NATIONAL UNIVERSITY OF SINGAPORE
NUS Business School
Department of Decision Sciences

DSC5103 Statistics

Lecturer : Associate Prof Wang Tong

Session : Semester I, 2016/2017

Aims & Objectives

This course aims to provide a holistic overview of the modern Statistical Learning toolbox. Different from traditional Statistics courses, this course (1) emphasizes on understanding the intuition behind the tools and not on deriving the underlying mathematics; (2) incorporates real-world datasets and analytics projects to help you bridge theories and practices; and (3) equips you with hands-on experiences in using data analysis software (R and/or Python) to visualize the concepts and ideas and also solve exercises.

The course covers most of the commonly used analytics tools such as linear/logistic regression and decision tree. Two exceptions, Support Vector Machine and Neural Network (which are arguably more of Computer Science tools), are left to other modules.

The students are expected to get their hands really dirty by applying (and even messing up) the tools in analytics software (R and/or Python).

Prerequisites

- Probability Theory and classic Statistics (College level)
    - Random Variables, Mean, Variance, Correlation
    - Conditional Probability, Bayes' Theorem
    - Basic Probability Distributions
    - Sampling, Confidence Interval, Hypothesis Testing, P-value
    - Time Series Analysis (Exponential Smoothing, ARIMA)
- R programming fundamentals

Note: I shall conduct a 3-day bootcamp for the above topics in late July or early August (to be announced on IVLE). Those without relevant background are strongly encouraged to participate. Materials (and perhaps video recording) will be available for those who cannot make it.

Topics

Week 1. Overview of Statistical Learning

1. Descriptive, Predictive, and Prescriptive Analytics
2. Supervised, Unsupervised, and Reinforcement Learning
3. Classification and Regression
4. Bias-Variance Trade-off, Under-fitting vs. Over-fitting

Week 2. K-Nearest Neighbors Algorithm

Week 3. Linear Regression

1. Simple and Multiple Linear Regression

2. Interpreting Regression Output
3. Model Selection
4. Introducing interactions and nonlinearity

Week 4-5. Generalizations of Linear Regression
1. Logistic Regression and Maximum Likelihood Estimation
2. Poisson Regression and other Generalized Linear Models

Week 6. Resampling Methods
1. Cross-validation
2. The Bootstrap

Week 7. Linear Model Selection Revisited
[In-class Test 1]

Week 8. Regularization
1. Ridge Regression
2. The Lasso
3. Elastic Net

Week 9. Tree-based Methods I
1. Decision Trees

Week 10. Tree-based Methods II
1. Bagging, Random Forest
2. Gradient Boosting Machines

Week 11-12. Unsupervised Learning
1. K-Means Clustering and Hierarchical Clustering
2. Gaussian Mixture Model and the Expectation-Maximization Algorithm
3. Dimension Reduction by Principle Component Analysis

Week 13. Review and Miscellaneous Topics
[In-class Test 2]

Text Book

An Introduction to Statistical Learning – with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 2013, Springer-Verlag New York.

http://www-bcf.usc.edu/~gareth/ISL/index.html

Reference Book (for those who want more theories and math)

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, by Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009, Springer-Verlag New York.

http://statweb.stanford.edu/~tibs/ElemStatLearn/

ASSESSMENT (100% CA)

Class Participation (Individual)	10%
Assignments (Group)	30%
In-class Tests (Individual)	40%
Final Project (Group)	20%