**Chapter 7: Clustering**

**Theoretical part:**

1. K-means

Read about K-means and Agglomerative clustering from

Data Mining (in books) pages 443-469 chapter 10-10.3 (not including 3.3)

Read about cluster evaluation chapter 10.6 pages 483-491

Apply semantic segmentation on the star image, try to separate it from background:

1. K - means (It is not hard to implement, if you struggle use sklearn)
2. Agglomerative clustering (there are multiple ways to solve issues you might encounter)  
   For each method use the next features:  
   • Color only  
   • Color and distance
3. Use 2 different cluster evaluation methods to evaluate your clusters
4. DB-scan

Read about the DBSCAN algorithm:

Data Mining (in books) pages 471-479 (10.4)

You will implement a variation of this algorithm to detect outliers.

Data.csv contains x (first column), y (second column) and class (third column):

1. Present the data with different colors for different classes
2. Apply dbscan algorithm on the data, present outliers in red and all other clusters in blue.

Read the next paper:

<https://bib.dbvis.de/uploadedFiles/17.pdf>

1. Implement the two algorithms presented and apply p-dbscan to find outliers now. Choose your parameters to present meaningful differences from the regular dbscan results.

**Practical part**

In this exercise you will be required to play football players in the game Fifa according to their ranking, position, action, other general data on the player. Data is at the address <https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset#players_20.csv>

Under the name players\_20.csv.

In addition, the information itself is also found in the datasets drive.

Unlike most of the cluster challenges, in this case the classes are clear: a defensive player, a Link player (center field), and an attack player (ignoring the robberies).

Therefore, the exercise will have two parts:

* In the first part, you will be required to remove the label and cluster the data. Note (!), you will still be measured by the some cluster criterion, or by the quality of the cluster **according to the customer's specific desire** in this case.
* In the second you get 15% of the labels, to check your cluster quality, and how it fits your customer's goals. (It will be **permissible** for you to run supervised methods, but if you do please consult with your instructor.)

**General Comments:**

**First, extract the labels from data for filtering purposes!**

Filter the label from the 'player\_positions' column:

The column contains all the positions where the player can play separated by a comma (if there are several). You only need to take the first one for each player (you can do this at a single line of code).

Next, take out of the data every row with label: 'GK' או NaN.

In addition, the following columns should be ignored with information throughout the exercise:

']gk\_diving', 'gk\_handling', 'gk\_kicking', 'gk\_reflexes', 'gk\_speed', 'gk\_positioning',

'goalkeeping\_handling', 'goalkeeping\_kicking', 'goalkeeping\_positioning', 'goalkeeping\_reflexes', 'goalkeeping\_diving',

'team\_position']

**Filter our the team position, as it may lead to leakage of the labels!!!**

Before starting work on the exercise, it is better to talk to the instructor about the evaluation process of the cluster, which you intend to perform.

**Part 1**

In this section, you should not use the Label column so please **put it aside**.

In addition, it is necessary to cluster only on relevant features (for example: A field sofifa\_id has no meaning in this context and should not be inserted)

In this section, you need to cluster the data in the optimal way, according to the customer's order (types of players). There is no prevention to use more than 3 clusters if you wish.

Assess how well the customer's requirements are satisfied (from an understanding of the domain).

For this part, two working days are given.

When you're done working on this part, go over the results (not yet CR) with the tutor and explain the work you've done, and how you estimate the cluster you received (again as a cluster, and also for the specific requirement).

**Part 2**

Now you can use 15% of the dataset’s labels, (set Seed to 4, using this to distribute the information in stratified form )

Before you determine the labels you see, theLabel must be mapped from the original value to the customer's goal according to the following dictionary:

position\_map = {

'RW':'ST', 'LW':'ST', 'CAM':'ST', 'ST':'ST',

'RM':'ST', 'LM':'ST', 'CF':'ST', 'LS':'ST',

'RES':'ST', 'RAM':'ST', 'RS':'ST', 'RF':'ST',

'LF':'ST', 'LAM':'ST',

'RCM':'B', 'LCB':'B', 'RCB':'B', 'LCM':'B',

'CB':'B',

'CDM':'MD', 'LDM':'MD', 'LB':'MD', 'RB':'MD',

'RDM':'MD', 'CM':'MD', 'RWB':'MD', 'LWB':'MD',

}

(You can actually copy it into code like this)

Now you need to cluster the data, given the new information added to you (for your consideration whether to use for model training/Just for testing/setting new metrics). You are welcome to consult your tutor about how to deal in such cases.

Again, at the end of the exercise, present all work done, including analyzing the quality of the cluster in the eyes of the client (the quality of the cluster is required as a cluster in addition)

Good luck 😊