Data Mining

Johnny Nguyen[1]

***Abstract*— This electronic document permits me to synthesis the first course of data mining.**

I. INTRODUCTION

This document explains what happened in the first course of Data Mining. We know the three ways to do a PhD :

- EDSTIC, subject on may
- CIFRE, find a supervisor on the academic, resume, subject, 3 monthers to get an answer, three months to improve later.
- PROJECTS, three students selected

Why doing pHd ? You can be in a good company (Google, Amazon) and get a nice job.

II. STATISTICS

The current challenges are Ultra high dimensions. The goal is to focus on statistics and not on algorithm. Data scientist exist to explore data on every company.

Data exploitation : analysis large amount of data in order to discover pattern and significant rules using Machine Learning.

This domain is ultra-competitive. Big data was used in 2013. It permits us to sell data.

A. Family

Statistic, Data warehouse, data visualization.
The scientific activities consists in retrieve information in large volume of multimedia.

B. OLAP

OLAP consists into high dimension data. We can make prediction on Data Mining but not in OLAP.

C. Process of knowledge extraction

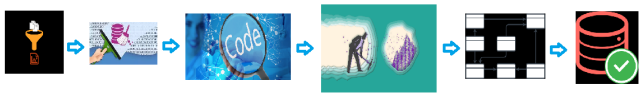

Fig. 1: Process of knowledge extraction

From a **data warehouse**, we need to **select** the data that we will **clean** before **transforming**. Then, we **mine** the data, to choose the correct **model** and **validate** it to extract **knowledge**.

To be sure that our **knowledge** is **efficient** we can **change each step** regarding the **validation**.
Hopefully, we don't want to change data.
1) **Set down the problem** : goals of the end user (query or prediction)
2) **Data selection** : parse the file, extract data into columns
3) **Data cleansing** : outliers, exclude incomplete records (fuzzy logic)
4) **Data coding** : transform data smartly (change type)
5) **Model** : unsupervised and supervised
6) **Validation** : quantitative assessment, standard deviation
7) **Knowledge integration**

*This work was not supported by any organization

III. DATA MINING

Algorithm to classify information.

- **Supervised** : predict one number or classes using training that we know the result.
- **Unsupervised**, we don't have the answer, we look for correlation. (Anti-spam)

The classification is a curve that slice data into different classes.

A. Five myths

- To prevent the future, like a crystal ball,
- Not yet viable for professional applications,
- Requires a separate and dedicated database,
- Polytech nicean is fine to make data mining,
- Dedicated for large companies that have a large volume of data client.

B. Expert system

We have a **need** that was **retrieve** and we **extract the knowledge** to build the system. The **computer scientist** will **build** this **system**.

C. Data mining vs expert system

Rules are created by an expert system. Whereas **data mining** is used to **extract strategies**.

D. Data science stack

Visualization (Kibana) Analysis (Scikit-learn) Storage/Access/Exploitation (Hadoop) Infrastructure (HPC)

We just need to know the idea of all layer to become expert in two consecutives ones. In our case, we will focus on Visualization and analysis.

IV. CONCLUSIONS

We must use R.

ACKNOWLEDGMENT

Thanks to Fredreric Precioso for his work.

REFERENCES

[1] https://www.quora.com/What-steps-should-be-included-in-a-data-cleansing-process
[2] http://apro-outsourcing.com/services/custom-solutions-with-big-data/data-preparation
[3] http://crosnt.com/medical-coding-data-management-manual-vs-auto-coding/
[4] https://lab.getapp.com/what-is-data-mining-small-business/
[5] https://www.lucidchart.com/pages/database-diagram/database-models
[6] https://corevalue.net/data-validation-testing/