课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	金融数据分析和数据挖掘 Financial data analysis and Data Mining
2.	**授课院系 Originating Department**	金融系 Department of Finance
3.	**课程编号 Course Code**	FIN208
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业核心课 Major Core Courses
6.	**授课学期 Semester**	春季 Spring
7.	**授课语言 Teaching Language**	中英双语 English & Chinese
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	陈琨，助理教授，金融系 Kun CHEN, Assistant Professor. Chen Kun, Department of Finance 邮箱/Email:chenk@sustech.edu.cn 办公室/office:慧园 3 栋 319, Wisdom Valley 3#319
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	滕琪，金融系 Teng Qi, Department of Finance 邮箱/Email:tengq@mail.sustech.edu.cn
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11.		讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
	授课方式 Delivery Method					
	学时数 Credit Hours	32	N/A	32	N/A	64

12. 先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	概率论与数理统计 Probability and Statistics MA212
13. 后续课程、其它学习规划 Courses for which this course is a pre-requisite	无 None
14. 其它要求修读本课程的学系 Cross-listing Dept.	无 None

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

此课程的目的是讲授数据分析以及数据挖掘的基本过程、模型和工具，及其在金融中的应用。此课程将培养学生软件包（如 Excel 和 weka 软件）的实用技巧和一些必要扩展的应用来分析和解决金融数据问题。

The course aims to teach students the process, models, and tools for data analytics and data mining in finance. The course will teach students the practical skills to employ software packages (such as Excel and weka) and apply necessary extensions to analytic framework and tackle financial data analysis problems.

16. 预达学习成果 Learning Outcomes

1、能够描述在金融领域数据分析与挖掘的主要任务和内容。
2、通过此课程的学习，学生能够完成金融以及其他领域数据的分析与挖掘，形成系统的数据分析知识，应用于实践操作中。
3、创造性地应用所讲述的建模技术，并灵活解决所发现的实际数据分析与挖掘问题。
4、以口头、书面或电子表格的形式灵活有效的表述分析过程及其结果。

1. Describe the target and requirements for a spectrum of business data analysis and data mining problems in finance, marketing, etc.
2. Develop the ability to employ data mining algorithms to discover patterns in data to address the selected problems.
3. Creatively apply and adapt the introduced modeling techniques to propose original findings for practical organizational data analysis problems.
4. Creatively communicate analytic procedure and results effectively in presentations with oral, written and electronic formats.

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

理论（32 学时）
第一章 数据 （4 学时）
1.1 数据类型、数据质量、数据预处理（2 学时）
本部分主要讲解如何从属性与度量来描述数据对象，数据质量问题的检测和纠正，以及对数据进行预处理的一些思想和方法。
1.2 相似性和相异性（2 学时）
本部分主要讲解简单属性之间的相似度和相异度，数据对象之间的相似度和相异度，以及邻近度计算问题。
第二章 描述分析 （2 学时）
本章主要介绍汇总数据的方法，一是描述统计查看数据的基本分布，二是以图形或表格的形式将数据转换成可视化的形式。
2.1 描述统计
2.2 图表和技术
第三章 回归和决策 （4 学时）
3.1 线性回归（2 学时）
本部分主要介绍线性回归的基本概念，并重点介绍一元线性回归和非线性回归的基本方法以及回归方程的方差分析和显著性检验。
3.4 决策及优化（2 学时）
本章主要讲解最优化问题的定义、分类以及求解优化问题的各种技术
第四章 分类：基本概念、决策树与模型评估 （5 学时）

