תרגיל בית 6 –Clustering,PCA

יש להגיש שני קבצים נפרדים: קובץ PDF ובו פתרון התרגיל כולל הפלטים של החלק המעשי וקובץ נוסף ובו הקוד שכתבתם. יש להקפיד על תשובות ברורות ומסודרות ועל קוד מסודר ומתועד היטב. שמות שני הקבצים צריכים להיות HW6_ID1_ID2, כאשר ID היא תעודת הסטודנט של המגישות. רק את מבנות הזוג צריכה להגיש את התרגיל

שאלות על התרגיל יש לכתוב בפורום תרגילי הבית באתר הקורס. התרגיל מנוסח בלשון נקבה אך מתייחס לשני המינים.

שאלה 1

נתונות m תצפיות $x_1, \dots, x_m$ ממימד $p > 2$ אשר מקיימות : $\frac{1}{m}\sum_{i=1}^{m} x_i = 0$

נסמן ב- X את המטריצה שמכילה את התצפיות בעמודות ואת מטריצת השונויות של התצפיות ב- $XX^T = \Sigma$.

עבור כל וקטור u נגדיר את $y = u^T x$ להיות הטרנספורמציה של x באמצעות u .

נחפש את וקטור היחידה u ($u^T u = 1$) אשר ממקסם את השונות של התצפיות $y_1, \dots, y_m$:

$$u_1 = argmax\left\{\sum_{i=1}^{m}(y_i - \bar{y})^2\right\} \ , s.t. u^T u = 1$$

ראינו בהרצאה כי u_1 הינו הוקטור העצמי של Σ אשר מתאים לערך העצמי המקסימלי λ_1

א. הסבירי לשם מה נחוץ האילוץ ש- u הינו וקטור יחידה.

ב. בבואנו להוסיף את הכיוון השני המשמר את מירב השונות ($Principal\ Component\ 2$), אנחנו ראשית אוכפים את האילוץ כי המשתנים $u_1^T x_i$ ו- $u_2^T x_i$ יהיו בלתי מתואמים: $\sum_{i=1}^{m}(u_1^T x_i)(u_2^T x_i) = 0$.

הסבירי בקצרה את הדרישה והוכיחי כי היא גוררת את האילוץ ש u_1 ו- u_2 יהיו ניצבים זה לזה $u_1^T u_2 = 0$

רמז: זכרי את הגדרת Σ ואת העובדה ש- u_1 הינו וקטור עצמי של Σ .

שאלה 2- נכון לא נכון

עבור כל טענה קבעו האם היא נכונה/לא נכונה. אנא ספקו הסבר לכל תשובה

1.יש ברשותנו קופסא שחורה ALG אשר, בהינתן קבוצה של וקטורים $\mathcal{S}$ ומספר k, מריצה את אלגוריתם *k-means* שנלמד בכיתה. הקריאה $ALG(\mathcal{S}, k)$ מחזירה את ערך פונקציית המטרה המתאים.

טענה: "אם $k_1 < k_2$, בהכרח מתקיים ש $ALG(\mathcal{S}, k_1) \geq ALG(\mathcal{S}, k_2)$".

2. לקחנו את נתוני *USArrests*, הכוללים 50 תצפיות וארבעה משתנים מסבירים. לאחר נירמול ביחס לממוצע ולסטיית התקן, הפעלנו אלגוריתם *PCA* על המדגם. המודל שחזר הינו

```
pr.out=prcomp(USArrests,center=TRUE, scale=TRUE)
pr.out
```

```
## Standard deviations (1, .., p=4):
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
##
## Rotation (n x k) = (4 x 4):
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

<u>טענה:</u> "שימוש בשלושת ה *PC* (*principal components*) הראשונים יביא לשחזור של מעל 90% מן השונות".

<u>שאלה 3- קוד</u>

בשאלה זו תממשו את אלגוריתם K-means. מצורף קוד ראשוני

א. טענו את קובץ הנתונים data.csv

ב. ממשו פונקציית dist המקבלת זוג תצפיות (lists) ומחזירה את המרחק (האוקלידי) ביניהן

ג. כתבו פונקצייה K-means המקבלת כקלט : K- מס' קלאסטרים , data- dataframe ובו הנתונים.
על הפונקציה לממש את האלגוריתם החל משלב האתחול ועד לשלב ההתכנסות.
בסיום פעולתה הפונקציה תחזיר את ערך פונקציית המטרה של האלגוריתם (כפי שראינו בתרגול 12)
<u>שימו לב:</u> אתחלו את המרכזים הראשוניים בצורה רנדומלית אך בטווח הערכים של הנתונים שקיבלתם על מנת להימנע מקלאסטרים ריקים.

ד. השתמשו בפונקצייה מסעיף ג' והציגו (וצרפו לPDF) גרף של ערך פונקציית המטרה כתלות ב-K. מהו ערך ה-K הנכון לדעתכם? נמקו

<u>הערה:</u> מכיוון שערך פונקציית המטרה נובע ממיקומי האתחול הראשוניים, אנא הריצו עבור כל K מספר פעמים ובחרו את הערך הנמוך יותר.