Data Mining Project

Work Summary

Autor

Machine Learning
Master's degree in Informatics and Computing Engineering

Porto, 2021

Conteúdo

Capítulo 1

Business Understanding

- Analysis of requirements with the end user.
- Definition of business goals.
- Translation of business goals into data mining goals.

Capítulo 2

Data Understanding

- Diversity of statistical methods.
- Complexity of statistical methods.
- Interpretation of results of statistical methods.
- Knowledge extraction from results of statistical methods.
- Diversity of plots.
- Complexity of plots.
- Presentation.
- Interpretation of plots.
- Visual knowledge extraction.

Capítulo 3

Data Processing

- Data integration.
- Assessment of dimensions of data quality.
- Cleaning redundancy.
- Cleaning missing data.
- Cleaning outliers.
- Data transformation for compatibility with algorithms.
- Feature engineering from tabular data.
- Sampling for domain-specific purposes.
- Sampling for development.
- Imbalanced data.
- Feature selection.

Capítulo 4

Descriptive

- Diversity of algorithms.
- Parameter tunning.
- Understanding algorithm behaviour.
- Performance measure.
- Correct interpretation of performance measures.
- Comparative analysis of results.
- Model improvement.
- Analysis of resuls.
- Diversity of tasks.
- Diversity of algorithms.

Capítulo 5

Predictive

- Parameter tunning.
- Undertanding algorithm behavior.
- Performance estimation: training vs test.
- Performance estimation: other factors (time, ...).
- Performance estimation: performance measure.
- Performance estimation: correct interpretation of performance measures.
- Performance estimation: analysis of results.
- Model improvement.
- Feature importance.
- Analysis of "white-box"models.

Capítulo 6

Project

- Management methodology.
- Management plan.
- Project management tools.
- Collaboration tools.

We are using Github for version control and collaboration. We regularly do pair programming using Visual Studio Code live sharing feature.

Capítulo 7

Tools

- Analytics.
- Database.
- Other tools (data cleaning, visualization).

Capítulo 8

Presentation

- Quality of layout.
- Quality of content in slides.
- Delivery.
- Use of time.