4.1 预备知识（1 学时）
本部分主要阐述分类的基本概念，介绍解决分类问题的一般方法，
4.2 决策树归纳（2 学时）
本部分主要讲解决策树分类法的工作原理，以及建立决策树的方法。
4.3 模型的过分拟合（1 学时）
本部分主要讲解模型过分拟合问题，以及如何处理决策树归纳中的过分拟合。
4.4 评估分类器的性能（1 学时）
本部分主要介绍了保持方法、随机二次抽样、交叉验证等一些常用的评估分类器性能的方法。
第五章 分类：其他技术（5 学时）
5.1 基于规则的分类器（1 学时）
本部分主要讲解基于规则的分类器的工作原理和特征，以及规则提取的方法。
5.2 最近邻分类器（1 学时）
本部分主要讲解最近邻分类的算法原理，以及最近邻分类器的特征。
5.3 贝叶斯分类器（1 学时）
本部分主要讲解贝叶斯定理以及贝叶斯定理在分类中的应用。
5.4 人工神经网络（1 学时）
本部分主要讲解人工神经网络的结构及特点。
5.5 支持向量机（1 学时）
本部分主要讲解支持向量机的基本思想，以及在线性可分数据上、非线性可分数据上训练 SVM。
第六章 关联分析：基本概念和算法（5 学时）
6.1 问题定义（1 学时）
本部分主要讲解关联规则挖掘问题的形式描述，以及关联规则挖掘算法通常采用的策略。
6.2 频繁项集与规则的产生（2 学时）
本部分主要讲解关联规则挖掘算法中频繁项集和规则产生的有效技术，重点介绍 Apriori 算法的频繁项集和规则的产生。
6.4 FP 增长算法（1 学时）
本部分主要介绍了产生频繁项集的其他算法—FP 增长算法。
6.5 关联模式的评估（1 学时）
本部分主要讲解评估关联模式质量的方法。
第七章 聚类分析：基本概念和算法（4 学时）
7.1 K 均值、凝聚层次聚类、DBSCAN 算法（3 学时）
本节主要介绍了 K 均值、凝聚的层次聚类算法、DBSCAN 算法的基本原理，以及各算法的优缺点。
7.4 簇评估（1 学时）
本部分主要介绍评估聚类算法产生的簇的方法。
第八章 应用 （3 学时）
8.1 财务报表分析：比率和预测（1 学时）
本部分主要介绍数据挖掘技术在财务报表分析中的应用。
8.2 财务预测：销售、收入和股票（1 学时）
本部分主要介绍数据挖掘技术在财务预测中的应用。
8.3 市场营销中的商业智能：普查、分割和购物篮分析（1 学时）
本部分主要介绍数据挖掘技术在市场营销决策中的应用。
实践（32 学时）
第一章 描述分析（4 学时）
1.1 描述统计（2 学时）
本部分主要是辅导学生使用 Excel 中的“数据分析”功能对数据进行描述统计。
1.2 数据透视表（2 学时）
本部分主要是辅导学生使用 Excel 中的“数据透视表”功能对数据进行可视化分析。

第二章 Python 初始（2 学时）
本章主要讲解 Python 的安装以及 Python 的基础语法
2.1 Python 的安装
2.2 Python 的基础语法

第三章 网络爬虫（6 学时）
8.1 Requests 库（2 学时）
本部分主要介绍 HTML 标签，以及 Python 中 Requests 库的使用
8.2 BeautifulSoup 库——图片抓取（2 学时）
本部分主要介绍 Python 中 BeautifulSoup 库，并使用其来抓取网络上的图片。
8.3 BeautifulSoup 库——文本抓取（2 学时）
本部分主要是辅导学生编写爬虫程序从网络上抓取文本信息保存到本地。

第四章 文本挖掘（4 学时）
4.1 jieba 分词库（2 学时）
本部分主要讲解 Python 中 jieba 分词库对文本进行处理和关键词的提取。
4.2 TF-IDF 算法（2 学时）
本部分主要讲解 TF-IDF 算法原理，以及利用 TF-IDF 算法提取文本关键词，求文本相似度。

第五章 Weka 初始（2 学时）
本章主要讲解 weka 软件中的数据格式以及它的基本功能。

第六章 分类（4 学时）
6.1 Weka+分类（2 学时）
本部分主要讲解运用 weka 软件中的分类算法对数据集进行分析，应用不同的分类算法，绘制多条 ROC 曲线，比较他们之间的不同
6.2 Weka API—J48 分类（2 学时）
本部分主要讲解如何使用 JAVA 调用 weka 包实现 J48 分类算法。

