光华管理学院 Guanghua School of Management | UG&PG 本科研究生项目

课程大纲

数据挖掘与应用

Data mining and application

课程编号：02817130
学　　分：3
课程类型：必修
先修课程：概率论、数理统计

授课对象：研究生
任课教师：张俊妮
开课学期：2017 年秋

任课教师简历（500 字左右）:
张俊妮博士现任北京大学光华管理学院统计学副教授。她 1998 年毕业于中国科学技术大学，获计算机软件学士学位；2002 年毕业于美国哈佛大学，获统计学博士学位。

她的研究领域为因果推断、贝叶斯分析、蒙特卡洛方法、数据挖掘以及统计在经济、金融、营销中的应用。曾在国际主要学术期刊 Journal of the American Statistical Association, Journal of Educational and Behavioral Statistics, Statistica Sinica, Computational Statistics and Data Analysis, Journal of Chemical Physics, 管理世界, 经济学季刊等国内外刊物上发表文章。并有中文专著《数据挖掘与应用》。

她曾参与国家自然科学基金项目“品牌个性维度及其测量量表研究”，负责过北京大学光华管理学院与中国信达资产管理公司合作的“金融不良资产定价”课题研究项目和北京天健兴业资产评估公司“统计估值模型”的项目研究，主持国家自然科学基金项目“使用倾向分和主分层进行因果推断”，并且担任过担任美国国立卫生学院（NIH）国际研究合作基金项目“生活质量研究中的因果推断”的中方负责人。在 2010 年“计量方法在经济中的应用”国际大会上，她是组委会成员之一。她于 2004-2009 年担任 Computation Statistics 编委（Associate Editor）。她至今仍担任北京哈佛校友会理事，是美国统计学会和全球华人统计学会成员。

任课教师联系方式：
光华管理学院 2 号楼 473 办公室，电话：62757922，
邮箱：zjn@gsm.pku.edu.cn

助教姓名及联系方式：

辅导、答疑时间：

一、项目培养目标

1 **Learning Goal 1** Graduates will be thoroughly familiar with the specialized knowledge and theories required for the completion of academic research.

1.1 Objective 1 Graduates will have a deep understanding of basic knowledge and theories in their specialized area.

1.2 Objective 2 Graduates will be familiar with the latest academic findings in their specialized area and will be knowledgeable about related areas.

1.3 Objective 3 Graduates will be familiar with research methodologies in their specialized

area, and will be able to apply them effectively.

2 **Learning Goal 2** Graduates will be creative scholars, who are able to write and publish high-quality graduation dissertation and research papers.

2.1 Objective 1 Graduates will write and publish high-quality graduation dissertation and research papers

2.2 Objective 2 Graduates will be critical thinkers and innovative problems solvers.

3 **Learning Goal 3** Graduates will have a broad vision of globalization and will be able to communicate and cooperate with international scholars

3.1 Objective 1 Graduates will have excellent oral and written communication skills

3.2 Objective 2 Graduates will be able to conduct efficient academic communication in at least one foreign language

4 **Learning Goal 4** Graduates will be aware of academic ethics and will have a sense of social responsibility.

4.1 Objective 1 Graduates will have a sense of social responsibility.

4.2 Objective 2 Graduates will be aware of potential ethical issues in their academic career.

4.3 Objective 3 Graduates will demonstrate concern for social issues.

二、课程概述

本课程主要从统计学的角度探讨与数据挖掘相关的理论与应用。课程中将使用 SAS 等统计软件。

三、课程目标

在学习完本课程之后，学生应理解数据挖掘的大框架以及数据挖掘中各种常用的方法，并可熟练使用统计软件进行数据挖掘。

四、内容提要及学时分配

课号	主题
1	数据挖掘案例及数据挖掘框架
2	数据理解与数据准备（1）
3	数据理解与数据准备（2）
4	关联规则挖掘
5	多元统计中的降维方法
6	聚类分析
7	判别分析；朴素贝叶斯分类算法；k 近邻法；线性模型与广义线性模型（1）
8	线性模型与广义线性模型（2）
9	神经网络（1）
10	神经网络（2）、决策树（1）
11	决策树（2）

12	模型比较与评估
13	模型组合与两阶段模型
14	期末项目口头报告

期末考试时间：

五、 教学方式

课程讲授为主。

六、 教学过程中 IT 工具等技术手段的应用

课程提供 pdf 讲义，上课时用计算机演示讲义，并讲授及演示统计软件的使用。

学生做作业和期末项目时需要使用统计软件分析数据。

七、 教材

张俊妮（2009），《数据挖掘与应用》，北京大学出版社。

八、 参考书目

（1）实际应用

1. Berry, M. J. A. and Linoff, G. (1997). Data mining techniques for marketing, sales and customer relationship management. Wiley Publishing Inc.
2. Berry, M. J. A. and Linoff, G. (2000). Mastering data mining: the art and science of customer relationship management. John Wiley & Sons.
3. Kudyba, S. (Eds) (2004). Managing Data Mining: Advices from Experts. Cybertech Publishing.

（2）数据挖掘综述

4. Dunham, M. H. (2003). Data mining introductory and advanced topics. Pearson Education.
5. Han, J. and Kamber, M. (2001). Data mining: concepts and techniques. Morgan Kaufmann Publishers.
6. Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of statistical learning. Springer-Verlag.
7. Weiss, S. M. and Indurkhya, Nitin (1998). Predictive data mining: a practical guide. Morgan Kaufmann Publishers.

（3）数据挖掘的某些方面或某些数据挖掘方法

8. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and regression trees. Chapman & Hall, New York.
9. Kohonen, T. (1995). Self-organizing maps, 3rd edition. Springer-Verlag.
10. McCullagh, P. and Nelder, J. A. (1989), Generalized linear models, second edition. Chapman & Hall/CRC.
11. Pyle, D. (1999). Data preparation for data mining. Morgan Kaufmann Publishers.
12. Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge University Press.
13. Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc.
14. Samarasinghe, S. (2006). Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition. Auerbach.
15. Zhang, C. and Zhang, S. (1998). Association rule mining. Springer-Verlag.

九、 教学辅助材料，如 CD、录影等

十、 课程学习要求及课堂纪律规范

学生分为各小组，每个组由 3~4 人组成，在整个学期中互相合作与支持，并共同完成课程要求的作业和期末项目。

每次作业按照规定时间上交。迟于规定时间 24 小时内上交的作业，扣除该次作业总分的 50%。迟于规定时间 24 小时以外上交的作业，以零分计。

十一、 学生成绩评定办法（需详细说明评估学生学习效果的方法）

先如下计算小组成绩：

作业： 50%

期末项目： 50%

期末时每位同学需要对所在小组内各位成员的贡献进行评分，每位同学的最终成绩在小组成绩的基础上根据贡献分进行调整。