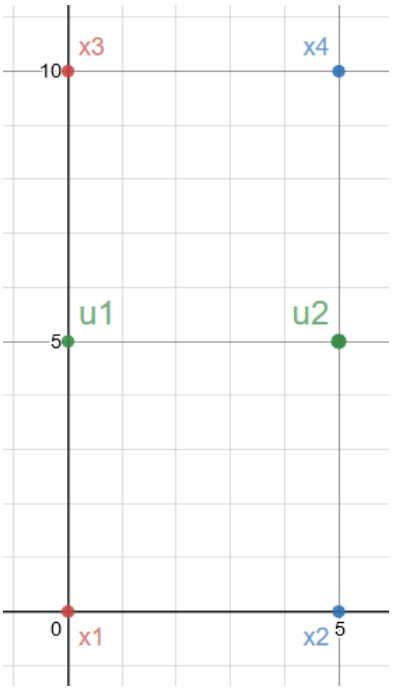
שאלות תאורטיות:

1.1 התכנסות של k-means:

בכל צעד של האיטרציה של k-means אנו מעדכנים את הקבוצות כך שתוצאת פונקציית ההפסד תהיה קטנה או שווה לתוצאת פונקציית ההפסד באיטרציה הקודמת. תנאי העצירה של האלגוריתם הוא כאשר התוצאות בין האיטרציות יהיו שוות. מכיוון שבסה"כ יש $k^{|X|}$ השמות אפשריות לקבוצות, לאחר מקסימום של $k^{|X|}$ אנו נתכנס. לכן נתכנס לאחר מספר סופי.

1.2 סאב-אופטימליות של k-means:

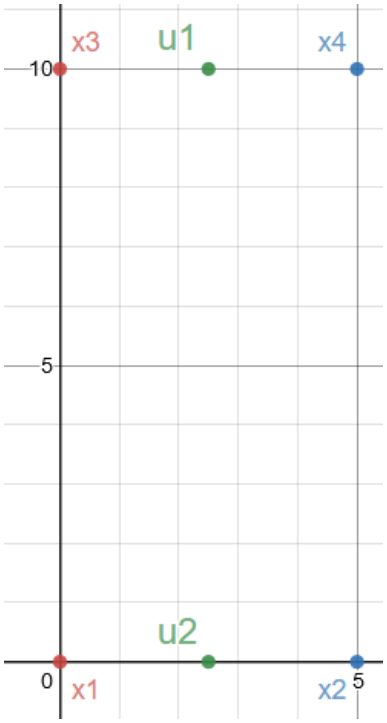
1. נסתכל על המקרה הבא עבור k=2:



נשים לב כי לפי האתחול שלנו, המרכזים u1,u2 לא יתעדכנו יותר. לכן התכנסנו לפונקציית הפסד עם התוצאה

$$\sum_{i=1}^{k=2} \sum_{x_j \in C_i} \left\| x_j - u_i \right\| = 20$$

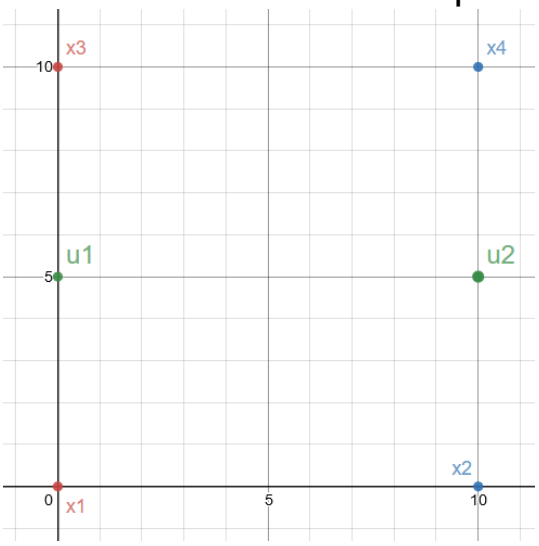
לעומת זאת, אם היינו מאתחלים את המרכזים בצורה הבאה:



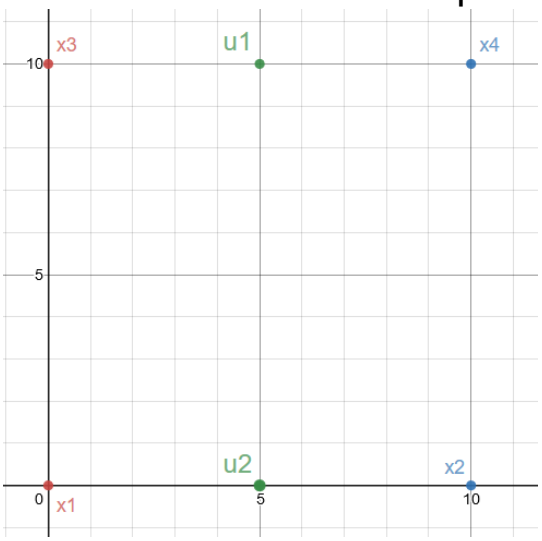
היינו מקבלים פונקציית הפסד עם הערך 10. שהיא האופטימלית האמיתית.

