MSc in Bioinformatics for Health Sciences

DMI. Data Mining and Data Integration in Biomedicine

Syllabus Information

Academic Course: 2019/20

Academic Center: 804 - Official Postgraduate Programme in Biomedicine

Study: 8045 – Bioinformatics for Health Sciences - MSc

Subject: 32548 - DMI Data Mining and Data Integration in Biomedicine

Credits: 5.0

Course: 1st

Teaching languages: English

Teachers: Janet Piñero

Teaching Period: 1st term

Presentation

This course will cover a selection of commonly employed data mining techniques and will also cover the steps of data pre-processing. Given the importance of data integration to understand the complexity of biological systems, the course will also include examples of data integration. These topics will be illustrated using real case examples, such as mining scientific publication data, electronic health records data, and drug response data.

The course comprises 5 ECTS credits, implying plenary lectures, and hands-on sessions (please bring your own laptop for the hands-on sessions).

Associated skills

General competences:

1. Learning different strategies for preprocessing, integrating and visualizing different types of biomedical data

2. Developing abilities to reading, writing and listening scientific English related to the subject.

3. Communicating scientific research by means of presentations.

Specific competences:

Developing R programming skills

Performing exploratory data analysis for different types of biomedical data, and choosing the most suitable visualization.

Deciding the most appropriate data mining methods to use in different knowledge discovery in biological data.

Learning strategies to integrate heterogeneous types of biological data.

Understanding the uses and limitations the data mining algorithms employed

Learning outcomes

By taking this course, the students will be able to understand and apply different data mining algorithms to real-world problems.

Contents

Contents section 1: Introduction to Data mining

1.1 What is Data mining

Contents section 2: Data Analysis and Integration

2.1 Data Understanding and Preparation

2.2 Anomaly Detection

2.3 Data transformations and dimensionality reduction

2.4 Ontologies and vocabularies and examples of ontologies: the disease ontology

Contents section 3: Cluster analysis

3.1 Distance measures

3.2 Clustering algorithms: K-means, hierarchical clustering, partitioned-based clustering, and fuzzy clustering.

3.3 Mining graph data: clustering protein-protein interaction data

Contents section 4: Data Classification

4.1 Introduction to machine learning

4.2 Mining text data using Rule based approaches

4.3 Mining text data using machine learning approaches

4.4 Introduction to Deep Learning

4.5 Evaluation and validation of data mining results

Contents section 5: Visualization

5.1 Exploring techniques to visualize different types data in R

Teaching methods

The course is mainly a practical course, and each topic is organized around a brief lecture to present the basic concepts behind each topic, followed by hands-on exercises using publicly available resources. At the end of some of the sessions, the student will be asked to deliver the answer of specific exercises. In addition, each student will present a short seminar of 20-40 minutes of a scientific article of relevance for the field.

Evaluation

The evaluation will consist of a final exam at the end of the course, representing 60%, the evaluation of the exercises performed during the course (25%) and the presentation of the scientific article (15%).

Grading system:

A minimum performance of 50% on each item is required to pass the subject.

Bibliography and Information Resources

Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison Wesley.