Lecture 1

Introduction to CMP713

Assoc. Prof. Dr. Burkay Genç

19 Feb, 2024

Who am I?

- I am Assoc. Prof. Dr. Burkay Genç
- BSc Industrial Engineering, Bilkent Univ.
- MSc Computer Engineering, Bilkent Univ.
- PhD Computer Science, Waterloo Univ. Canada
- Izmir Economy University
- TED University
- Hacettepe University, Inst. of Population Studies
- Hacettepe University, Computer Engineering Dept.
    - Data mining, Algorithms, Discrete Math
- Hacettepe University, Informatics Institute
    - Game Technologies

What is this course about?

- We will learn the **practice** of data mining
    - This is not a machine learning course (see CMP712)
    - This is not a neural networks course (see CMP684)
    - This is not a deep learning course (see CMP784)
    - This is not a text mining course (see CMP614)
    - This is not a graph mining course (see CMP615)
    - But, we will talk about all of them, briefly.
- We will use the **R language**

What is data mining?

- Gartner Group definition:

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

What will you learn?

1. Develop understanding of application, goals
2. Create dataset for study
3. *Data Cleaning and Preprocessing*
4. *Data Reduction and projection*
5. *Choose Data Mining task*
6. *Choose Data Mining algorithms*
7. *Use algorithms to perform task*
8. *Evaluate and iterate through 3-7 if necessary*
9. *Reporting*
10. Deploy: integrate into operational systems

Data Cleaning and Preprocessing

- Data collected from the field has problems
  - Missing data
  - Anomalies
  - Repetitions
  - Fake values
- Also, contains data in unexpected, useless forms
  - Reformat
  - Feature engineering

Data Reduction and projection

- Sometimes data contains too much information that bloats the analysis
- You need to find less useful parts of data and reduce its size
    - Dimensionality reduction
        - PCA
        - Remove features
- You may also have too many examples
    - Too many examples mean slower analysis
    - Find an eliminate repetitive examples

Choose Data Mining task

- Classification vs Regression
  - Classification is for data where the target is discrete
    - Binary if two
    - Else
  - Regression is for data where the target is continuous
    - Discrete data with many classes can also be considered as continuous
- Applying the wrong task can completely ruin your analysis

Choose Data Mining algorithms

- Classification
  - Decision Tree, Random Forest, SVM, Boosting, ANN, kNN, Bayes
  - Logistic regression
- Regression
  - All of the above
  - Linear, multiple, hedonic, lasso, piecewise linear, …

Use algorithms to perform task

- Build a model
  - Build more models
- Tune parameters
- Find best combinations

Evaluate and iterate

- Check if the results are pleasing
- Are they speaking wisdom?
- Can you develop policies based on the results?
- Are the policies applicable?
- If not, iterate from the beginning

Reporting

- When all looks good, you have to report your results
- Be concise
  - Short
  - But, full
- Be visual
  - Explain with plots, not words
- Know your audiance
  - Usually you don’t present to scientists and data engineers
    - They won’t be interested in which technique you used to optimize models

What is required?

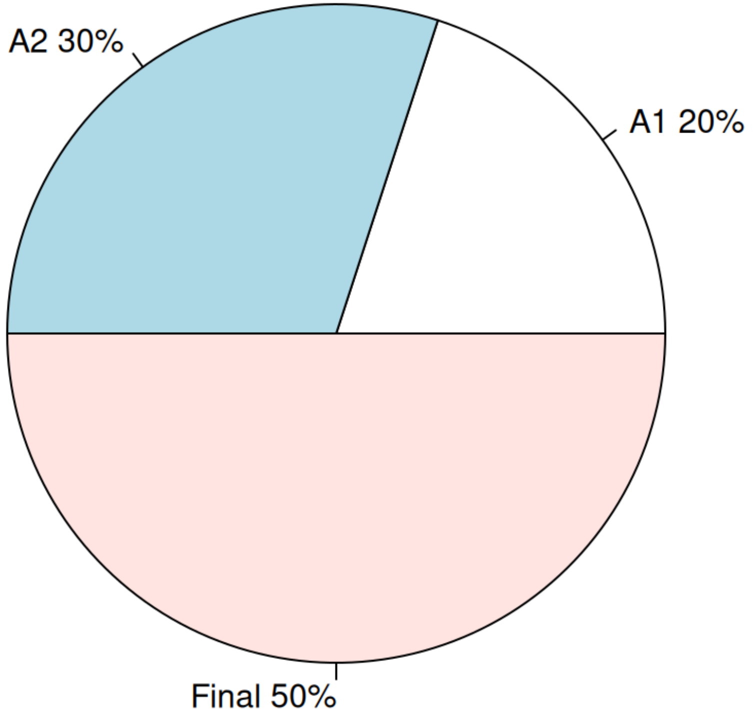
- A modern PC
- R programming language installation
    - CRAN
- RStudio IDE installation
    - RStudio
- Both are free software, available for all OSs.