2. דוגמה למקרה בו יש שני פתרונות אופטימליים:

פתרון א':

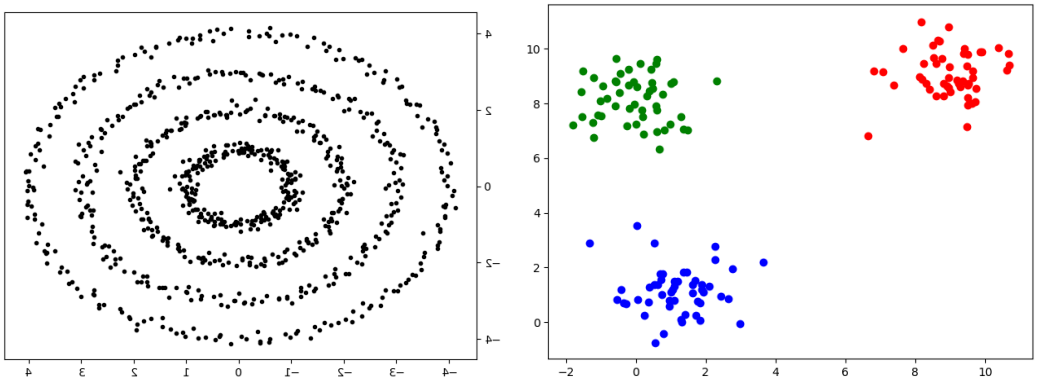


פתרון ב':

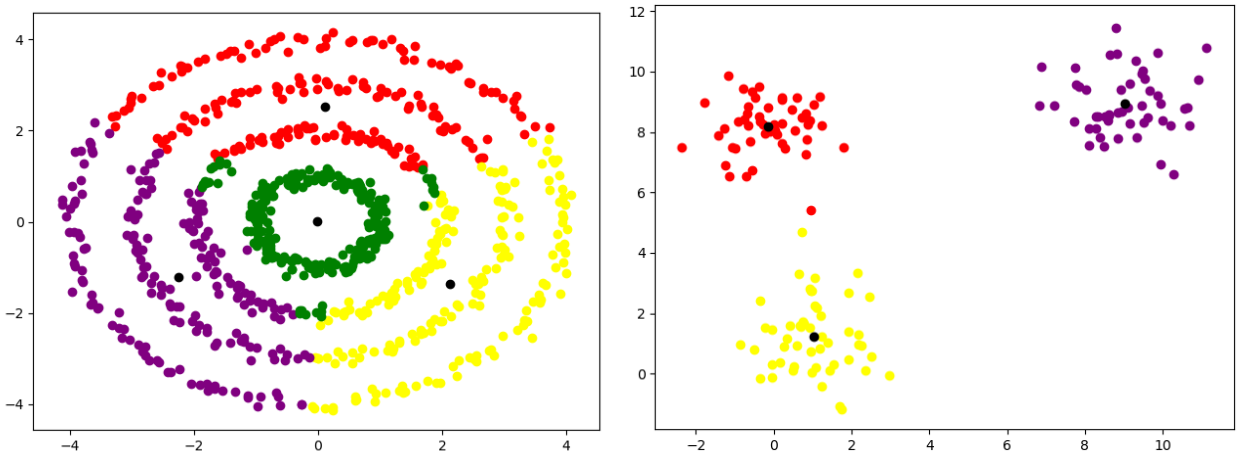


2.4 Spectral Clustering:

לצורך הדגמת המודלים השתמשתי בשני סוגי דאטא – אחד הוא הcircles שנתתם לנו, ואחד הוא דאטא סיננטית שיצרתי בעצמי. הדאטא נראים ככה:



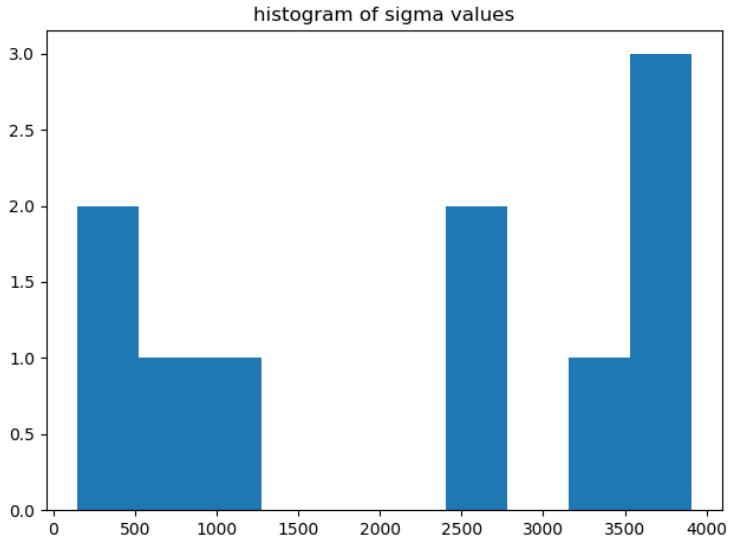
<u>תוצאות K-means:</u>

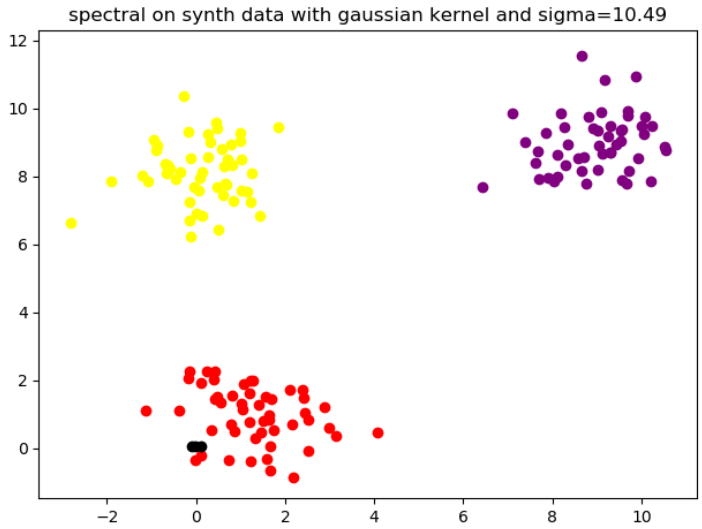


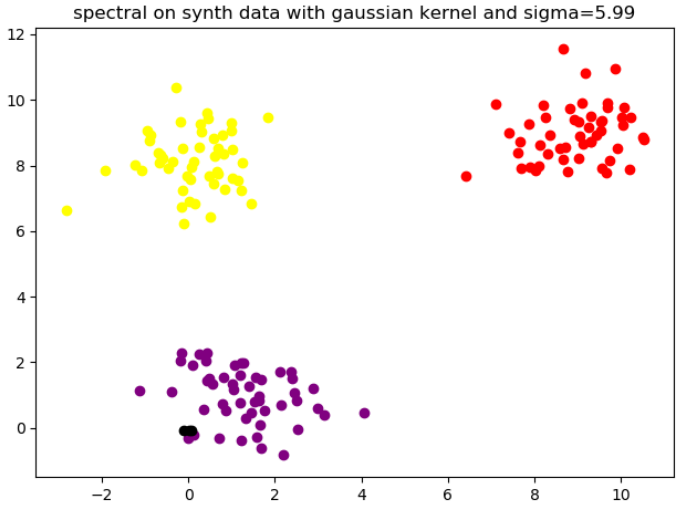
ניתן לראות כי בדאטא הסינטטי k-means השיג תוצאות מושלמות, אך בדאטא המעגלי הוא קיבל תוצאות לא טובות בכלל (שכן הוא מתבסס על מרחק, והמרחק במקרה הנ"ל לא מסווג טוב את הקבוצות).