第七章 关联规则（2 学时）
本章主要讲解运用 Weka 软件中的关联规则对数据集进行分析。

第八章 聚类（4 学时）
8.1 Weka+聚类（2 学时）
本章主要讲解运用 Weka 软件中不同的聚类算法对数据集进行分析，比较他们之间的不同。
8.2 Weka API—EM 聚类（2 学时）
本部分主要讲解如何使用 JAVA 调用 weka 包实现 EM 聚类算法。

第九章 PageRank（2 学时）
本章主要介绍 PageRank 算法的来源和原理，以及算法的实现。

第十章 Final Project（2 学时）
本章主要是辅导学生完成最终项目报告。

Lecture (32 hours)
Chapter 1： Data （4 hours）
1.1 Data type, data quality, data preprocessing (2 hours)
This section explains how to describe data objects from attributes and metrics, explains the detection and correction of data quality problems, and the ideas and methods for preprocessing data.
1.2 Similarity and dissimilarity (2 hours)
This section mainly explains the similarity and dissimilarity between simple attributes, the similarity and dissimilarity between data objects, and the problem of proximity calculation.
Chapter 2： Descriptive Analysis (2 hours)
This chapter mainly introduces the methods of summarizing data. One is to describe the basic distribution of statistics and the other is to convert the data into a visual form in the form of graphs or tables.
2.1 Descriptive Statistics
2.2 Charts
Chapter 3： Regression and Decision (4 hours)

This chapter mainly introduces the theoretical model of linear regression and various techniques for solving optimization problems.

3.1 Liner Regression (2 hours)

This section mainly introduces the basic concepts of linear regression, and focuses on the basic methods of linear regression and nonlinear regression, as well as the analysis of variance and significance of regression equations.

3.2 Decision and Optimization (2 hours)

This chapter mainly explains the definition, classification and various techniques for solving optimization problems.

Chapter 4： Classification: Basic Concepts, Decision Trees, and Model Evaluation (5 hours)

4.1 Preliminaries (1 hours)

This section mainly describes the basic concepts of classification and introduces the general approach to solving classification problems.

4.3 Decision Tree Induction (2 hours)

This part mainly explains the working principle of the decision tree classification method and the method of establishing the decision tree.

4.4 Model Over fitting (1 hours)

This section focuses on the model over-fitting problem and how to deal with over-fitting in decision tree induction.

4.5 Evaluating the Performance of a Classifier (1 hours)

This section mainly introduces some commonly methods for evaluating the performance of a classifier, such as retention methods, random subsampling, and cross-validation.

Chapter 5： Classification: Alternative Techniques (5 hours)

5.1 Rule-Based Classifier (1 hours)

This section mainly explains the working principle and characteristics of the rule-based classifier and introduces the method of rule extraction.

5.2 Nearest-Neighbour classifiers (1 hours)

This section mainly explains the algorithm of the nearest neighbour classification and the characteristics of the nearest neighbour classifier.

5.3 Bayesian Classifiers (1 hours)

This section mainly explains Bayesian and the application of Bayesian in classification.

5.4 Artificial Neural Network (ANN) (1 hours)

This section mainly explains the structure and characteristics of artificial neural networks.

5.5 Support Vector Machine (SVM) (1 hours)

This section mainly explains the basic idea of support vector machine, and explains how to train SVM on linear separable data and nonlinear separable data.

Chapter 6： Association Analysis: Basic Concepts and Algorithms (5 hours)

6.1 Problem Definition (1 hours)

This section mainly explains the formal description of the mining rules of association rules and the strategies commonly used in association rules mining algorithms.

6.2 The Generation of Frequent Itemset and Rule (2 hours)

This section mainly explains the effective techniques for generating frequent itemsets and rules in the association rule mining algorithm, focusing on the frequent itemsets and rules generation of the Apriori algorithm.

6.4 FP-Growth Algorithm (1 hours)

This section mainly introduces other algorithms that generate frequent itemsets—the FP growth algorithm.

