UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data and Text Mining (blended)

2021-2-F1801Q105

Aims

To train the *expert of knowledge extraction from structured, un-structured and semi-structured data* according to the *data and text mining methodology*.
The goal is achieved by;

- teaching how to *design*, *develop* and *present* data mining and text mining projects,
- introducing the main learning algorithms and models for structured, un-structured and semi-structured data,
- exploiting *open source* platforms, languages and software,
- stimulating the *team working* methodology.

The student will be able to *design*, *develop*, *document*, and *present* data and text mining projects *solving real world problems*.

Contents

The course contents are the following;

- **Data Exploration** to inspect and summarize the available data and to design and develop a pre-processing workflow,
- **Classification**, to learn a mapping from input attributes to output or target attributes to be classified or predicted,

- **Clustering**, to form homogeneous groups of observations and/or attributes using a given proximity measure,

- **Association Rules**, to automatically extract rules hidden in the data with specific reference to transaction data.

- **Text Preprocessing**, to transform un-structured and semi-structured data to be processed by learning algorithms.
- **Text Classification**, to learn classifying social networks posts, news, ...
- **Topic Models**, to automatically extract hidden topics from textual sources.
- **Information Extraction**, to automatically extract entities, i.e. person, place, organization, ... and their relationships from un-structured and semi-structured data.

You will learn how to develop data and text mining workflows using the **KNIME open source software platform**. You are *not required to code any programs* while if you want KNIME allows to use powerful and professional open source programming languages and commercial software environments; R, Weka, Matlab, Python, Java, ...

Detailed program

- **Data Exploration and Preprocessing**

    - Data types and attributes
    - Graphical and tabular data exploration
    - Missing data treatment
    - Data Pre-Processing

- **Classification**

    - Introduction
    - Techniques, models and algorithms; artficial neural nets, Bayesian classifiers, decision trees, …
    - Performance measures to evaluate and compare classifiers
    - Unbalanced classes and non binary classification
- **Clustering**

    - Introduction
    - Proximity measures for nominal, ordinal and continuous attributes
    - Techniques, models and algorithms; partitioning, hierarchical, graph based, density based, …
    - Performance measures to evaluate and compare clustering solutions

- **Association Rules**

    - Introduction and basic definitions
    - Item and itemsets
    - Apriori, principle and algorithm
    - Performance measures to evaluate and compare association rules
- **Text Preprocessing**

    - Tokenization

    - Filtering and Stemming

    - the bag-of-words model, 0/1, term frequency
    - Term frequency inverse document frequency
- **Text Categorization**
    - binary classification
    - multi-class
    - multi-label
- **Topic Models**
    - Document clustering
    - Topic Models
    - Latent Dirichlet Allocation
    - Topic validation
- **Information Extraction**
    - Entity extraction
    - Entity relationship extraction
    - Sequence prediction
    - Industrial and commercial applications
- **Deep Learning**
    - Introduction
    - Feedforward neural network
    - Basics on Convolutional neural networks
    - Basics on Sequential neural networks

Prerequisites

Basic knowledge on; informatics, probability calculus and statistics.

Teaching form

Teaching happens in blended learning, while tutorial lectures will happen in classes. The entire course is also available in *digital* form consisting of *video lectures* for theory and *hands-on. All videos are in English.* The course material is organized through *learning paths* where *lecture modules* consist of theoretical lecture, hand-on lecture

and *self-evaluation sessions*. Self-evaluation session offers a powerful and effective resource to *online learning*, i.e. after the class has taken place, The course makes available 230 quizzes to allow students to fairly assess their understanding level and to train for the exam.

Textbook and teaching resource

Audiovisual, slides, dataset and workflow designed and implemented by the course teacher and instructor.
Furthermore, the following books are recommended

- http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf
- http://dsd.future-lab.cn/members/2015nlp/Machine_Learning.pdf
- https://www.researchgate.net/file.PostFileLoader.html?id=526bc9cfd3df3efa3ec519ee&assetKey=AS%3A272156041121792%401441898465154
- http://www.springer.com/us/book/9781447125655
- https://link.springer.com/book/10.1007%2F978-1-4020-4993-4
- http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470276800.html

Semester

Fall Semester

Assessment method

Assessment is based on two components, a *Data and/or Text Mining project* and a *methodology exam* which is performed in the laboratory by using a computer. Students and encouraged to *work in small teams* to design, develop and document their data and/or text mining project. The data and/or text mining project is usually *selected by the students team* by exploiting the *Kaggle platform* (**https://www.kaggle.com/**) where *Data Science* requests and offers meet.
The *machine learning project* gives a maximum of 21 points, assigned according to six criteria as follows:

- Technical merit: notably rigour, accuracy and correctness (maximum 5 points)
- Clarity of expression and communication of ideas; including readability and discussion of concepts (maximum 5 points)
- Appropriate referencing and the context of the present work (maximum 2 points)
- Overall balance and structure of report (maximum 3 points)
- Repetition; have significant parts of the manuscript already been published by other authors? (maximum 3 points)

- Diagrams, tables, captions; are they clear and essential (maximum 3 points)

The *methodology exam* gives a maximum of 11 points. according to the following; 6 points for 6 quizzes, one point for each quiz (*each quiz concerns concepts presented in the course*) and a maximum of 5 points for an open ended question having the goal to evaluate the *critical point of view of the candidate*. The candidate can also ask to undergo *oral examination*, consisting of questions about *reasoning and deduction about the concepts presented in the course*, which gives a maximum of 3 points.

Office hours

On dating