CS 451/686-02 Data Mining Syllabus

Fall 2016

Maria Daltayanni

Instructors & TAs

- Instructor: Dr. Maria Daltayanni
  - Office: HR 510A
  - Phone: 415-422-4917
  - Email address: mdaltayanni@cs.usfca.edu
  - Website: http://www.cs.ucsc.edu/~mariadal
  - Office hours:
    - MW 10.30am – 11.30am or by appt.
    - Extra office hours before midterm and final: TBD

- TA: Lyndon Ong Yiu
  - Email address: lcongyiu@dons.usfca.edu>
  - Office hours: TBD

Class details

- Time: 8:35am – 10:20am
- Days: Monday+Wednesday
- Classroom: Lo Schiavo Science G12
- Date Range: Aug 23 – Dec 07, 2016

- Class Website: datamining.cs.usfca.edu

  Can you access it?

Prerequisites

- Courses with minimum grade **C**:
  - MATH 230 **and**
  - CS 245

- Who can enroll:
  - Undergraduate students from CS
  - Undergraduate students from Data Science
  - Graduate students from CS

Textbook & Resources

- Data Mining:
  - An Introduction to Statistical Learning with applications in R
    - by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
  - Mining Massive Datasets
    - by Jure Lescovec, Anand Rajaraman, Jeff Ullman
- R:
  - An Introduction to R
  - R Cookbook
- Python:
  - How to think like a computer scientist
  - Python for Data Analysis

Evaluation

- To receive a final letter grade, you must submit all five below:

Weekly Homework	10%
Labs	20%
Project	30%
Midterm Exam	20%
Final Exam	20%

Homework & Labs

- The homework will be multiple-choice questions
- Out on Weds
- Due in 7 days sharp, i.e. next Wed

- One or two labs per week, mostly in R
- Depending on the topic
- In the end of the in-class lab, you will upload your results (e.g. plots, predictions, etc.) to Canvas
- You will take a similar lab home

- No make-ups, sorry!
- Notify me ahead of time if you must miss a day, and explain your very important reason.

Project

- You will pick a topic for your class project among a few suggested ones.

- The project will run through 4 phases:
  1. Form 2-3 member teams
  2. Project proposal (10%)
  3. Project progress report (30%)
  4. Code+report+presentation final submission (60% for 20% each)

Lecture Layout

- Each lecture runs for 105 mins

1. First 50 mins (8:35 - 9:25am):
   - new material, discuss topics
   - Q&A
2. 10 mins (9:25am - 9:35am):
   - Break
3. Rest 45 mins (9:35am - 10:20am):
   - lab
   - lab assignment

Course Description

- This is an introductory course to Data mining. We will start with refreshing some knowledge in basic probability. You will then learn how to model problems with real world datasets, including large scale data
- Topics include: Supervised Machine Learning, Unsupervised learning (Clustering), Association rules, Nearest Neighbor Search, Locality Sensitive Hashing, Dimensionality Reduction, Model Selection, Recommendation Systems, Link analysis

Course Outline (tentative)

- Introduction and R
- Statistical Learning
- Association Rules
- Supervised Learning:
  - Linear Regression
  - Classification and Regression Trees
  - K-Nearest Neighbors
  - Support Vector Machines (SVM)
- Resampling
- Model Selection
- Unsupervised Learning:
  - Clustering
  - Principal Components Analysis (PCA)
- Recommendation Systems - collaborative filtering
- Link Analysis
  - PageRank
  - Hubs and Authorities (HITS)
- Locality Sensitive Hashing (LSH)
- Dimensionality Reduction: Singular Value Decomposition (SVD)