6.5 Evaluation of Association Patterns (1 hours)

This section focuses on methods for assessing the quality of associated models.

Chapter 7： Cluster Analysis: Basic Concepts and Algorithms (4 hours)

7.1 K-means, Agglomerative Hierarchical Clustering, DBSCAN (3 hours)

This section mainly introduces the K-means, condensed hierarchical clustering algorithm, the basic principles of the DBSCAN algorithm, and the advantages and disadvantages of each algorithm.

7.2 Cluster Evaluation (1 hours)

This section mainly introduces methods for evaluating clusters generated by clustering algorithms.

Chapter 8: Applications (3 hours)

This chapter mainly introduces the application of data mining algorithms in practical cases.

8.1 Financial statement analysis: Ratios and predictions (1 hours)
This section mainly introduces the application of data mining technology in financial statement analysis.
8.2 Financial forecasting: Sales, revenue, and stock (1 hours)
This section mainly introduces the application of data mining technology in financial forecasting.
8.3 Business intelligence in marketing: Census, segmentation & basket analysis (1 hours)
This section mainly introduces the application of data mining technology in marketing decision-making.

LAB (32 hours)
Chapter 1 Description Analysis (4 hours)
1.1 Data analysis (2 hours)
This section mainly shows how to use the “data analysis” function in Excel to analyze data.
1.2 Pivot Table (2 hours)
This section mainly shows how to use the “Pivot Table” to visually analyze the data.

Chapter 2 Python (2 hours)
This chapter focuses on Python installation and the basic syntax of Python.
2.1 Python installation
2.2 Basic syntax of Python

Chapter 3 Web Crawler (6 hours)
3.1 Requests library (2 hours)
This section explains the HTML tags and the use of Requests library in Python.
3.2 BeautifulSoup Library - Picture Grab (2 hours)
This section focuses on the BeautifulSoup library in Python and uses of it to grab images from internet.
3.3 BeautifulSoup Library - Text Grab (2 hours)
This section is mainly to tutors students program web crawler to grab text information from the internet.

Chapter 4 Text Mining (4 hours)
4.1 jieba library (2 hours)
This section mainly explains the processing of jieba library.
4.2 TF-IDF algorithm (2 hours)
This section mainly explains how to use the TF-IDF algorithm to extract keywords and seek text similarity.

Chapter 5 Weka (2 hours)
This chapter mainly explains the data format and basic functions in Weka software.

Chapter 6 Classification (4 hours)
6.1 Weka— classification (2 hours)
This section mainly explains the use of different classification algorithms in Weka software to analyse the dataset, as well as in this chapter, students will learn analysing different classification algorithms,
6.2 Weka API—J48 Classification (2 hours)
This section introduces how to call the weka package to achieve J48 classification by java.

Chapter 7 Association Rules (2 hours)
This chapter mainly explains the analysis of dataset using the association rules in Weka software.

Chapter 8 Clustering (4 hours)
8.1 Weka+ clustering (2 hours)
This section mainly explains the use of different clustering algorithms in Weka software to analyse the dataset, as well as in this chapter, students will learn analysing different clustering algorithms.
8.2 Weka API—EM Clustering (2 hours)
This section introduces how to call the weka package to achieve EM clustering by java.

Chapter 9 PageRank (2 hours)
This chapter mainly introduces the source and principle of the PageRank algorithm and the implementation of the algorithm.

Chapter 10 Final Project (2 hours)
This chapter is mainly for students to complete the final project report.

18. **教材及其它参考资料 Textbook and Supplementary Readings**

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, posts & telecom press.

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance				
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects		30%		
平时作业 Assignments		20%		
期中考试 Mid-Term Test				
期末考试 Final Exam	2 小时 2 hours	50%		
期末报告 Final Presentation				
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. **记分方式 GRADING SYSTEM**

■ A. 十三级等级制 **Letter Grading**
☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. **本课程设置已经过以下责任人/委员会审议通过**
This Course has been approved by the following person or committee of authority

金融系课程规划与审核委员会
Curriculum Planning and Review Committee, Dept. Finance