Big Data Analytics

Course Teacher: Xiaojun WU

Course Code：

Teaching Language: English

Targeted Students: Master Candidate

Contact Hours: 36

Extracurricular class hours: 72

Prerequisites: Applied Statistics

Number of learners: Less than 40

Semester: Fall

Credits: 2

1. Course Description

"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application and relative software. Most of the methods of big data analysis are closely associated with the cutting-edge artificial intelligence, especially machine learning algorithms. The course Big Data Analytics (BDA) is designed for master candidate students in order to let them understand the evolution of BDA, tendency of BDA and major methods of BDA. By studying the course, the student will learn not only how to deal with big data in the business circumstances for deeply understanding the scenarios and making optimizing decision, but also how to apply these methods in the academic studies. The course concerns itself with major BDA algorithms, such as logistic regression, decision tree, random forest, support vendor machine and neural network. Although all these mainstream algorithms will be introduced in the course in a way that is easy to understand, the course will mainly focus on the applications of BDA, other than the mathematic details of these algorithms. It's not compulsory for students to have programming experience previously, however it would be better to understand the course if they have such experience. Weka, a mature computing software of BDA, will be used in the course as an important computing platform to apply the BDA algorithms without writing any codes.

2. Course Objectives and Requirements

1)**Course Objectives**：
This course exposes master candidates to essential knowledge, primary principles and research methods in BDA to enable them to know how to understand and analyze big data in business and management and have some insight into the application of intelligent decision.

In addition, this course aims to cultivate the perspective of cutting edge technologies with regard to artificial intelligence and data-driven management. The students will not only learn relevant research methodologies in the academic field, but also improve their ability to analyze and solve practical problems arising in business and management from theoretical and practical perspectives.

2)**Requirements**：

Although it's not necessary to have experience in computer programming, students are required to have basic mathematic knowledge as prerequisites such as probability theory, linear Algebra and applied statistics. Students are expected to get prepared for lectures, finish and submit assignments as required.

3. Course Arrangement

Unit	Credit hours	Contents	Preparation of class and reading materials
1	2	**Unit One Introduction to the BDA -1** 1.History of Big Data and Machine earning 2.Types of machine learning 3.Basic procedure of BDA application	Chapter 1 & 2 of Textbook（Gareth James etc.） Chapter 2 of Textbook（Aurélien Géron）
2	2	**Unit Two：Introduction to the BDA -2** 1.Bias-Variance trade-off 2.Over-fitting and under-fitting 3.Batch and Online learning	Chapter 2 of Textbook（Aurélien Géron）
3	2	**Unit Three：Regression** 1.Simple liner regression 2.Multiple linear regression	Chapter 3 of Textbook（Gareth James etc.）
4	2	**Unit Four：Classification -1** 1.Logistic regression 2.Bayes Classifier 3.KNN regression and Classifier	Chapter 4 of Textbook（Gareth James etc.）
5	2	**Unit Five：BDA Practice -1** 1.Introduction to weka 2.Practice of LR, KNN and Bayes Classifier	Install weka in the laptop Download the dataset
6	2	**Unit Six：Classification -2** 1.Linear discriminant analysis 2.Quadratic discriminant analysis 3.Confusion matrix	Chapter 4 of Textbook（Gareth James etc.）
7	2	**Unit Seven：Resampling Methods** 1.Cross-Validation 2.ROC curve 3. The Bootstrap	Chapter 5 of Textbook（Gareth James etc.）

8	2	**Unit Eight：Ridge and Lasso regression** 1.Penalty and regularization 2.Ridge regression 3.Lasso regression 4.Gradient Descent (SGD)	Chapter 6 of Textbook （Gareth James etc.）
9	2	**Unit Nine：Decision Tree** 1.Introduction to decision tree 2.Gini index and entropy 3.ID3, C4.5 and CART 4.Pruning	Chapter 8 of Textbook （Gareth James etc.） Chapter 6 of Textbook （Aurélien Géron）
10	2	**Unit Ten：Ensemble methods and RF** 1. Introduction 2. Random Forest 3.Adboost 4.GBDT and XGboost	Chapter 8 of Textbook （Gareth James etc.） Chapter 7 of Textbook （Aurélien Géron）
11	2	**Unit Eleven：BDA Practice -2** 1. Practice of LDA, QDA 2. Practice of ridge & lasso regression 3. Practice of GBDT and Random Forest	
12	2	**Unit Twelve：Support Vector Machine -1** 1. Hyperplane 2. Kernel function	Chapter 9 of Textbook （Gareth James etc.） Chapter 5 of Textbook （Aurélien Géron）
13	2	**Unit Thirteen：Support Vector Machine -2** 1.SVM for classification 2.SVM for regression 3.Parameter selection & grid search in Weka	Chapter 9 of Textbook （Gareth James etc.） Chapter 5 of Textbook （Aurélien Géron）
14	2	**Unit Fourteen：Clustering** 1.Introduction 2.K-means clustering 3.Expectation-Maximization Algorithm	Chapter 10 of Textbook （Gareth James etc.）
15	2	**Unit Fifteen：Multilayer Perceptron and Neural Network** 1.Introduction 2.The perceptron 3.Multilayer Perceptron 4.Backpropagation Algorithm 5.Introduction to Deep Neural Network	Chapter 10 of Textbook （Aurélien Géron）
16	2	**Unit Sixteen：ML Practice -3** 1. Practice of SVM 2. Practice of EM 3. Practice of K-means clustering 4. Practice of MLP	

