**Syllabus of "Data Mining"**

(The second semester of 2020-2021)

**Course** : Data Mining

**Instructor:** Professor Huang Mian

Q&A time: Appointment or 3:00 pm- 4 :00 pm on Mondays

10 , School of Statistics and Management , 65901057

E-mail: huamg.mian@mail.shufe.edu.cn

**Course category:** Statistics major elective course

**Course Target:** Senior Undergraduates in Statistics

**Course schedule description:** March 1, 2021-June 15, 2016

Week 1 p.m. 6:00-7:40 Classroom 3107

Offline teaching, subject to school arrangements

Use online teaching during the epidemic

Curriculum adjustment: The content of the holiday course will be postponed.

Final exam time: Estimated at the end of June 2021, subject to the time of the Academic Affairs Office.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| credit | total period | Whether it is a quality course | Theory teaching hours | Practical teaching hours | Experimental teaching hours |
| 2 | 3 2 | no | 16 | 16 | 0 |

**Courseware URL:** [bb.shufe.edu.cn](http://iclass.shufe.edu.cn/teacherweb/users/wumaths)

**Textbooks and bibliography:**

**Textbook** :

Machine Learning and Python Practice by Huang Mian

An Introduction to Statistical Learning, Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani

**Reference :**

1. The Elements of Statistical Learning, T. Hastie, Tibshirani, Friedman ;Modern

3. Machine learning. By Zhou Zhihua

4. Pattern Recognition and Machine Learning, Bishop

**Advanced Placement**

Calculus, Advanced Algebra, Probability Theory and Mathematical Statistics

**Pre-knowledge**

This course is a common course for majors in statistics. Students have a foundation in mathematical statistics, linear algebra, and computer programming , which will help the study and understanding of this course.

**Curriculum knowledge requirements:**

This course is a professional elective course in statistics. The basic knowledge requirements include advanced algebra, probability theory, mathematical statistics, and basic programming skills such as computer languages such as C or C++ , understanding of procedural programming, basic data types, custom functions, etc. The knowledge requirements of this course include data sorting and summary, simple descriptive statistical analysis, and mastering of commonly used unsupervised learning and supervised learning methods, such as kernel density estimation, K -means method, principal component analysis, regression analysis, logistic regression, linear Discriminant analysis, local modeling methods, model selection and inference, etc. Through this course, students can program to realize various data mining methods independently, and can apply these methods to actual data analysis.

**Course setting ability requirements:**

This course will learn how to use scripting language (P ython ), combined with classroom programming, and combined with real-time programming for teaching. In the classroom, first teach the theory and basic derivation process of data discovery method, and then implement procedural programming to realize the data mining method and analyze actual data. Deepen students' understanding of data mining methods by concretizing abstract concepts. Explain various methods of data mining by combining multiple examples, including K -means method, principal component analysis, regression analysis, logistic regression, linear discriminant analysis, partial modeling method, integrating actual cases in the economic and financial fields for teaching, and improving student learning Interest and practical ability and data analysis ability, combined with after-school homework training, consolidate students' actual analysis ability and programming ability.

**Course achievement goals:**

The data mining course emphasizes on viewing and solving practical problems from a statistical perspective, understanding and mastering common statistical methods for mining useful information from data . Learn to use computer programming languages , be able to program independently , and apply the learned methods to actual data analyze. Through this course study, master the commonly used unsupervised learning and supervised learning methods including data sorting and summary, simple descriptive statistical analysis, and master the commonly used unsupervised learning and supervised learning methods, such as kernel density estimation, K -means method, Principal component analysis, regression analysis, logistic regression, linear discriminant analysis, local modeling methods, model selection and inference, etc. Through the combination of classroom programming and actual case analysis, students' practical programming ability and data processing ability can be effectively improved. In addition, through the analysis of several ideological and political cases , students have a certain degree of understanding of the development of China’s big data industry and economic and financial industries, as well as the great achievements of China’s big data industry and related economic and financial industries, as well as the tremendous achievements of Chinese scientists in this industry. contribute.

**Assessment form**

Homework 30 %

Time Attendance/ Midterm Quiz 1 0%

Final Exam 60 %

**Test paper structure**

[E ssay question](http://dict.cn/essay%20question) 3 0 %

Analytic problem 50%

Programming 20%

**Academic honesty**

The academic dishonesty issues involving students mainly include cheating on exams; plagiarism; forging or improper use of school records; obtaining and using exam materials without the teacher’s permission. The minimum penalty for academic dishonesty is 0 points on the exam. Other punishments include reporting to the relevant departments of the school and handling them in accordance with relevant regulations.

**Essentials of Course Teaching**

**1. Introduction to Python**

Introduction to Python Programming

Basic data structure

Python scientific computing

data visualization

**2. Unsupervised learning**

Principal component analysis

Kmeans algorithm

Kernel density estimation

**3. Regression and classification**

regression analysis

LDA and Logistic regression

Regression variable selection and compression

**4. Partial modeling**

Spline method

Nuclear regression

**5. Model evaluation, selection**

Bias, variance and forecast error

Model selection, information standards and cross-validation

Bootstrap and Resample methods

**6. Discussion topics**

Neural network and deep learning

Recent developments in data science