise, piece-wise, and logistic regression.
Basic programming skills in R. An Introduction to R is available at http://cran.r-project.org/doc/manuals/R-intro.html

Evaluation:
Weakly quizzes 25%
R Lab 25%
Team Projects 50%
Quiz: There will be 12 in class quizzes. No make-up quiz will be provided regardless of reason. Two lowest scores will be dropped automatically in the final grade as a remedy for accidentally absence.

R Lab: There will be 10 sets of online R lab questions. The lowest score will be dropped.

There will be 4 projects which are done by team work (3 ~4 students).
Grading will be score between 1 to 10.
10: Excellent, 9: Very good, 8: Good, 7: need improvement, <=6: need substantial improvement.

**Final Grading Scale:**
A: 93%; A-: 90%; B+: 87%; B: 83%; B-: 80%; C+: 77%; C: 70%; D: 60%; F: <60%

**Academic Integrity:**
All Penn State and Eberly College of Science policies regarding academic integrity apply to this course. See [http://www.science.psu.edu/academic/Integrity/index.html](http://www.science.psu.edu/academic/Integrity/index.html) for details.

**ECOS Code of Mutual Respect and Cooperation:**
The Eberly College of Science Code of Mutual Respect and Cooperation: [http://www.science.psu.edu/climate/Code-of-Mutual-Respect final.pdf](http://www.science.psu.edu/climate/Code-of-Mutual-Respect final.pdf) embodies the values that we hope our faculty, staff, and students possess and will endorse to make The Eberly College of Science a place where every individual feels respected and valued, as well as challenged and rewarded.

“Penn State welcomes students with disabilities into the University's educational programs. If you have a disability-related need for reasonable academic adjustments in this course, contact the Office for Disability Services (ODS) at 814-863-1807 (V/TTY). For further information regarding ODS, please visit the Office for Disability Services Web site at [http://equity.psu.edu/ods/](http://equity.psu.edu/ods/).

In order to receive consideration for course accommodations, you must contact ODS and provide documentation (see the documentation guidelines at [http://equity.psu.edu/ods/guidelines/documentation-guidelines](http://equity.psu.edu/ods/guidelines/documentation-guidelines)). If the documentation supports the need for academic adjustments, ODS will provide a letter identifying appropriate academic adjustments. Please share this letter and discuss the adjustments with your instructor as early in the course as possible. You must contact ODS and request academic adjustment letters at the beginning of each semester.”