
COURSE SYLLABUS

FOR FULL-TIME UNDERGRADUATE PROGRAMS

(Issued under Decision No.1380/QĐ-ĐHKTQĐ on 15/8/2016 by the University President)

1. COURSE NAME: Principles of Data Mining

Code: TKKD1111

Total credits: 2

2. DEPARTMENT IN CHARGE OF INSTRUCTION: Business Statistics

Department

Office: Room No.401 – Block 7 – National Economics University

Office Hours: 8:00 – 17:00, from Monday to Friday

Office Telephone: 04.38693275

3. PREREQUISITES: Applied Informatics for Statistics.

4. COURSE DESCRIPTION:

Today, the collection and storage of data is easier than ever. The giant and diverse data warehouses are increasingly popular in all areas of economic, business administration, medicine, social studies,... Analysis of a large amount of data is very necessary, but the problem is how to exploit and analyze these data warehouses. This module will give a general understanding of the concepts and data mining, with particular emphasis on data analysis. Specifically, the concepts and technical basis of the classification, forecasting, combination, and grouping are included. The content is presented with examples, the approach to the principles of the basic methods such as data preprocessing, descriptive statistics, cluster analysis, peripheral element analysis ...

The basic principles of data mining is the subject of selected blocks depth knowledge of statistics, and also the subjects needed for other disciplines in the fields of business administration, commerce and finance, insurance

5. COURSE OBJECTIVES:

After completing the course, learners need to achieve the following knowledge and skills:

- Understanding the basic concepts of data mining, tasks of data mining and all kinds of application cases.

- Understanding and being skilled at planning, market researching from defining the problem to the design of research content, selection methods of collecting information and conducting planning.

- Ability to build a survey plan and design tools to collect information consistent with the object of study and practical conditions.

- Having proficient skills at basic methods such as interviews, observation and document analysis.

- Applying the statistic methods proficiently in presenting and analyzing the market through the information collected.

- Training the ability to write the research reports and presentation skills.

6. COURSE CONTENT:

TENTATIVE SCHEDULE

<i>No</i>	<i>Content</i>	<i>Total hours</i>	<i>In which</i>	
			<i>Theory</i>	<i>Practical exercise</i>
1	Chapter 1	4	3	1
2	Chapter 2	4	3	1
3	Chapter 3	4	3	1
4	Chapter 4	6	4	2
5	Chapter 5	6	4	2
6	Chapter 6	5	3	2
	Mid-term test	1		1
	Total	30	20	10

CHAPTER 1 – INTRODUCTION TO DATA MINING

This chapter provides an overview of data mining on many different aspects: the need to mention, the definition of data mining specifically on the process of information discovery. Next is the issue of data mining from several aspects, such as the type of data that can be exploited, the types of information that can be exploited, the techniques used and the application target. Finally, this chapter gives us the issue of research and development of main data mining.

Content:

1.1. The essential of data mining

1.2. What is data mining

1.2.1. The concept of data mining

1.2.2. What kinds of data can be mined?

1.2.3. What kinds of patterns can be mined?

1.2.4. What technologies are used?

1.2.5. What kinds of applications are targeted?

1.3. Major issues in data mining

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

CHAPTER 2 – GETTING TO KNOW DATA

We often want to exploit the data immediately, but first we need to prepare the data available. This involves considering the specific properties and data values. There typically exist in actual data the interference elements, accounting for a large capacity (usually several gigabytes or more) and these data come from many different sources and heterogeneous. This chapter refers to data describing issues. The knowledge of your data will take effect for the period of data preprocessing (this period being studied in Chapter 3) - the first phase of data mining process including 3 phases. We would like to know about the following issues: What types of properties and fields create data? What kind of value each property has? Which attributes are discrete and which ones are continuous? What do data sources look like? How are the values distributed? Is that the way we can visualize the data and have a better judgment on them? Can we realize any outliers; can we measure the similarity of the data object that is not identified with specific subjects? A deep understanding of data will suggest the later analysis. All these issues will be resolved in the following content.

Content:

2.1. Data objects and attribute types

2.2. What is an attribute?

2.1.2. Nominal attributes

2.1.3. Binary attributes

2.1.4. Ordinal attributes

2.1.5. Numeric attributes

2.1.6. Discrete and continuous attributes

2.3. Basic statistical descriptions of data

2.3.1. Measuring the central tendency: mean, median, and mode

2.3.2. Measuring the dispersion of data: range, quartiles, variance, standard deviation, and interquartile range

2.3.3 Graphic displays of basic statistical descriptions of data

2.4. Data visualization

2.5. Measuring data similarity and dissimilarity

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

CHAPTER 3 – DATA PREPROCESSING

In Chapter 2, we have studied the various types of criteria and how to use the basic descriptive statistics to study the characteristics of data. That can help determine the volume of errors or boundary element value, which will have the effect of cleansing and data synthesis. Technical data preprocessing is applied prior to mining, which basically can improve the whole quality of the mining model and / or the time required for actual mining.

