数据挖掘与应用

课程编号：02817130 **授课对象： 研究生**

学　　分：3 **任课教师：张俊妮**

课程性质：专业必修 **开课学期：2013 年秋**

先修课程：概率论、数理统计

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

一、 项目培养目标

学习目标 1 系统掌握从事学术研究所需要的专业知识及理论。 具体目标 1、系统掌握本学科基础知识及基本理论 具体目标 2、掌握本学科前沿知识和理论、具有足够的相关领域的知识 具体目标 3、熟练掌握本学科的研究方法
学习目标 2 具有从事创新性研究的能力；能够撰写并发表高质量的毕业论文和学术论文 具体目标 1、撰写高质量的毕业论文和学术论文 具体目标 2、具有高水平的分析能力和批判思维能力，能够创造性地解决问题
学习目标 3 具有宽阔的国际视野，能够与国际学者进行交流、合作的能力。 具体目标 1、具有优秀的口头交流和文字交流能力 具体目标 2、能够熟练地运用至少一门外语进行学术交流与沟通
学习目标 4 了解学术伦理，具有强烈的社会责任感、关注社会问题 具体目标 1、了解社会责任感的重要性 具体目标 2、了解学术生涯中的学术道德问题 具体目标 3、关注现实社会问题

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（包括学生所提高的技能要求）

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

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

无。

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。