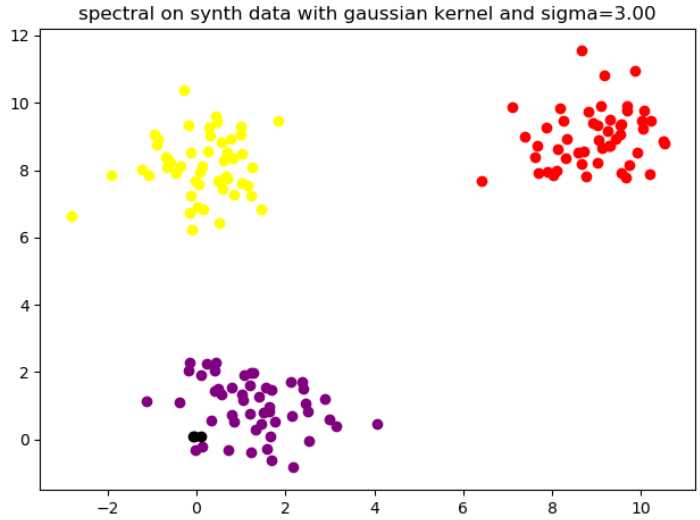
תוצאות spectral clustering:

היסטוגרמה על ערכי סיגמה של הדאטא הסינטטי והתוצאות עם ערכים שונים:



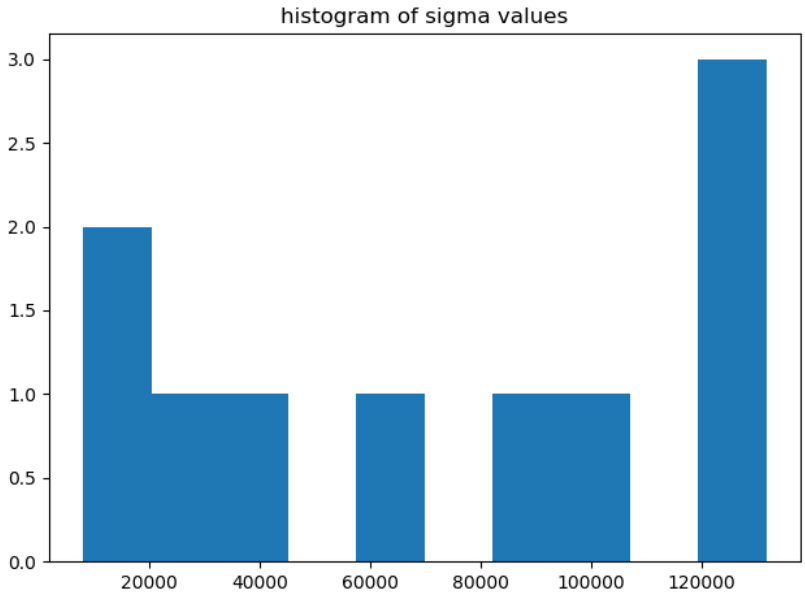


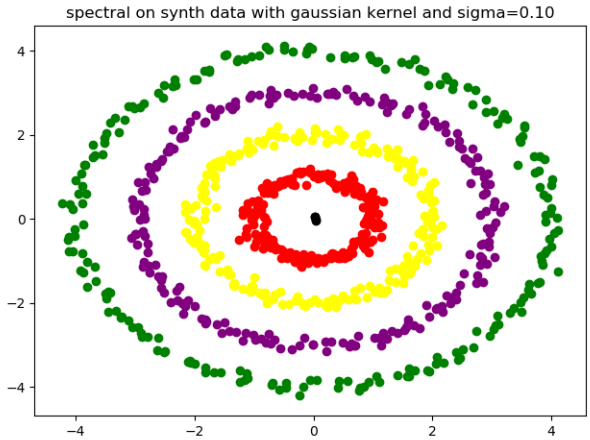


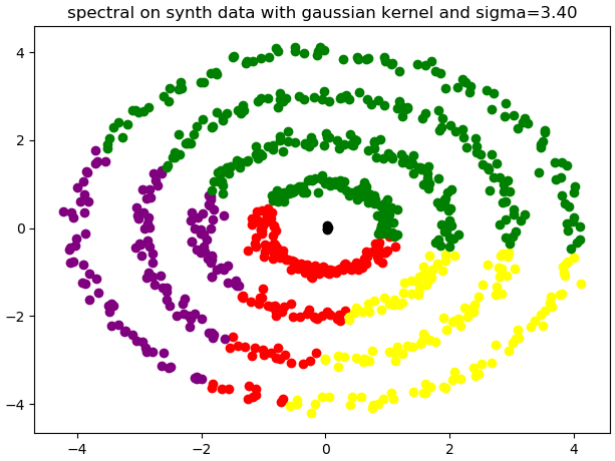


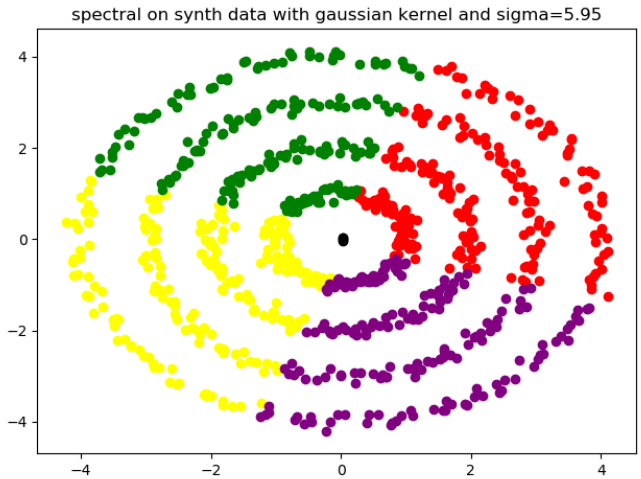
ניתן לראות כי בכל הגרפים הcentroids (המיוצגים בשחור) אינם משקפים את הcentroids האמיתיים. כמו כן עבור כל הערכים של סיגמה סיווגנו באופן מושלם.

תוצאות על הדאטא של המעגלים עם ערכי סיגמה שונים (שימוש בgaussian kernel):









ניתן לראות כי עבור ערכי סיגמה קטנים מאוד (קטנים יותר מהעשירון הראשון) אנו מקבלים clustering מושלם, אך עבור ערכים גדולים יותר אנו מקבלים clusters המתבססים על מרחק ואינם מדויקים. כמו כן, הcentroids בכולם אינם מדויקים.

האם הsimilarity graph הוא ייצוג טוב של הclustering?

כן, אם נסתכל על המקרה בו אנו משתמשים בmnn בשביל יצירת הsimilarity graph, נראה כי עבור כל נקודה – השכנים שלה הם בתוך הcluster שאנו רוצים עבורה, שכן בכל מעגל הנקודות הקרובות ביותר לכל נקודה הן בתוך אותו המעגל.

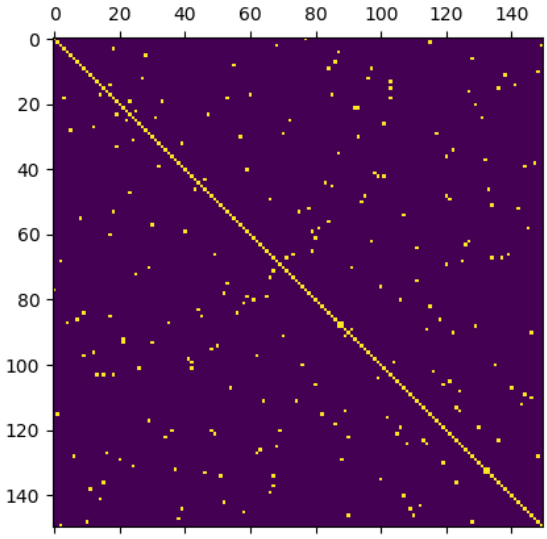
גם בדאטא הסינטטי נראה אותו הדבר, מכיוון שהנקודות הקרובות לכל נקודה נמצאות בcluster שלה.

2.4.1 – similarity graph:

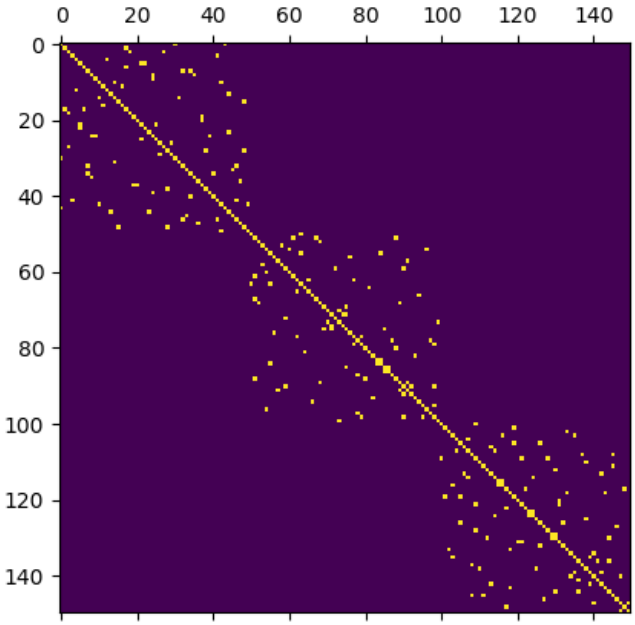
נסתכל על הדאטא הסינטטי מכיוון שיש בו פחות נקודות ולכן יהיה יותר קל להסתכל על הגרף.

השתמשתי כאן בmnn עם m=2.

כאשר הדאטא מעורבב נקבל את המטריצה הבאה:



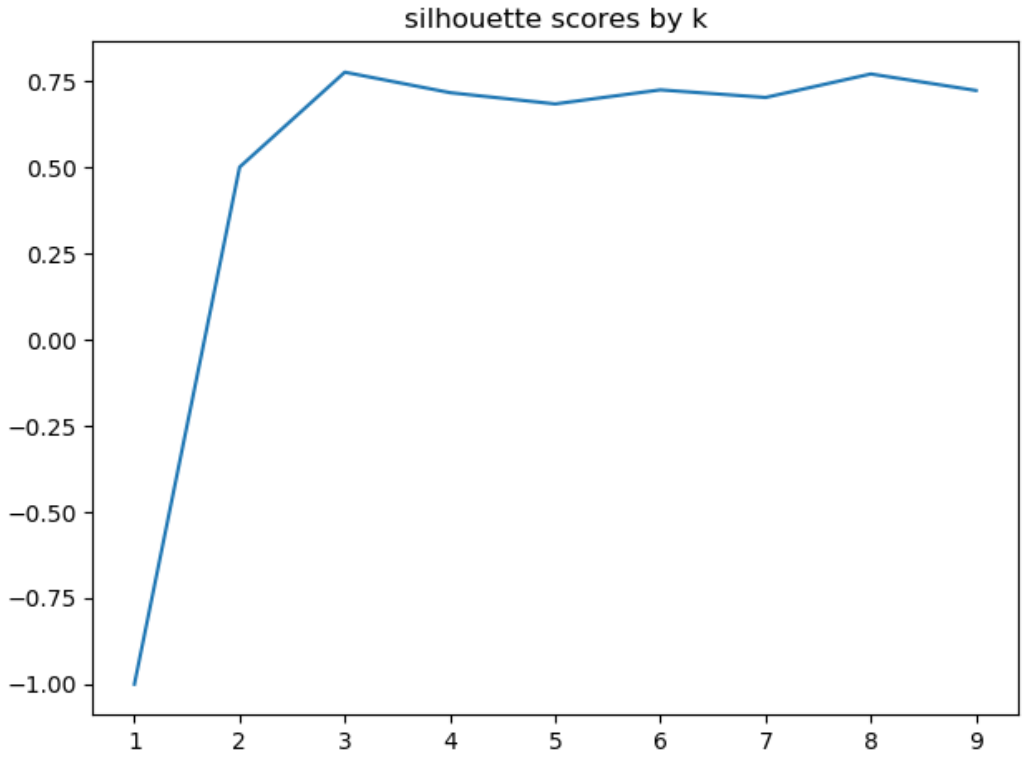
לעומת זאת כאשר הדאטא מסווג נקבל את המטריצה:



ניתן לראות בבירור כי כאשר המטריצה מסודרת השכנים של כל נקודה נמצאים באותו cluster שלה.
לכן אפשר להסיק כי במקרה הזה הsimilarity graph באמת מספק חיזוי טוב לclustering של הנקודות.

2.5 בחירת K:

כדי לבחור את הK הרצתי את k-means ואת silhouette על ערכים שונים של K על הדאטא הסינטטי (שהK האופטימלי שלו הוא 3, כי ככה יצרתי אותו) וקיבלתי את הגרף הבא:

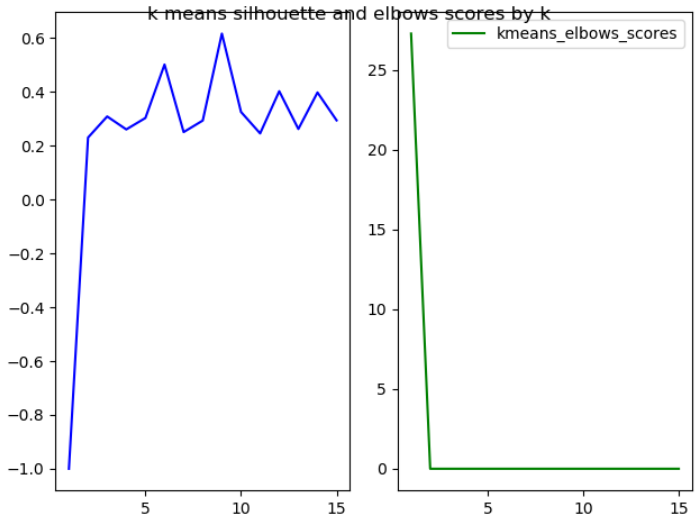


ניתן לראות כי גם בגרף הנקודה המקסימלית היא בK=3, ולכן גם הsilhouette מזהה כי הK האופטימלי הוא אכן 3.

2.6 biological clustering:

בחלק הזה עשיתי השוואה בין k-means רגיל לבין spectral עם mnn עבור 5 שכנים.

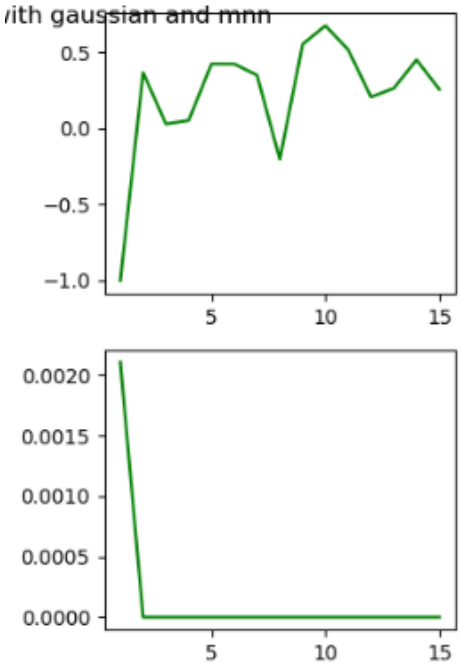
<u>K-means:</u>



(מצד ימין – elbow, מצד שמאל – silhouette)

לפי k-means הK הטוב ביותר הוא K=9 עם תוצאת silhouette של 0.61.

<u>Spectral:</u>



(למעלה – הsilhouette, למטה – הelbow)

לפי הspectral הK הטוב ביותר הוא K=10 עם תוצאת silhouette של 0.67.

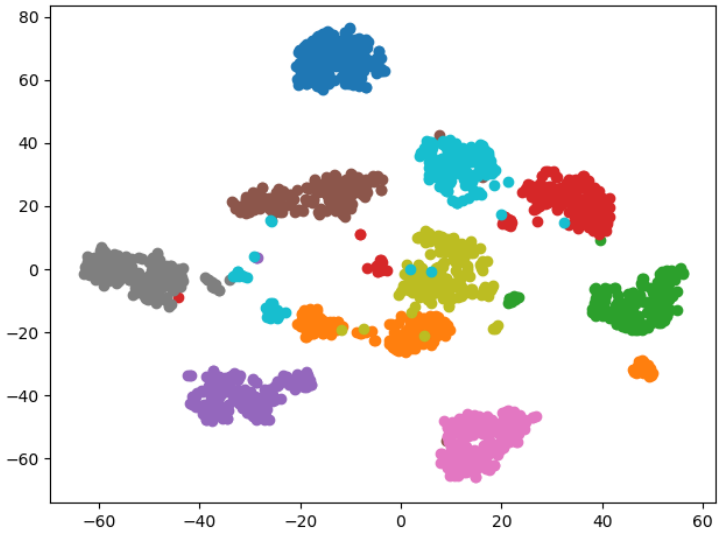
לפי דעתי הspectral יותר משכנע במקרה הזה, שכן באזור הK שהוא מראה הsilhouette נותן ערכים גבוהים (כלומר גם עבור k=9 וk=11 מתקבלים ערכים גבוהים). ואילו בk-means הערכים באזור הם נמוכים משמעותית (~0.3) ולכן מראים פחות עקביות.

TSNE 2.7:

השתמשתי ב2 דאטאסטים כדי להראות את ההבדל בהורדת המימד בין השיטות. הדאטאסט הראשון הוא הMNIST הקטן (תמונות 8x8), והדאטאסט השני הוא סינטטי שאני יצרתי (מופיע בעמודים הקודמים).