4. Teaching Methods

Lectures, Discussions, Case Analysis, Debate, etc.

5. Learning Outcomes

Category	Learning Outcomes
Master of Knowledge	1. Master the basic theory and development trend of Big Data Analytics, 2. Master the basic principles and methods of Big Data Analytics, 3. Master the basic approaches of Big Data Analytics, such as decision tree, random forest, SVM and MLP.
Intellectual abilities learned	4. Have the ability to select the applicable Big Data Analytics methods in face of a business or an academic problem. 5. Have the ability to apply the Big Data Analytics in the business and management 6. Have the ability to apply the Big Data Analytics in the academic study
Practical skills learned	7. Familiar with the skills of application of mainstream software of Big Data Analytics 8. Master the skills to build model based on the data and issues in the business and management 9. Master the skills to build model based on the data and issues in the academic study
Personal competences and characters Cultivated	10. The competence of solving problem by using data driven methods 11. The management competence of being good at decision making by Big Data Analytics 12. Open mind for the cutting edge Big Data Analytics methods

6. Performance Evaluation: Means & Ratio

Evaluation Means	Ratio （%）	Link with learning outcomes expected
Team and individual assignments	40	Focusing on exercises analysis, evaluating the abilities of using BDA knowledge related to analyzing and solving the problems

Attendance and Engagement	20	Evaluating the engagement in the course study and participation in the course discussion
Final paper	40	Evaluating the degree of mastering the concepts, methodologies in a real case of business or academic problem. Through the paper writing, the level in mastering and applying the knowledge and methodology of students can be assessed.

Preparation:

Students are requested to read all the pre-reading materials for preparation of each session.
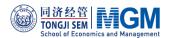
7.Textbook，References and Reading Materials

1) Textbook

[1]. Gareth James,Daniela Witten,Trevor Hastie,Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer. 2013

[2]. Aurélien Géron. Hands-On Machine Learning with Scikit-Learn & TensorFlow. OREILLY. 2017

[3]. Mattew Kirk. *Thoughtful Machine learning with Python: A test-driven approach*. O'REILLY,2017.

2) Online resources

■ Textbook #1 related resources

[1]. http://www-bcf.usc.edu/~gareth/ISL/

[2]. https://github.com/qx0731/ISL_python

[3]. https://github.com/JustinMShea/StatisticalLearning

■ Python

[4]. https://wiki.python.org/moin/BeginnersGuide

■ Weka

[5]. https://www.cs.waikato.ac.nz/ml/weka/

[6]. http://weka.wikispaces.com/

[7]. online course: https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/

■ Machine Learning online course materials

[8]. http://cs229.stanford.edu/syllabus.html

■ Others

[9]. http://scikit-learn.org/stable/index.html

[10]. https://github.com/

3) Other References

[1]. Sebastian Raschka.*Python Machine Learning*. Packt Publishing 2015. UK

[2]. Ethem Alpaydin. *Introduction to Machine Learning (Third Edition)*. MIT Press, 2014

[3]. Yuxi(Hayden) Liu. *Python Machine Learning by example*. Packt Publishing 2017

[4]. John Paul Mueller, Luca Massaron. *Machine Learning for dummies*. John Wiley & Sons, Inc. 2016

8. Assignment Requirements

Two assignments will be assigned during the course. Students should analyze the assignments, conduct the data processing and give answers.

1) individual work
2) submit the word file of answer to professor on time

The criteria of assignment evaluation (100 points)

1) Analysis and Data Analysis   80 points
2) Submission on time   20 points

9. Final Paper Requirements

The students should select a case or interesting problem in your academic study or business practice, then define the problem and analyze the data with methods you have learned in the course. A formal paper is required to write based on the data analysis. The paper should include introduction, literature review, methodology, analysis, discussion, conclusion and references.

Requirements

1) Individual work
2) Submit the word file of paper and dataset to professor on time

The criteria of final paper evaluation (100 points)

3) Literature Review   20 points
4) Structure of the paper   15 points
5) Analysis and Data Analysis   30 points
6) Reference   20 points
7) Formats   15 points

Appendices:

1. The course PPT

2. Other teaching material (such as reading materials, the articles)