

# Syllabus (Spring 2018-2019)

## DA510 Data Mining

---

### Meeting Times

Saturday 13:00 - 16:00

Wednesday 19:00 - 22:00

### Instructor

İnanç Arın e-mail: [inancarin@sabanciuniv.edu](mailto:inancarin@sabanciuniv.edu)

### Course Description, Aim and Content

Data collection has become easier and cheaper with the advances in technology which motivate data mining research and applications. Data mining aims to turn the collected massive raw data into valuable knowledge, very similar to what conventional mining does. In the scope of this course we will cover various data mining methods, namely the associations, sequential patterns, temporal patterns, clustering and classification applied to different data types such as time series and textual data. We will look at recommendation systems and personalization where data mining models play an important role. Students are expected to understand the fundamental theory behind each technique, as well as implementing them using an environment such as Python and RapidMiner. Some machine learning concepts will be recovered by applying them on Python.



## Schedule<sup>1</sup>

1. Some Machine Learning Concepts (Recap)
  - a. Linear Regression, (Stochastic) Gradient Descent
  - b. Overfitting/Underfitting
  - c. End to end machine learning application
2. Association Rules & Sequential and Temporal Patterns
3. Clustering Algorithms
  - a. Partition Based Clustering Algorithms
  - b. Hierarchical Clustering Algorithms
  - c. Density Based Clustering Algorithms and Outlier Detection with DBSCAN
4. Project Proposal Presentation
5. Text Mining
  - a. Text Clustering
    - i. Longest Common Subsequence/Substring
  - b. Text Classification
  - c. Data Structures/Indexing Methods to Improve Text Clustering (Suffix Trie, Suffix Tree, Locality Sensitive Hashing)
6. Data Mining for Recommendation Systems and Personalization
7. Time Series Analysis
8. Dimensionality Reduction (SVD)
9. Search Engines - Elasticsearch/Logstash/Kibana
10. Project Final Report Presentation
11. Final Exam

---

<sup>1</sup>Instructor may change the contents and the order of the topics during the semester

---

## Grading

- Assignments 30% (In-class & take-home, open everything)
- Final 40% (Closed book, written exam)
- Project 30% (Proposal Report & Presentation, Final Report & Presentation, Implementation)

## References

- Data Mining Concepts and Techniques, Jiawei Han & Micheline Kamber & Jian Pei, Morgan Kaufmann (3rd edition)
- Scikit-Learn: Machine Learning in Python <https://scikit-learn.org/stable/index.html>
- RapidMiner: rapidminer.com <https://docs.rapidminer.com/latest/studio/getting-started/>

## Announcements and SuCourse

Students are responsible for all announcements made during the regular class meetings. Students should check their Sabanci University e-mail accounts ([mymail.sabanciuniv.edu](mailto:mymail.sabanciuniv.edu)) and follow the SuCourse site (<https://www.sabanciuniv.edu/sucourse/eng/>) for this class regularly as they are responsible for all announcements and postings delivered through these sources.

**Important Note:** This document may be modified during the semester (Spring 2018-2019) due to unforeseen reasons.