In this chapter, will introduce the basic concepts of data preprocessing in the position; The methods for data pre-processing are held in the following categories: data cleaning, data aggregation, data reduction and data transformation.

Content:

3.1. Data preprocessing: an overview

3.1.1. Data quality: why preprocess the data?

3.1.2. The main task of data preprocessing

3.2. Data cleaning

3.2.1. Missing values

3.2.2. Noisy data

3.2.3. Data cleaning as a process

3.3. Data integration

3.3.1. Entity identification problem

3.3.2. Redundancy and correlation analysis

3.3.3. Other issues

3.4. Data reduction

3.4.1. Overview of data reduction strategies

3.4.2. Data reduction methods

3.5. Data transformation and data discretization

3.5.1 Data transformation strategies overview

3.5.2 Data transformation and discretization methods

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

CHAPTER 4 – CLASSIFICATION AND CLUSTER ANALYSIS

Classification and cluster analysis are the techniques applied in many statistical studies to explore large data sets, helping to split the individuals (the observers) or the analysis variables into groups based on similarities on certain characteristics. With the aim of dividing the variables that have similar characteristics to a group for the analysis. This chapter focuses on classification and clustering technique to split the individuals (observers), including the concepts of analytical methods and performance of analysis steps.

Content:

4.1. Discriminant analysis

4.1.1. General concept

4.1.2. The order of implementation

4.2. Cluster Analysis

4.2.1. General concept

4.2.2. The order of implementation

4.2.2.1. Select variables to cluster

4.2.2.2. Select distance measurement parameters

4.2.2.3. Select clustering methods

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

3. Hoang Trong, Chu Nguyen Mong Ngoc (2008) - *Analyzing research data with SPSS*, University of Economics Ho Chi Minh city Publishing House.

4. Nguyen Cao Van (2012) - *Theory of Probability and Statistics Syllabus*, National Economics University Publishing House.

5. Tran Thi Kim Thu (2011), *Theory of Statistics Syllabus*, National Economics University Publishing House.

CHAPTER 5: PRINCIPAL COMPONENT ANALYSIS

PCA is one of the methods of multivariate data analysis that is used much in the statistics by analyzing the mutual impact relationship between a large number of variables in order to summarize information into smaller data in homogeneous groups (known as the Principal Component) while minimizing the loss of the original information. This chapter presents the technique of analyzing principal components as well as the steps taken to analyze the principal components.

Content

5.1. Data summarising Techniques

5.2. The general concept of PCA

5.2.1. Concept

5.2.2. Advantages and limitations of PCA

5.2.3. Application of PCA

5.3. The order of implementation of PCA

5.3.1. Determine the purposes of research, select analysis variable

5.3.2. Set up correlation coefficient matrix

5.3.3. Determine number of major components

5.3.4. Rotate the principal component shaft

5.3.5. Select name and explain the principal component

5.3.6. Interpret the factor matrix

5.3.7. Perform the variables in the plane created by the main component

5.3.8. Determine multiple

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

3. Hoang Trong, Chu Nguyen Mong Ngoc (2008) - *Analyzing research data with SPSS*, University of Economics Ho Chi Minh city Publishing House.

4. Nguyen Cao Van (2012) - *Theory of Probability and Statistics Syllabus*, National Economics University Publishing House.

5. Tran Thi Kim Thu (2011), *Theory of Statistics Syllabus*, National Economics University Publishing House.

CHAPTER 6: CORRESPONDENCE ANALYSIS

Correspondence analysis method is widely used in the analysis of large data sets. This is a visual method for analyzing multi-dimensional data. The results of the analysis are cognitive maps that help perform the position of all the manifestations of the variables on the same plane. The contents of the chapter will present analytical methods corresponding to bivariate Correspondence and multiple Correspondence analysis cases.