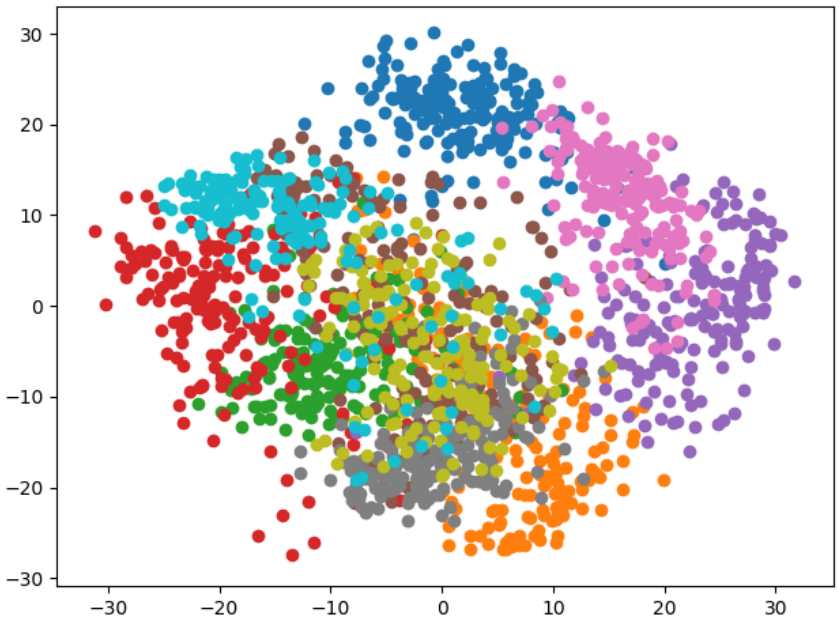
<u>דאטאסט ראשון – MNIST קטן:</u>

שימוש בTSNE עבור הורדת מימד למימד 2:



הורדת המימד שמרה באופן טוב על יחסית על המבנה של הדאטא, כאשר ניתן לראות חריגות יחסית קטנות (סה"כ 1800 דגימות, כמה עשרות חריגות).

שימוש בPCA עבור הורדת מימד למימד 2:



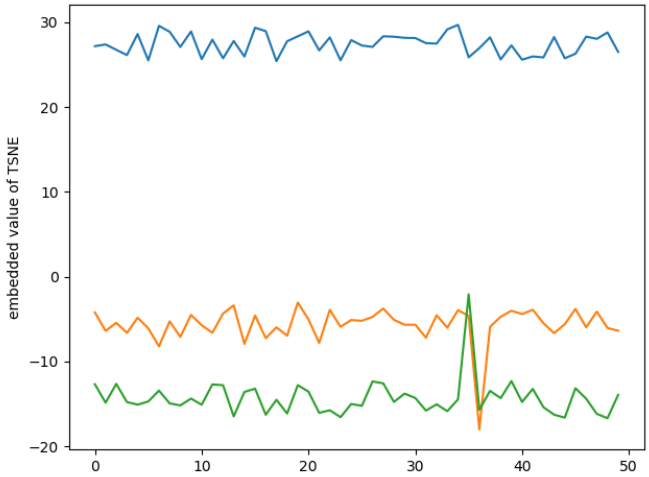
שאלוהים יעזור. PCA לא שמר בכלל על המבנה.

לסיכום – TSNE שמר על המבנה הרבה יותר טוב מPCA במקרה הזה.

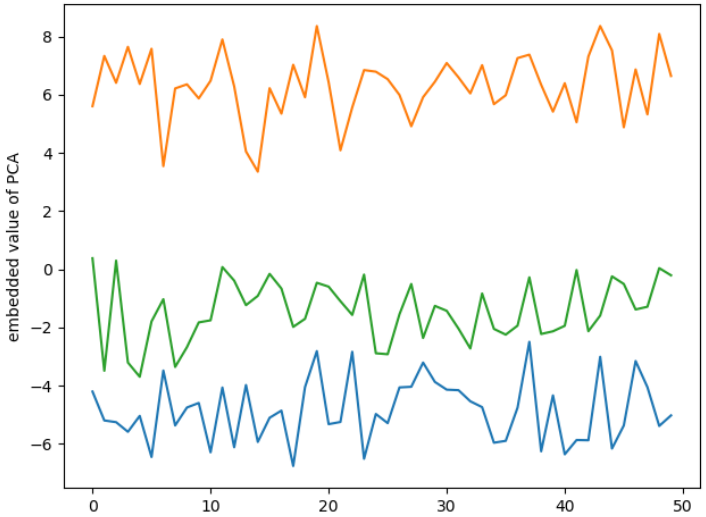
דאטאסט שני – דאטא סינטטי:

הפעם לקחתי את הדאטא שהמימד ההתחלתי שלו הוא 2, והורדתי אותו למימד 1 בעזרת TSNE וPCA. ולהלן התוצאות:

TSNE:



PCA:



ציר הy מראה את הערכים (ממימד 1) לאחר ההורדה. ציר הx לא מראה כלום.

ניתן לראות כי בדוגמה הזאת דווקא PCA פעל טוב יותר – והצליח להפריד באופן מוחלט בין הדאטא מהסוגים השונים, לעומת TSNE שנפל קצת באזור x=35.