

统计学习和数据挖掘

乔星烨

Binghamton University, State University of New York

先修科目

1. R 编程语言
2. 线性回归
3. 矩阵代数

课程描述

这是统计学习和数据挖掘的综述课程。课程将涵盖主要的有监督和无监督统计学习方法和概念，包括罚回归和罚分类(penalized regression and classification)，分类分析(LDA, QDA)，支持向量机(support vector machine)，核函数方法(kernel methods)，非线性方法，模型选择，矩阵分解，主成分分析，图模型(graphical model)，聚类分析，回归分类树，随机森林，提升方法(boosting)，和集成学习(ensemble learning)。课程的资料都是英文形式存在；授课以中文为主，辅以英语专有名词。

学习目标

学生们将会学到如何以及何时使用正确的统计学习方法，它们之间相对的优点和缺点，批判性地衡量学习算法的优劣。完成课程以后，学生们将会

- 处理和显现不同的数据类型
- 使用基本的学习方法来构建预测模型或者进行探索分析
- 理解预测模型的内在机理，并且评价和解释此类模型
- 合理地调整选择和审核统计学习方法
- 使用广泛使用的分析工具和软件
- 独立地或者在团队中解决问题

推荐教材（大部分教材都可以在网上合法免费下载）

- Hastie, Trevor, Tibshirani, Robert, and Friedman, J. H. 2009. The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer New York.
- James, Witten, Hastie and Tibshirani, 2014. An Introduction to Statistical Learning with Applications in R. [Book Home Page](#). The [PDF](#) file of the book can be downloaded for free. There is also a [R library](#) for this book.
- Hastie, Trevor, Tibshirani, Robert, and Wainwright, Martin. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC; 1 edition (May 7, 2015). The [PDF](#) file of the book can be downloaded for free.
- Bühlmann, Peter, and van de Geer, Sara. Statistics for High-Dimensional Data. Springer-Verlag Berlin Heidelberg.

- Boyd, Stephen, and Vandenberghe, Lieven. Convex Optimization. Cambridge University Press. The [PDF](#) file of the book can be downloaded for free.