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

STAT3612 Data Mining (2018-19 Semester 2)

Course Oultine

Instructor:	Dr. Aijun Zhang (RR224)
Email:	ajzhang@hku.hk
Lecture Hours:	Monday 5:30pm – 6:20pm (T3)
	Thursday 4:30pm – 6:20pm (T3)
Tutor:	Dr. Gilbert Lui (RR118) (csglui@hku.hk)
Tutorial Hours:	TBD (Starting from Week 2)
Course Website:	http://stat3612.saas.hku.hk & http://moodle.hku.hk

Course Objectives:
Machine learning is the study of computer algorithms that build models of observed data in order to make predictions or decisions. Statistical machine learning emphasizes the importance of statistical theory and methodology in the algorithmic development. This course provides a comprehensive and practical coverage of essential machine learning concepts and a variety of learning algorithms under supervised and unsupervised settings. The course materials are presented with lots of examples and reproducible codes.

Contents and Topics:
Data science, data exploration, generalized linear models, variable selection, basis expansion, regularization, cross-validation, tree-based methods, kernel methods, neural networks, dimension reduction, principal component analysis, cluster analysis, stochastic optimization, interpretable machine learning.

Intended Learning Outcomes:

1. Get familiar with the workflow of a data science or machine learning project;
2. Understand and apply a wide range of statistical machine learning methods, and recognize their characteristics, strengths and weaknesses;
3. Identify and use appropriate techniques for a particular data science project;
4. Evaluate the quality of the resulting model in terms of prediction accuracy and model explainability;
5. Apply R/Python programming for solving data-scientific problems.

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

Assessment:

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

Software/Programming:
R (R Studio), Python (Notebook), TensorFlow/Keras.

References:

1. James, G., Witten, D, Hastie, T. and Tibshirani R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, New York. `http://www-bcf.usc.edu/~gareth/ISL/`

2. Hastie, T, Tibshirani, R. and Friedeman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition, Springer, New York. `https://web.stanford.edu/~hastie/ElemStatLearn/`

3. Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, OReilly. `https://github.com/ageron/handson-ml`

4. Chollet, F. (2018). *Deep Learning with Python.* Manning. `https://www.manning.com/books/deep-learning-with-python`

For future data scientisits: