

SYLLABUS DEL CORSO

Unsupervised Learning

2526-1-F9103Q045-F9103Q04502

Obiettivi

To develop skills for solving real world unsupervised learning problems.

The goal is achieved by;

- Teaching how to design, train, deploy and monitor unsupervised learning models. **(DdD 1, DdD 2)**
- Exploiting open source platforms, languages and software, **(DdD 1, DdD 2)**
- Stimulating team working. **(DdD 4)**
- Reading and discussing scientific papers made available by the teacher **(DdD 3, DdD 5)**

Contenuti sintetici

The course contents are the following;

- **Data Types**; to list different types of data and to learn how they must be used for unsupervised learning.
- **Data Preprocessing**; to preprocess data in such a way it can be used by unsupervised learning tasks,
- **Clustering Learning**; to form homogeneous groups of observations and/or attributes using a given proximity measure,
- **Clustering Validation**; to evaluate and compare different clustering solutions to select the one to deploy.
- **Anomaly Detection**; to find anomalous observations, to discover outliers observations, under different theoretical settings.
- **Bayesian Networks**; to learn probabilistic/causal structure from data and to make decisions under uncertainty.

You will learn how to design, train, validate and deploy unsupervised learning models using Python.

Programma esteso

1. Data

- 1.1 Data types and attributes
- 1.2 Proximity measures for nominal, ordinal and continuous attributes
- 1.3 Data Pre-Processing

2. Cluster Analysis

- 2.1 Introduction
- 2.2 Clustering algorithms
 - 2.2.1 *Partitioning*
 - 2.2.2 *Hierarchical*
 - 2.2.3 *Graph-based*
 - 2.2.4 *Density-based*
 - 2.2.5 *Time-series*
- 2.3 Comparing clustering solutions
 - 2.3.1 *Performance measures*
 - 2.3.2 *Evaluation*
 - 2.3.3 *Comparison*

3. Anomaly Detection

- 3.1 Introduction
- 3.2 Anomaly detection algorithms
 - 3.2.1 *Statistical approaches*
 - 3.2.2 *Proximity-based approaches*
 - 3.2.3 *Clustering-based approaches*
 - 3.2.4 *One-class classification*
 - 3.2.5 *Information theoretic approaches*

4. Bayesian Networks

- 4.1 Introduction
- 4.2 Bayesian network models
 - 4.2.1 *Discrete variables*
 - 4.2.2 *Continuous variables*
 - 4.2.3 *Mixed variables*
- 4.3 Learning
 - 4.3.1 *Parameters*
 - 4.3.2 *Structure*
- 4.4 Inference
 - 4.4.1 *Exact*
 - 4.4.2 *Approximate*

Prerequisiti

Conoscenza base di: calcolo delle probabilità, statistica, matematica.
capacità di progettare e implementare progetti software

Modalità didattica

Il corso è organizzato come segue:

- 16 lezioni da 2 ore di teoria di natura erogativa in presenza
- 12 lezioni da 2 ore di esercitazione di natura interattiva in presenza

Materiale didattico

- **Introduction to Data Mining** (<https://www-users.cse.umn.edu/~kumar001/dmbook/index.php>)
- **Bayesian Networks and Decision Graphs** (<https://link.springer.com/book/10.1007/978-0-387-68282-2>)

Periodo di erogazione dell'insegnamento

Primavera

Modalità di verifica del profitto e valutazione

L'esame è strutturato come segue:

- **Project work**; Lo studente è invitato a sviluppare e/o ad applicare uno o più algoritmi per analizzare un caso di studio assegnato dal docente. **(Assegna un massimo di 13 punti)**.
- **Colloquio sulla relazione di laboratorio**; Sui temi presentati a lezione e collegati al project work **(assegna un massimo di 20 punti)**.

Non sono previste prove intermedie

Orario di ricevimento

Da concordare inviando una mail a fabio.stella@unimib.it

Sustainable Development Goals
