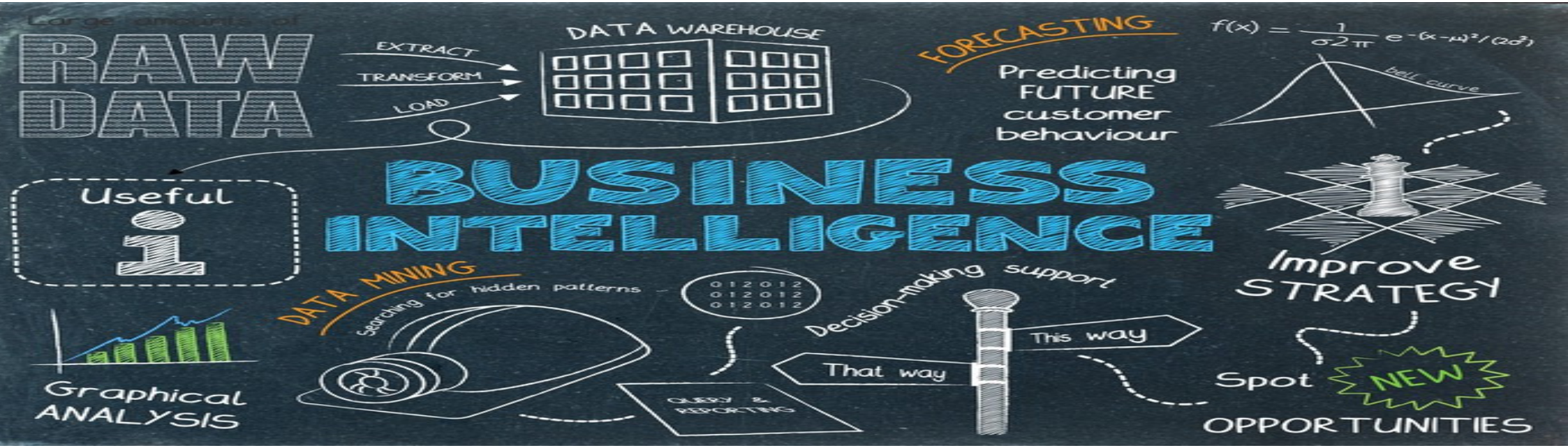
UECS3213 / UECS3453 Data Mining

Introduction



Dr. Simon Lau Boung Yew

About me

- Dr. Simon Lau Boung Yew 刘本佑
- Room: KB 8th Floor FE3(1)
- Email: simonlau@utar.edu.my
- Consultation Hour:
  - Every Thursday 9am - 1pm

Synopsis

- Advances in data generation and collection are producing **data sets** of **massive size** (big data) in commerce and a variety of scientific disciplines.
- **Data analysis techniques** are becoming a necessity. This course presents fundamental concepts and algorithms of data mining techniques in detail thus providing the students with the background for the application of data mining to real problems.
- It introduces **programming tools** (Python libraries) for students to apply their concepts to implement and evaluate data mining techniques to solve problems.

Pre-requisite

- UECS1203 Database System Fundamentals OR
- UECS1403 Database System Fundamentals OR
- UECS1004 Programming and Problem Solving OR
- UECS1643 Fundamentals of Programming

Class Schedule

- Duration: 14 (16 Jan – 23 Apr 2019) + 3 weeks
- Lectures:
  - Every Monday 12pm-2pm (2 hours) Venue: KB316
- Practical:
  - Every Friday 8:30am-10:30am (2 hours) Venue: KB606 (UECS3213)
  - Every Friday 8:30am-10:30am (2 hours) Venue: KB606 (UECS3453)

Objectives

- To highlight the **importance** of data mining and its applications
- To introduce data mining **concepts**, **theories** and **implementations**.

Course Learning Outcomes

- **CO1:** Identify the key **technological foundations** of data mining
- **CO2:** Create **programming** solutions using data mining **techniques** for given problem
- **CO3: Evaluate performance** of data mining solutions for a given problem
- **CO4:** Construct a data mining **project** as a team
- **CO5:** Recognize the importance of data mining techniques and its **applications** in the industry

Outline

- Topic 1: Introduction to Data Mining
- Topic 2a: Data
- Topic 2b: Data Exploration
- Topic 3a: Classification - k-NN & Decision Tree Classifier
- Topic 3b: Classification - Naive Bayes, Support Vector Machine & Ensemble Methods
- Topic 4: Regression Analysis
- Topic 5: Cluster Analysis and Anomaly Detection
- Topic 6: Association Analysis

Topic 1: Introduction to Data Mining

- Overview of Data Science
- What is data mining? What is not data mining?
- DIKW Knowledge Pyramid
- Terminologies
- Origin of data mining
- Why data mining is important?
- Data Mining Methods
- Challenges of Data Mining
- Data Mining Applications

Topic 2a: Data

- What is data?
- Attribute and Attribute Value
- Types of Data Sets
- Structured vs Unstructured Data
- Data Processing
- Curse of Dimensionality
- Similarity and Dissimilarity Measures

Topic 2b: Data Exploration

- What is data exploration?
- Exploratory Data Analysis (EDA)
- Summary Statistics
- Visualization
- Visualization Techniques
- OLAP

Topic 3: Classification

- Theories
- Examples
- K-Nearest Neighbours
- Decision Trees
- Naïve Bayes
- Support vector machine
- Ensemble Method

Topic 4: Regression

- What is a model?
  - Deterministic Models
  - Probabilistic Models
- Regression model
- Correlation
- Correlation vs Regression
- What is regression?
- Simple Linear Regression
- Model Assessment
- Multiple Linear Regression
- Logistic Regression

Topic 5: Cluster Analysis and Anomaly Detection

- What is Cluster Analysis?
- Applications of Cluster Analysis
- Types of Clustering
- Types of Clusters
- Clustering Algorithms
- Anomaly / Outlier Detection
- Causes of Anomalies
- Distinction Between Noise and Anomalies
- Model-Based Anomaly Detection
- Anomaly Detection Techniques

Topic 6: Association Analysis

- Association Rule Mining
- Association Rule Mining Task
- Mining Association Rules
- Frequent Itemset Generation Strategies
- Apriori principle

Course Delivery

- Classroom Lectures (2 hours per week)
- Practical / Tutorial Sessions (2 hours per week)
- Student-Lecturer consultation / discussion
- Self Study

Course Assessment

Assessment Type	Percentage (%)
Coursework	**50**
Assignment 1 • Summary and critical analysis of talk	10
Assignment 2 • Hands-on assignment	10
Assignment 3 • Hand-son programming assignment • Technical report • Group Presentation	30
Final Exam	**50**
Total	**100**

Attendance

- Attendance is compulsory.
- Students with attendance < 80% will be barred from the final exam.

References (Lectures)

- C. C. Aggarwal. (2015). Data Mining: The Textbook. Springer
- Harrington, P (2012). Machine Learning in Action. Manning Publications.
- Pang-Ning Tan, Michael Steinbach and Vipin Kumar. (2006). Introduction to Data Mining (2nd / 3rd Edition). Pearson Addison-Wesley
- Jiawei Han and Micheline Kamber. (2006). Data Mining: Concepts and Techniques (2nd Edition). Morgan Kaufmann.

References (Labs)

- J. Grus (2015). Data Science from Scratch: First Principles with Python. O'Reilly Media.
- Richert, W. and Coelho, L.P. (2013). Building Machine Learning Systems with Python. Packt Publishing.

Important Dates

- **Semester:** 14/1 - 19/4
- **Chinese New Year**: 4/2 - 8/2
- **Assignment 1**: 18/2 (Talk), Week 10 22/3 (Report Submission)
- **Assignment 2:** Week 10: 22/3 (Report Submission)
- **Assignment 3:** Week 13-14 (Group Presentation), 12/4 (Report/Code Submission)
- **Final Exam:** 22/4 - 10/5

Small Survey

- Why Data Mining? Do you know?
- Ability in Programming
- Ability in Mathematics
- Taken Artificial Intelligence before?
- Python programming language

Teaching Plan

Week	Lecture Topic	Tutorial / Practical Topic	Assessments / Specific Task
1	Topic 1: Introduction to Data	Lab 1: Familiarization with Data Mining Tools: Python, RAnaconda Python, Jupyter NotebookR, RStudio Tutorial 1 - Introduction to Data Mining	Assignment 1 & 2 question release
2	Topic 2a: Data	Lab 2: Introduction to Python Programming & Python Data Science Libraries: Pandas, NumPy, SciPy, matplotlib, scikit-learn Tutorial 2 - Data	Assignment 3 question release

Teaching Plan

3	Topic 2b: Data Exploration	Lab 3: Introduction to NumPy Library Tutorial 2 - Data	
4	Chinese New Year	Revision	-
5	Topic 3a: Classification (Part 1)	Lab 4: Introduction to Pandas DataFrame Tutorial 3 - Classification	
6	Topic 3a: Classification (Part 1)	Lab 5: Introduction to Data Visualization Methods in Python using matplotlib Tutorial 3 - Classification	Assignment 1: Industry TalkSummary of TalkOpinionsTentative date: 18/2/2019

Teaching Plan

7	Topic 3b: Classification (Part 2)	Lab 6: Implementing K-Nearest Neighbors in scikit-learn Tutorial 3 - Classification	Assignment 2 releaseProgramming in Python, numpy, pandas, matplotlib, scikit-learn
8	Topic 3b: Classification (Part 2)	Lab 7: Implementing Decision Tree using scikit-learn Tutorial 3 - Classification	

Teaching Plan

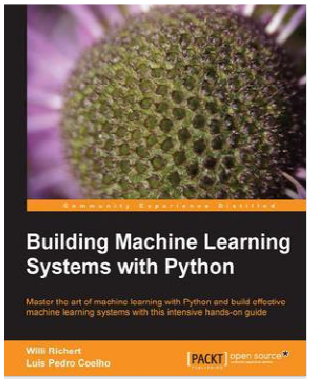
9	Topic 4: Regression Analysis	Lab 8: Naïve Bayes implementation using scikit-learn Tutorial 4 - Regression Analysis	
10	Topic 5: Cluster Analysis and Anomaly Detection	Lab 9: SVM implementation using scikit-learn Tutorial 5 - Cluster Analysis	Assignment 1 & 2 submission Tentative date: 22/3/2019

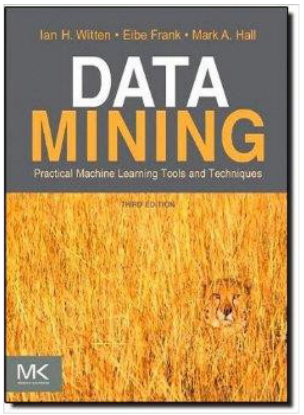
Teaching Plan

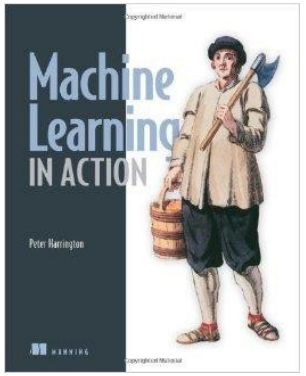
11	Topic 5: Cluster Analysis and Anomaly Detection	Lab 10: k-Means clustering implementation using scikit-learn Tutorial 5 - Cluster Analysis	
12	Topic 6: Association Analysis	Tutorial 6 - Association Analysis	
13	Project Presentation		Assignment 3 submission & Group Presentation Tentative date: 12/4/2019
14	Project Presentation / Revision		

Main References

- Harrington, P (2012). Machine Learning in Action. Manning Publications.
- Richert, W. and Coelho, L.P. (2013). Building Machine Learning Systems with Python. Packt Publishing.
- Witten, I.H, Franck, E, and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. (3rd ed.). Morgan Kaufmann.
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar (2018). Introduction to Data Mining (2nd Edition), Pearson







Other References

- J. Grus (2015). Data Science from Scratch: First Principles with Python. O'Reilly Media.
- C. C. Aggarwal. (2015). Data Mining: The Textbook. Springer
- Richert, W. and Coelho, L.P. (2013). Building Machine Learning Systems with Python. Packt Publishing.
- Russel M.A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. (2nd Ed). O'Reilly Media.

Tools

- Python 3
- Anaconda Python
  - numpy, scipy, matplotlib, pandas, scikit-learn, etc.
  - Jupyter notebook
  - Spyder notebook

What you will learn: Theory and Concept

- Classification
- Regression
- Clustering
- Hands-on Practicals:
  - Python and its environments
    - numpy
    - scipy
    - matplotlib
    - pandas
    - scikit-learn

Thank you