Content:

6.1. Bivariate correspondence analysis

- 6.1.1. General introduction of the bivariate correspondence analysis
- 6.1.2. Terms used in the correspondence analysis
- 6.1.3. The order of implementation of BCA

6.2. Multiple correspondence analysis

- 6.2.1. Concept of multiple correspondence analysis
- 6.2.2. The order of implementation of MCA
- 6.2.3. Rule explained in the multiple correspondence analysis

Texts and readings for the chapter:

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.
2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.
3. Hoang Trong, Chu Nguyen Mong Ngoc (2008) - *Analyzing research data with SPSS*, University of Economics Ho Chi Minh city Publishing House.
4. Nguyen Cao Van (2012) - *Theory of Probability and Statistics Syllabus*, National Economics University Publishing House.
5. Tran Thi Kim Thu (2011), *Theory of Statistics Syllabus*, National Economics University Publishing House.
6. Brigitte Le Roux, Frederic Lebaron, Johannes Hjellbrekke (2012) *Multiple Correspondence Analysis (MCA)*, *Workshop GDA in Social Science*, Berkeley, October 1-5.
7. Brigitte Le Roux, Frederic Lebaron, Johannes Hjellbrekke, *Multiple Correspondence Analysis (MCA)* (2010), QASS series No 163; SAGE.
8. Ngo Van Thu (2005), *Practical Statistics Syllabus*, Publishers of Hanoi scientific and technical.

9. Ludovic Lebart, Marie Piron, Mireille Razafindrakoto, François Roubaud and Jean-Pierre Cling (2008), *Documents in Tam Dao courses - Data Analysis Level II: Consolidation and Application for analyzing the job market and informal sector in Vietnam*

10. Vu Nguyen, Dao The Anh, (2005), *Training materials on multivariate statistical analysis applied in the restructuring of agriculture and rural economy*, Institute of Agricultural Science and Technology of Vietnam.

11. ALVIN C. RENCHER, 2002, *Methods of Multivariate Analysis*, Brigham Young University, A JOHN WILEY & SONS, INC. PUBLICATION Second Edition.

12. Jacqueline J. Meulman, Willem J Heiser, 2004, *SPSS Categories® 13.0*, SPSS Inc.

7. REQUIRED TEXTBOOK & COURSE MATERIALS:

Data Mining – translation Book, Tran Thi Kim Thu (2013)

8. RECOMMENDED TEXTS & OTHER READINGS

1. Jiawei Han, Micheline Kamber, Jian Pei (2011), *Data Mining Concepts and Technique, Third Edition*, Morgan Kaufmann.

2. Nguyen Hoang Tu Anh (2008), *Lectures on Data mining and applications*, Ho Chi Minh City University of Natural Sciences.

3. Hoang Trong, Chu Nguyen Mong Ngoc (2008) - *Analyzing research data with SPSS*, University of Economics Ho Chi Minh city Publishing House.

4. Nguyen Cao Van (2012) - *Theory of Probability and Statistics Syllabus*, National Economics University Publishing House.

5. Tran Thi Kim Thu (2011), *Theory of Statistics Syllabus*, National Economics University Publishing House.

6. Brigitte Le Roux, Frederic Lebaron, Johannes Hjellbrekke (2012) *Multiple Correspondence Analysis (MCA), Workshop GDA in Social Science*, Berkeley, October 1-5.

7. Brigitte Le Roux, Frederic Lebaron, Johannes Hjellbrekke, *Multiple Correspondence Analysis (MCA)* (2010), QASS series No 163; SAGE.

8. Ngo Van Thu (2005), *Practical Statistics Syllabus*, Publishers of Hanoi scientific and technical.

9. Ludovic Lebart, Marie Piron, Mireille Razafindrakoto, François Roubaud and Jean-Pierre Cling (2008), *Documents in Tam Dao courses - Data Analysis Level II: Consolidation and Application for analyzing the job market and informal sector in Vietnam*.

10. Vu Nguyen, Dao The Anh, (2005), *Training materials on multivariate statistical analysis applied in the restructuring of agriculture and rural economy*, Institute of Agricultural Science and Technology of Vietnam.

11. ALVIN C. RENCHER, 2002, *Methods of Multivariate Analysis*, Brigham Young University, A JOHN WILEY & SONS, INC. PUBLICATION Second Edition.

12. Jacqueline J. Meulman, Willem J Heiser, 2004, *SPSS Categories® 13.0*, SPSS Inc.

9. COURSE ASSESSMENT METHOD

Comply with the current regulations of the National Economics University:

- Teachers' evaluation: 10%
- Mid-course test: 30%
- Final examination: 60%

(Students are eligible to take the final test if: the evaluation of teachers is at least 5, the minimum mid-course test score is 3)

HEAD OF DEPARTMENT

(signed)

MSc. Do Van Huan

Hanoi, 2016

PRESIDENT

(signed)

Prof.Dr. Tran Tho Dat