Department of 統計及精算學系
Statistics & Actuarial Science

STAT3612 Data Mining (2016-17 Semester 2)

Course Oultine

Instructor:	Dr. Aijun Zhang
Email:	ajzhang@hku.hk
Office:	RR224
Telephone:	3917 1984

Lecture Hours:	Monday 12:30pm – 2:20pm (LE7/RR101)
	Thursday 12:30pm – 1:20pm (LE7/RR101)
Tutor:	Mr. Jia You (RR116)
Email:	u3005315@hku.hk
Tutorial Hours:	TBD (RR101)
Course Website:	http://www.statsoft.org/teaching/stat3612 + moodle.hku.hk

Course Objectives:
With an explosion in information technology in the past decade, huge amount of data appears in various fields such as finance, marketing research, customer relationship management, medicine and healthcare. The challenge of understanding these data with the aim of creating new knowledge and finding new relationships among data attributes has led to the innovative usage of statistical methodologies and development of new ones. In this process, a new area called data mining is spawned. This course provides a comprehensive and practical coverage of essential data mining concepts and statistical models for data mining.

Prerequisites:
STAT2602 (Probability and Statistics II) or STAT3902 (Statistical Models).

Intended Learning Outcomes:

1. Understand and apply a wide range of data mining techniques, and recognize their characteristics, strengths and weaknesses.
2. Be proficient with the leading data mining software - R, Python, or SAS Enterprise Miner.
3. Identify and use appropriate data mining techniques for a data mining project, taking into account both the nature of the data to be mined and the goals of the user of the discovered knowledge.
4. Evaluate the quality of discovered knowledge, taking into account the requirements of the data mining task being solved and the goals of the user.

Contents and Topics:
Data science, machine learning, data manipulation, exploratory data analysis, linear methods, variable selection, regularization, model assessment, Bayes classifier, decision trees, ensemble methods, support vector machines, neural networks, principal component analysis, clustering, sparse coding.

Assessment:

Assignments:	Three assignments (Hands-on data analysis and modeling)	30%
Tests:	Two in-class tests (Problem solving and calculations)	40%
Group Project:	Project proposal, oral presentaiton and written report	30%

Software/Programming:
R (R Studio), Python (IPython), or SAS (Enterprise Miner)

References:

1. James, G., Witten, D, Hastie, T. and Tibshirani R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

2. Hastie, T, Tibshirani, R. and Friedeman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition, Springer, New York.

3. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

4. Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H. T. (2012). *Learning From Data: A Short Course.* `http://www.amlbook.com/`

5. Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective.* MIT Press.

6. Tan, P. N., Steinbach, M. and Kumar, V. (2014). *Introduction to Data Mining.* Second Edition, Addison Wesley.

7. Wickham, H. and Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly. `http://r4ds.had.co.nz/`

8. Raschka, S. (2015). *Python Machine Learning.* Packt.

9. SAS Institute (2015). *Getting Started with SAS Enterprise Miner 14.1.*