יהלי צופים  
נועם סטולרו  
מנחה: ד״ר עופר שיר.

קלאסטרינג

kmeans, cmeans, gap statistics

# תוכן עינינים

# מבוא:

ניתוח אשכולות הינו המשימה של קיבוץ אובייקטים לקבוצות, כך שהאובייקטים הנמצאים באותה קבוצה דומים זה לזה יותר מאשר לאובייקטים השייכים לקבוצות אחרות.

קלאסטרינג הינו הלחם והחמאה של תחום ה Data Science.

בתחומים רבים כגון: למידה חישובית, זיהוי תבניות, ניתוח תמונה, איסוף נתונים, ביו אינפורמטיקה ודחיסת מידע, יש כמות אדירה של מידע. על מנת שנוכל לעבוד עם כמות כזו של מידע, עלינו לבצע עיבוד מקדים. אחד הפתרונות הפופולריים הינו – ניתוח אשכולות באופן בלתי מונחה (unsupervised).

## מטרה:

מטרת הפרויקט, הינה מימוש יוריסטיות (אלגוריתם) אשר מציעות פתרון לבעיה המורכבת של ניתוח אשכולות, תוך כדי החלטה על מדד הדמיון לאובייקטים השונים.

את המימושים בחרנו לכתוב בשפת פייטון עקב השימוש ההולך וגובר בשפה ומשום שפייטון הינה שפת קוד פתוח ולכן הקוד שכתבנו יכול להיות לשימוש לכל אחד.

## רקע:

# מהלך העבודה:

## קשיים:

כאשר ניגשנו למטלת מימוש אלגוריתם ה-kmeans, ניצבו בפנינו הקשיים הבאים:

כיצד קובעים דימיון?

בשביל להחליט מה יהיה מדד הדימיון, יש להכיר את המידע אותו הולכים לנתח, קשה ואולי בלתי אפשרי למצוא מאפיינים אשר לפיהם ניתן לקבוע דימיון בצורה חד משמעית בין אובייטים.

למשל, בפרחים נרצה אולי להשוות בין המאפיינים הבאים:

* צבעים
* מספר עליי כותרת
* עובי גבעול
* גובה
* שורשים.

ואילו בבני אדם:

* צבע עיניים
* גובה
* משקל
* צבע עור
* מידת נעליים

לאחר שנקבע מדד הדימיון, נדרשת שיטה לקביעת שייכות לקבוצה, ישנם מספר דרכים לקבוע שייכות בין אובייקטים, לדוגמא: מרכז משותף, צפיפות. כל שיטה בעלת יתרונות וחסרונות ובהתאם ייבחר האלגוריתם. כאמור אנו נשתמש בשייכות לפי מרכז משותף, אובייקט ישויך לקבוצה אשר אובייקט המרכז שלה הוא הדומה לו ביותר, כמו כן, המדד הדימיון שלנו יהיה מרחק אוקלידי.

בעיה נוספת הנה ממד נתונים גבוה.

כאשר מימד הנתונים גבוה אי אפשר להציג את המידע באופן ויזואלי, כמו כן זמן העיבוד גדל. ייתכן גם כי ישנם מאפיינים אשר חשיבותם פחותה ורק יכולים להפריע. זו בעיה שהיקפה חורג מגבולות ולכן לא ניכנס אליה, נציין כי השתמשנו באלגוריתם ה PCA (principle component analysis).

הבעיה האחרונה אשר עלולה לעלות הנה הבעיה של קביעת מספר האשכולות.

כאשר לא ידוע דבר על הנתונים, לא ניתן לקבוע מראש כמה קבוצות קיימות לנתונים. בהמשך נציג פתרון לבעיה זו.

## kmeans:

### מימוש:

אנו רוצים למצוא k אשכולות באופן איטרטיבי על פי מרחק אוקלידי וצמצום הפונקציה הבאה:

כאשר:

למעשה הפונקציה הזו הנה סכום המרחקים הפנימיים של כל קבוצה, כמו כן אובייקט יכול לתרום אך ורק לקבוצה אליה הוא שייך.

אופן הפעולה:

* 1. הכנס k רצוי.
  2. הגרל k מרכזים באופן אקראי.
  3. שייך כל אובייקט למרכז הקרוב ביותר.
  4. חשב מחדש את כל המרכזים.
  5. חזור על צעדים 3 ו 4 עד שהמרכזים אינם משתנים.

שלב 4 של האלגוריתם מבוצע בצורה הבאה:

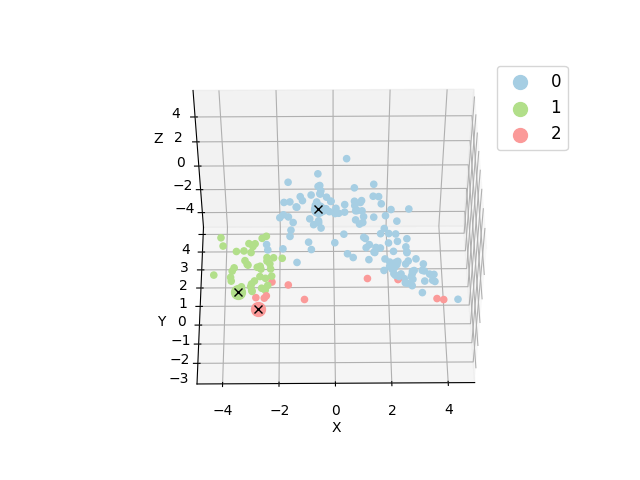
עבור כל קבוצה, נחשב את הממוצע של המאפיינים ולפיו נבחר את האובייקט הקרוב ביותר לממוצע.

### kmeans++:

בשלבים הראשונים של הפרוייקט נתקלנו בבעיות של זמן ריצה גבוה ונכונות התוצאות. חיפשנו דרכים לשפר את הדברים הללו ומצאנו שכאשר בוחרים מרכזים ראשוניים רחוקים יותר אז התוצאות וזמן הריצה משתפרים. התוספת של בחירת המרכזים הראשוניים באופן מושכל נקראת kmeans++

### השוואה בין kmeans ל kmeans++:

Figure 1 kmeans



‘

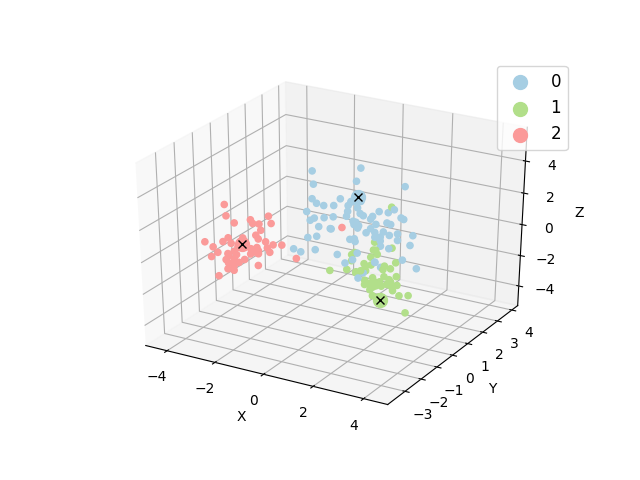


Figure 2 Kmeans++

## מימוש Gaps-Statistics:

### מימוש Monte-Carlo:

## מימוש Cmeans:

# סיכום: