

IE 582 STATISTICAL LEARNING FOR DATA MINING

Fall 2018

Instructor: Mustafa Gökçe Baydoğan (mustafa.baydogan@boun.edu.tr)

Office hrs: To Be Decided (TBD)

Lecture hours: Tuesday: 13:00-13:50 and Thursday: 9:00-10:50 (Room: IE Undergraduate Computer Lab)

Teaching assistant: TBD

Course website: <http://moodle.boun.edu.tr/>

Textbooks:

- *An Introduction to Statistical Learning with Applications in R*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, NY, 2013
(available online: <http://www-bcf.usc.edu/~gareth/ISL/>)
- *Elements of Statistical Learning : Data Mining, Inference, and Prediction*, T. Hastie, R. Tibshirani, J. Friedman, Springer, NY, 2009
(available online: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>)

About the course:

- A survey course for topics from data mining and machine learning is presented. Advantages and disadvantages of methods are discussed. The emphasis of this course is on data models and concepts, rather than inference.
- R software will be used for examples. Any other software, programming language, etc. that you feel yourself comfortable are welcome.
- Prerequisite background: A working knowledge of basic statistical methods. A formal course in engineering statistics at the level of IE 256 is the official prerequisite. A previous course in empirical modeling such as regression analysis or design of experiments is recommended. Some experience with matrix algebra is required.

Course objectives:

The field of data mining spans a large collection of different models. These models look fragmented and disconnected—without a route to solve a particular problem. The aim of this course is:

- To learn selected data mining models, the model objectives, steps, inputs, outputs, assumptions, advantages, disadvantages, and relationships of methods.
- To learn important concepts such as the nature of data, over and under fitting, how to evaluate a model.
- To build a basic understanding of the methods through calculations with a set of very simple data sets as a roadmap (or guidance) is needed to develop solution strategies. These examples are provided as the homework exercises, and they illustrate the objectives, steps, inputs, outputs of each model.
- To understand the assumptions, advantages, disadvantages, and relationships of methods. This provides insight into the role of the various tools in a solution (guidance). This approach allows us to focus on the key characteristics of models. We do not focus on software implementations of algorithms. Pure operation of a software package does not provide insight into a model for a particular problem. Instead, a roadmap (or guidance) is needed to develop solution strategies.

Grading Criteria

Your course grade is determined from one mid-term exam (20%), a final exam (30%), six homework (5% each) and an individual final analysis project (20%).

Requirements:

- Class attendance is very important and strongly encouraged. Some material that is not included in the book may be covered in the class.
- As part of a group, you will be asked to present homework or examples.
- An individual final project is used to apply techniques from the course on a larger data set. Here, the learning is put to use as a sequence of steps is developed and implemented for a provided data set and problem objective. In the project, when one attempts to develop and implement a model, the complexities of an analysis can arise. Still, the modeling fundamentals provide the guidance for an effective solution. Computer software is expected to be used in the project, and options will be discussed.
- Exams are open book and notes and emphasize the calculation and interpretation of model-building concepts. Computers are not permitted. A calculator is needed. The final exam is comprehensive.
- Academic integrity is expected. Your work is to be your own.

Topics to be covered

Introduction to data mining, basic notation, data size and R

Details for supervised and unsupervised learning

Supervised learning

- Classifiers: nearest-neighbor, Bayesian
- Classification and decision trees, modifications for regression
- Overfitting and evaluating performance
- Support vector machines
- Ensemble methods, boosting, multiclass analysis, priors
- Feature selection in supervised learning

Unsupervised learning

- Dimensionality reduction
- Cluster analysis and evaluation– hierarchical, K-means, alternatives

Association analysis and rule mining