

# Corso di Programmazione 3

## Progetto Esame

Docente: Angelo Ciaramella

A.A. 2015/2016

### Studente

- **Cognome:** Accardo
- **Nome:** Giuseppe
- **Matricola:** 124/879

### Traccia - Data Mining

Si vuole sviluppare un sistema di *Data Mining* per l'elaborazione di dati. Il *Data Mining* comprende un insieme di tecniche e metodologie che hanno per oggetto l'estrazione e visualizzazione di informazioni da grandi quantità di dati (vedi sotto). Le fasi principali del sistema che si vuole sviluppare sono: *selezione delle caratteristiche*, *clustering* e *visualizzazione*.

Si suppone di avere un data set contenuto in un file `.data` (vedi sotto). Il data set è relativo alla classificazione di 3 tipi di rose (*setosa*, *versicolour*, *virginica*) mediante sue quattro caratteristiche (*feature*): lunghezza e larghezza del *sepal*, lunghezza e larghezza del *petal*.

Nella fase di *selezione delle caratteristiche* un utente può scegliere il numero di caratteristiche da selezionare per l'analisi (le colonne del data set).

Nella fase di *clustering* viene usato un algoritmo per "agglomerare" dati simili. Nel caso specifico viene usato l'algoritmo *K-Means* (vedi sotto). L'utente può scegliere il numero di "cluster" da usare.

Nella fase di *visualizzazione* i dati "agglomerati" sono visualizzati in 2 e 3 dimensioni. Nel caso in cui il numero delle feature è più grande di 3 viene applicato un algoritmo di Analisi delle Componenti Principali (vedi sotto) per la visualizzazione.

### Data Mining

Per maggiori dettagli vedere: [https://it.wikipedia.org/wiki/Data\\_mining](https://it.wikipedia.org/wiki/Data_mining)

## Data set

Il data set è disponibile a questo link:

<http://mlr.cs.umass.edu/ml/machine-learning-databases/iris/>

Considerare le prime quattro colonne del data set `iris.data`

## Clustering

Il *clustering* o analisi dei gruppi è un insieme di tecniche di analisi dei dati volta alla selezione e raggruppamento di elementi omogenei in un insieme di dati.

In questo contesto usare l'algoritmo *K-Means* ([http://www.saedsayad.com/clustering\\_kmeans.htm](http://www.saedsayad.com/clustering_kmeans.htm)).

## Principal Component Analysis

E' possibile usare delle librerie che implementa la PCA. Per maggiori dettagli vedere:

<https://java.net/projects/cougarsquared/sources/svn/content/trunk/core/libraries/flanagan/analysis/PCA.java?rev=2236>

## Note di sviluppo

La prova d'esame richiede la progettazione e lo sviluppo della traccia proposta. Il progetto deve essere sviluppato secondo le seguenti linee:

- usare almeno un pattern tra i **design pattern** noti;
- attenersi ai principi della programmazione **SOLID**;
- usare il linguaggio **Java**;
- inserire sufficienti **commenti** (anche per Javadoc) e **annotazioni**;
- usare l'**interfaccia grafica** con **gestione degli eventi**;
- gestione delle eventuali **eccezioni**;
- usare i **file** o **database**.

Lo studente deve presentare una relazione sintetica (per chi usa *latex* è possibile scaricare un template dalla piattaforma e-learning). La relazione deve contenere:

- una breve descrizione dei **requisiti** del progetto;
- il **diagramma UML** delle classi;
- altri diagrammi se opportuni;
- parti rilevanti del **codice** sviluppato.

## Consegna progetto

La relazione deve essere inviata al docente per e-mail `angelo.ciaramella@uniparthenope.it` entro la **data di scadenza della prenotazione on-line** dell'esame e deve essere inviata esclusivamente dall'indirizzo e-mail personale dello studente `nome.cognome@studenti.uniparthenope.it`.

Il codice Java del progetto deve essere consegnato contestualmente alla relazione secondo le modalità ritenute più opportune (e-mail, Dropbox, Piattaforma Sebeto, Pendrive, CD, ...).

## Modalità di esame

La prima parte della prova di esame verterà sulla discussione del progetto. Lo studente deve preparare una **presentazione sintetica** (slide) per descrivere il progetto svolto. La seconda parte della prova verterà sulla discussione degli argomenti affrontati a lezione.