CORSO: Unsupervised Learning

DOCENTI: Alessio Farcomeni (Ph.D., 2004); Marco Stefanucci (Ph.D, 2018)
EMAIL: alessio.farcomeni@uniroma2.it
marco.stefanucci@uniroma2.it
PAGINE WEB: https://economia.uniroma2.it/faculty/563/farcomeni-alessio
https://economia.uniroma2.it/faculty/725/stefanucci-marco

DESCRIZIONE DEL CORSO

Il corso discute delle principali tecniche statistiche per identificare gruppi nei dati (ovvero, identificare strutture discrete latenti, non direttamente osservate od osservabili) e, anche quando vi sia un solo gruppo, valori anomali. Il corso discute inoltre la riduzione della dimensionalità dei dati a scopi descrittivi, di costruzione di indici e ranking da osservazioni multivariate, sintesi e interpretazione dell'informazione. Principi di stima robusta verranno introdotti, e conseguentemente le varianti formalmente robuste delle tecniche discusse precedentemente.
Come esempio, si consideri il database dei clienti di una certa azienda, dove per ogni cliente (unità) vengono misurate una serie di caratteristiche (variabili) associate al comportamento e preferenze dei consumatori: numero di visite, ammontare speso per unità di tempo, soddisfazione in termini di voto su diversi aspetti del servizio, ecc. Le tecniche di classificazione non supervisionato ci permettono di rispondere a domande come: ci sono tipi distinti ti consumatori? Quanti sono, e quali sono I loro profili? Ci sono alcuni consumatori inusuali? Come possiamo costruire un ranking dei consumatori rispetto alla loro propensione a far business con noi? Come possiamo presentare l'intero data set in un grafico? Quale informazione è disponibile nei dati a disposizione?

OBIETTIVI DI APPRENDIMENTO

- ✓ abilità ad usare tecniche di apprendimento statistico in presenza di etichette non osservate
- ✓ abilità di identificare anomalie
- ✓ abilità di identificare gruppi e assegnare ciascuna osservazione ad un gruppo di unità simili
- ✓ abilità di ridurre la dimensionalità dei dati con minima perdita di informazione

METODOLOGIA

L'enfasi è sui principi e sulle tecniche statistiche specifiche. Ciascun metodo è introdotto tramite esempi e approfondito da un punto di vista tecnico. Una base di statistica matematica è necessaria, ma le derivazioni verranno ridotte al minimo indispensabile. Le metodologie sono discusse da un punto di vista teorico e pratico, con forte enfasi sulla parte pratica.
Vengono descritte le definizioni, assunzioni, proprietà, implementazione, e interpretazione di ciascuna metodologia. L'intero corso è basato sul software *R.*

VALUTAZIONE

Esame scritto, basato su domande chiuse (con la possibilità di sporadiche domande aperte). L'esame verterà sugli aspetti di specificazione e interpretativi delle metodologie discusse. Alcune domande riporteranno anche codice o output ottenuto dal software *R*, su cui verterà lo specifico quiz.

PROGRAMMA

- 1. Introduzione ed overview
  - 1.1 Subsampling
- 2. Metodi di segmentazione non gerarchica
  - 2.1 K-means
  - 2.2 PAM, Clara
  - 2.3 Misture di distribuzioni Gaussiane multivariate: Mclust
  - 2.4 Metodi robusti
    - 2.4.1 Trimmed K-means
    - 2.4.2 tclust
  - 2.5 Anomaly detection
- 3. Metodi di segmentazione gerarchica
  - 3.1 Single linkage e le sue proprietà
  - 3.2 Altri metodi di linkage
- 4 Riduzione Dimensionale
  - 4.1 Analisi delle Componenti Principali
  - 4.2 Analisi delle Componenti Principali robusta
  - 4.3 Cenni all'Analisi delle Componenti Principali sparsa

Se il tempo lo permette, possono essere discussi altri argomenti aggiuntivi o propedeutici alla comprensione dei contenuti del corso.

LIBRI DI TESTO

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, Springer Series in Statistics. https://web.stanford.edu/~hastie/ElemStatLearn/

Farcomeni, A. and Greco, L. (2015) Robust Methods for Data Reduction, Chapman & Hall/CRC Press

LETTURE SUGGERITE

Chatfield, C. and Collins, A. J. (1981) Introduction to Multivariate Analysis, Chapman & Hall/CRC Press

Witten J.D., Hastie T., Tibshirani R. (2014). An Introduction to Statistical Learning with Applications in R. Springer, Springer Series in Statistics.