Statistical Learning: Theory and Practice
統計學習初論

Spring, 2019
725U3550 / IM 5044

Instructor: Hsin-Min Lu (盧信銘)
Email: luim@ntu.edu.tw
Office Phone: 33661184
Office Hours: By appointment

Class Meetings: Tue. 9:10pm-12:10pm at Mgmt Bldg I, Room 402 (管一 402)

Required Text: Pattern Recognition and Machine Learning by Christopher M. Bishop; ISBN 0-387-31073-8.
Hands-on Machine Learning with Scikit-Learn & Tensorflow by Aurelien Geron; ISBN 978-1-491-96229-9.

Course Description

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer sciences and machine learning. The field encompasses many methods such as the regularized regression, classification, graphic models, and approximation inference. This course is appropriate for master's students and advanced undergraduates who wish to use statistical learning and machine learning tools to analyze their data.

The goal of this course is to introduce a set of tools for data analytics. We will cover the principles and applications of these models/tools. These tools will not be viewed as black boxes. Instead, students will be exposed to the details, not just the use, of these tools. The main reason is that no single approach will perform well in all possible applications. Without understanding how a tool work, it is impossible to select the best tool.

Grade Distribution

The course grades will be determined by the following percentages:

Homework (Python-based) (7 Assignments)	60%
Attendance, participation & quizzes	10%
Final Project 1 (Relative prediction performance)	15%
Final Project 2 (Presentation)	15%
Total	100%

Computational Tools

Students are required to use Python 3 (with scikit-learn, pandas, matplotlib, numpy, etc.) in selected homework problems.

Homework

There are about seven graded assignments. Unless otherwise specified, students are required to organize their code and results using Jupyter Notebook/Lab and submit their homework to Ceiba using the HTML format. An assignment is due at the beginning of the first class in the following week. Late submissions will not be accepted. Homework assignments play a very important role in the learning process, and students are expected to spend a significant amount of time in solving homework problems. Students are allowed to discuss about homework questions. However, each student must turn in her/his own homework. Cheating will result in severe penalty for everyone involved.

Final Project 1

Students are going to work on a well-defined prediction problem as a team. A team may consist of three to four students. Each team can submit their prediction to a ranking system. The score a team receive will depend on the relative performance among all participating teams. Details will be given in class.

Fianl Project 2

Students are expected to form teams of three to six people and work on a data analytics problem that is interesting and challenging for you. Details will be given in class.

Approximate Course Schedule

Week	Date	Topic	Note
1	2/19	Introduction	
2	2/26	Regressions and Regularization (HW1)	
3	3/5	Regression and Linear Classification Models	
4	3/12	Linear Classification Models (HW2)	
5	3/19	Performance Evaluation	
6	3/26	Feature Selection (HW3)	
7	4/2	Holiday. No class.	
8	4/9	Dimension Reduction (HW4)	
9	4/16	Gaussian Process Regression and Classification	
10	4/23	Tree-based Models (HW5)	
11	4/30	Tree-based Model, Graphical Models (期末報告題目確定)	
12	5/7	Graphical Models (HW6)	
13	5/14	Graphical Models, Chinese Word Segmentation	
14	5/21	Mixture Models and EM Algorithm (HW7)	
15	5/28	Topic Models	
16	6/4	No class.	
17	6/11	Final Project 2 Presentation	
19	6/28	Final Project 1 Due (Competition-based)	