How the course will be taught?

- I will follow slides
- I will show live examples
- I will ask you to do small tasks

Grading (may be subject to changes)

- 20% : First Assignment (Data preprocessing/cleaning)
- 30% : Second Assignment (Exploratory Data Analysis)
- 50% : Final Assignment (Modeling, Evaluation, Deployment)



Grading (may be subject to changes)

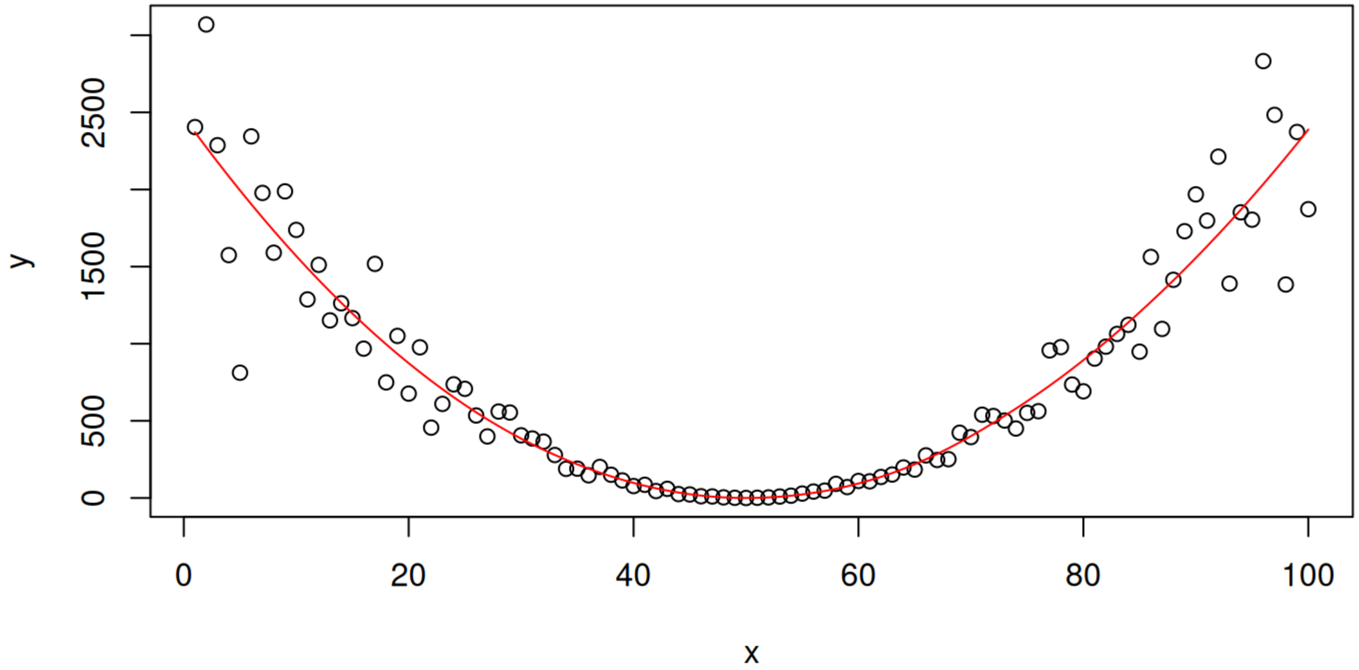
- 0-69: F3
- 70-73: B3
- 74-77: B2
- 78-81: B1
- 82-85: A3
- 86-89: A2
- 90+ : A1

Task

- Install R
- Install R-Studio
- Define two vectors `x` and `y` from 1 to 100
- Subtract 50 from `y` and take squares
- Define a third vector `z` of 100 values using normal distribution with mean 0, sd `y/5`.
- Add `z` to `y` and store in `y`
- Plot `x` vs `y`
- Draw a polynomial regression on the same plot

Task

```
x <- 1:100
y <- ((1:100)-50)^2
z <- rnorm(100, 0, y/5)
y <- y + z
plot(x, y)
l <- loess(y~x, degree=2)
lines(x, l$fitted, col="red")
```



Until next week

- Start learning R!
- Become a useR!
-