Lesson 01 Intro to Data Mining and Data

Lusine Zilfimian

February 10 (Monday), 2020

Welcome to the world of Data Mining!

- Syllabus Highlights and Info about the Course
- Intro to Data and Data Types
- Exploring Data: Summary Statistics
- Exploring Data: Visualization
- Ungraded Quiz

Syllabus highlights

- Course name: **Data Mining**
- Number of Credits: **3**
- Github Page: **click on me**
- Syllabus: **uploaded to our Github page**
- Textbooks: **let me know if you can not find/download them**
- Language and Software: **R** and **R Studio**
- R Textbooks: **see in syllabus**
- Question: **Do we need lab sessions for R, R Markdown, R Shiny, and Github? (think about it).**

Syllabus more important highlights

- HW: **(almost) weekly**
- Exams: **Midterm + Final Exam**
- Final Project: **the guideline will be uploaded on Github**
- Quizzes: **:)**
- Grading policy:

$$Final\ Grade = 0.2(HW + ME + FP) + 0.3FE + 0.1Q$$

Based on my previous experience...

- No Makeups for Quizzes.
- No late HWs: I respect your time and expect the same from you.
- Cheating: Be honest! Any similarities, which can be considered as cheated, will not be graded.
- Feel free to ask questions and have comments.

Questions?

Intro to Data Mining

What is DM? What is the difference between DM and Statistics?

- Data mining is the process of automatically discovering useful information in large data repositories.
- **Not all information** discovery tasks are considered to be data mining.

The process of converting raw data into useful information:

- Data Cleaning
- Data Integration and Selection
- Data Transformation
- Data Mining Algorithms
- Evaluation (+ Interpretation)
- Visualization

Example of using DM in

- Bioinformatics
- Marketing
- Macroeconomics
- Education

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

- Predictive task
- Descriptive task

There are two types of predictive modeling tasks:

- Classification
- Regression

Course Structure: Topics at a glance

- Types of Data, Data Preprocessing
- Exploring Data
- Linear Regression
- Logistic Regression
- Poisson Regression
- Regularization
- Classification
- Cluster Analysis
- Dimensionality Reduction
- Hmm... other interesting topics

Types of Data, Data Preprocessing

What are an attribute and a measurement scale?

- **An attribute** is a property or characteristic of an object that may vary; either from one object to another or from one time to another.
- **A measurement scale** is a rule (function) that associates a numerical or symbolic value with an attribute of an object.

Types of Attributes

We **CAN** define 4 types of attributes:

- Nominal
- Ordinal
- Interval
- Ratio

The definition of the attribute types is cumulative.

Types of Data Sets

- **Record data**

Long format:

```
##         Name  HW Grade
## 1     Lusine HW1    15
## 2      David HW1    16
## 3 Shoghakat HW1    17
## 4     Lusine HW2    18
## 5      David HW2    19
## 6 Shoghakat HW2    20
## 7     Lusine HW3    18
## 8      David HW3    17
## 9 Shoghakat HW3    20
```

Types of Data Sets

Wide format:

```
##         Name HW.1 HW.2 HW.3
## 1     Lusine   15   18   18
## 2      David   16   19   17
## 3 Shoghakat   17   20   20
```

Transaction data

```
##   ID                  Items
## 1  1     Lays, Coca-Cola
## 2  2 Lays, Beer, Sprite
## 3  3     Chocolate, Milk
```

Document-term Matrix

```
##   Document math is life
## 1       D1    1  2    3
## 2       D2    4  5    6
## 3       D3    0  7    8
```

- etc.

Data Quality

- Noise and Outlier
- Missing Value
- Inconsistent Value
- Duplicated data and Deduplication

Data Preprocessing (see Lesson 2)

- Aggregation: to reduce the memory and provide high-level view
- Sampling (types)
- Feature subset selection (Redundant and Irrelevant features)
- Feature creation
- Discretization and binarization
- Variable transformation (Normalization or Standardization)