**פרק 7 : אשכול**

**חלק תיאורטי:**

1. K-means

Read about K-means and Agglomerative clustering from

Data Mining (in books) pages 443-469 chapter 10-10.3 (לא כולל .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.

**חלק מעשי**

בתרגיל זה תידרש לאשכל שחקני כדורגל במשחק fifa על פי הדירוג שלהם על פי עמדה / פעולה + נתונים כלליים על השחקן. המידע ומידע אודות העמודות של המידע נמצא בכתובת <https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset#players_20.csv>

תחת השם players\_20.csv.

בנוסף, המידע עצמו נמצא גם בdrive של ההכשרה בתיקייה של התרגיל.

באופן לא אופייני לרוב אתגרי האשכול, במקרה זה הclasses הרצויים ברורים: שחקן הגנתי, שחקן קישור (מרכז שדה), ושחקן התקפי (מתעלמים מהשוערים).

לכן לתרגיל יהיו 2 חלקים:

* בראשון, תידרש לאשכל ללא label כלל. שים לב(!), אין זה אומר שלא תימדד על טיב האשכול **כאשכול**, או על טיב האשכול **בהתאם לרצון הספציפי של הלקוח** במקרה זה, גם בחלק זה של התרגיל.
* בשני תקבל 15% מהתיוגים, לבדיקת טיב האשכול שלך, ועד כמה הוא קולע למטרות הלקוח. (בחלק זה יהיה **מותר** (לא חובה) לך להפעיל שיטות supervised אם תבחר, ואם זה אכן הכיוון נכון שתתייעץ עם החונך להכוונה רלוונטית).

**הערות כלליות:**

**תחילה יש לחלץ את הlabel מהמידע, לטובת סינונים!**

האופן שבו עלייך לפרסר את הlabel מתוך העמודה 'player\_positions':

העמודה מכילה את כל העמדות בהן השחקן יכול לשחק מופרדות ע"י פסיק (אם יש כמה). עלייך לקחת את הראשונה בלבד עבור כל שחקן (ניתן לעשות זאת בשורת קוד אחת).

כעת יש להוציא מהמידע כל שורה שהlabel שלה הוא: 'GK' או NaN

בנוסף, יש להתעלם מהעמודות הבאות במידע לכל אורך התרגיל:

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

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

'team\_position']

**בדגש חזק על העמודה team\_position שעלולה להביא לleakage של הlabel האמיתי!**

לפני תחילת העבודה על התרגיל, מומלץ לדבר עם החונך על תהליך ההערכה של האשכול, אותו אתה מתכוון לבצע.

**חלק 1**

בחלק זה אין לעשות שימוש בעמודת הlabel שהוצאת **ולשים אותה בצד**.

בנוסף יש לאשכל על פיצ'רים שיהיו רלוונטיים למטרת האשכול (למשל: לשדה sofifa\_id אין משמעות בהקשר זה ואין להכניס אותו בצורתו המקורית כפרמטר לאלגוריתם clustering)

בחלק זה עלייך לאשכל את המידע בצורה הטובה ביותר, לפי הוראת הלקוח (סוגי השחקנים). אין מניעה לאשכל את המידע ליותר מ-3 אשכולים אם תרצה.

למרות שאין label נדרש ממך לעמוד על טיב האשכול אליו הגעת, ולהעריך עד כמה הוא עומד בדרישות הלקוח (מתוך הבנה של המידע).

לחלק זה מוקצבים כשני ימי עבודה.

כאשר תסיים לעבוד על חלק זה, עבור על התוצאות (עדיין לא CR) עם החונך והסבר לו את העבודה שעשית, וכיצד הערכת את האשכול שקיבלת (שוב - גם כאשכול, וגם עבור הדרישה הספציפית).

**חלק 2**

כעת ניתן להשתמש ב15% מהתיוגים של המידע, (קבע Seed ל-4 והשתמש בכך על מנת לחלק את המידע בצורה stratified על גבי התיוגים)

לפני שתקבע את התיוגים שתראה, יש למפות את הLabel מהערך המקורי למטרת הלקוח על פי המילון הבא:

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',

}

(אתה יכול ממש להעתיק אותו לקוד ככה)

כעת עלייך לאשכל את המידע בהתחשב במידע החדש שהתווסף לך (לשיקולך אם לשימוש לאימון מודל / רק לבדיקה / להגדרת מטריקה). מוזמן להתייעץ עם החונך שלך כיצד מתמודדים עם מידע במקרים כאלה.

שוב, בסוף התרגיל הצג לחונך את העבודה שנעשתה, כולל ניתוח טיב האשכול בעיניי הלקוח (נדרשת הערכה של טיב האשכול כאשכול בנוסף)

בהצלחה 😊