Data Analysis in Fundamental Sciences (MAT 388E)

Lecturer: Atabey Kaygun (`kaygun@itu.edu.tr`)

Lectures: Friday, 14:30-17:30 at D106

Books

T. Hastie, R. Tibshirani, J. Friedman	The Elements of Statistical Learning (Available on the web)
C.M. Bishop	Pattern Recognition and Machine Learning
E. Alpaydin	Machine Learning

Other resources

The books I listed are mostly theoretical. But for the computational homeworks you may need the following:

1. M. Kirk, *Thoughtful Machine Learning with Python*.
2. C. O'Neil, R. Schutt, *Doing Data Science*.
3. N. Zumel, J. Mount, *Practical Data Science with R*.
4. J. VanderPlas, *Python Data Science Handbook*.

Also, there are excellent resources on the web. I would recommend:

- edX: `https://edx.org`
- MIT-X: `https://openlearning.mit.edu`

Enroll to any of the data, machine learning, statistics, R or python classes that catches your fancy, or you think might be useful for you.

Requirements

I do require that every student installs python and R languages, and jupyter notebook environment on their own machine as I am going to give and collect homeworks online. Preparing the required computing environment on the machines you will use to write your code for your homeworks is your own responsibility. The students are also required to open a github account because I am going to collect their homeworks from their github accounts.

- The R language: `https://r-project.org`
- The python language: `https://python.org`
- Jupyter environment: `https://jupyter.org`
- Github: `https://github.com`

Homeworks and Attendance

I am not going to take attendance. But you are strongly encouraged to attend classes. **I will give 6 homeworks and collect them from github on each Friday midnight when they are due.** Check the dates below. Late homeworks or homeworks submitted via e-mail **are NOT** accepted. Debugging and testing your code is your own responsibility. For each question that involves writing and executing computer code, you will get no points if the code does not work.

Quizzes and Exams

I will give in total of 4 quizzes (on paper) and one midterm exam (on paper.) The core component of this class is data *analysis* even though we will learn about the foundations underlying the data questions we consider in mathematical, machine learning, optimization, and even linear algebraic terms. So, on paper exams and quizzes will be an important part of this class.

4 Quizzes	5% each (20% in total)
6 Homeworks	6% each (24% in total)
1 Midterm	21%
1 Final	35%

Exam Schedule

Quiz 1	October 12
Quiz 2	November 2
MIDTERM	November 16
Quiz 3	November 30
Quiz 4	December 21

Homework Schedule

Homework	Given on	Collected on
Homework 1	September 28	October 5
Homework 2	October 5	October 12
Homework 3	October 19	October 26
Homework 4	November 16	November 23
Homework 5	December 7	December 14
Homework 6	December 21	December 28

Tentative Course Plan

These are the main topics I am going to cover. As this class is given for the first time this list is *tentative* is definitely subject to change. Buyer be aware: I may add new topics, or remove some of these in the future.

1. Differences between data science, statistics and machine learning
2. A crash course in python
3. A crash course in R
4. Best practices in data science
5. Basic statistical concepts
6. Univariate and multivariate numerical data problems
7. Boolean decision problems
8. Unsupervised clustering problems
9. Supervised classification problems
10. Neural network models