Universidade do Minho
Departamento de Informática
Mestrado [Integrado] em Engenharia Informática
Mestrado em Matemática e Computação

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2024/2025

Practical Exercise no. 9

Theme

Clustering: K-means, K-medoids and DBSCAN

Exercise

Unsupervised learning is essentially used to obtain inferences from data sets without human intervention, in contrast to supervised learning, in which the labels are provided together with the data. The three techniques applied in this context are K-means, K-medoids and DBSCAN. The first two algorithms aim to group a set of unlabelled case studies according to the similarity of their characteristics. However, while K-means tries to minimise the distances within the cluster, K-medoids tries to minimise the sum of the distances between each point and the 'medoid' of the respective cluster. On the other hand, DBSCAN algorithm finds the core samples of high density and expands clusters from them.

The dataset has 777 observations and 18 columns:

- **Private** - A factor with levels No and Yes indicating private or public university
- **Apps** - Number of applications received
- **Accept** - Number of applications accepted
- **Enroll** - Number of new students enrolled
- **Top10perc** - % of new students from top 10% of H.S. class
- **Top25perc** - % of new students from top 25% of H.S. class
- **F.Undergrad** - Number of full-time undergraduates
- **P.Undergrad** - Number of part-time undergraduates
- **Outstate** - Out-of-state tuition
- **Room.Board** - Room and board costs
- **Books** - Estimated book costs
- **Personal** - Estimated personal spending
- **PhD** - % of faculty with Ph.D.'s
- **Terminal** - % of faculty with terminal degree
- **S.F.Ratio** - Student/faculty ratio
- **perc.alumni** - % of alumni who donate
- **Expend** - Instructional expenditure per student
- **Grad.Rate** - Graduation rate

Tasks

This practical exercise is intended to group a set of universities into two groups: private institutes or public institutes. Given the characteristics presented, it was decided to apply a set of non-supervised models, specifically K-means, K-medoids and DBSCAN, as a way of solving this binary classification problem. It includes the following tasks:

T1. Load the dataset and apply methods for data exploration and visualization;

T2. Train three unsupervised learning models using K-means clustering (*sklearn.cluster.KMeans*), K-medoids (*sklearn_extra.cluster.KMedoids*) and DSCAN (*sklearn.cluster.DBSCAN*), classifying each case study as *'private institute'* or *'public institute'* ;

Note: The *'Private'* attribute indicates the labelling of each university, showing whether the university is a private institute. For training purposes, this attribute should be *removed* from the dataset.

T3. Given the value of the *'Private'* attribute, evaluate the clustering performance *of* each model by creating a confusion matrix *(sklearn.metrics.confusion_matrix(...))* and a classification report *(sklearn.metrics.classification_report(...))*;

T4. Given the results obtained in **T3**, what conclusions have you drawn? In which situations do the models get it right/fail? How can the proposed learning models be improved?