Statistical Methods for Data Mining

Instructor: Kuangnan Fang

Email: xmufkn@xmu.edu.cn

Office hours: Tuesday 2:30-4:30

Office: Room B406, Economics building

Class time&place: Mon:19.10-20.50 N202

Wed: 10.10-11.50 N202

Download Website: http://kuangnanfang.com/?id=25

Course description: This course provides an accessible overview of the field of data mining and statistical learning, an essential toolset for making sense of the vast and complex data sets that have emerged in fields ranging from biology to finance to marketing to astrophysics in the past twenty years. This course presents some of the most important modeling and prediction techniques, along with relevant applications. Topics include linear regression, classification, resampling methods, shrinkage approaches, tree-based methods, support vector machines, clustering, and more. Color graphics and real-world examples are used to illustrate the methods presented. Since the goal of this course is to facilitate the use of these statistical learning techniques in science, industry, and other fields, it contains a tutorial on implementing the analyses and methods presented in R, an extremely popular open source statistical software platform.

Prerequisites: Probability and Mathematical Statistics, R programming skill

Assessment:
Homework: 20%
Attendance: 10%
Mid-term exam: 30%
Project and presentation: 40%
The assignments are to be done by each student individually. Projects may be done individually or in groups of two (possibly more than two, with special permission). More will be expected of a group project than an individual project.

Course Text:
1. Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

2. James G, Witten D, Hastie T, et al. An introduction to statistical learning. New York: springer, 2013.

Contents: