

פרק 3: עיבוד מקדים וניתוח מידע

- 3.1 מבוא: סוגי מידע, פעולות בסיסיות, מאפיינים
- 3.2 הורדת מימד: ניתוח רכיבים עיקריים (PCA)
- 3.3 אישכול (clustering) – אלגוריתם K-means

בפרק זה נתבונן באוסף נתונים כללי, מהצורה $D_X = \{x_i\}_{i=1}^n$.

בהקשר של למידה מודרכת, נתונים אלה עשויים להיות רכיבי הקלט של סדרת האימון המתויגת $D = \{x_i, y_i\}_{i=1}^n$. השלבים הראשונים של תהליך הלימוד כוללים באופן טיפוסי עיבוד ראשוני של הקלט D_X , ניתוח הקלט להבנת אופיו, ובחירת ייצוג מתאים עבורו.

באופן כללי יותר, ניתן להתבונן באוסף כלשהו של נתונים בעלי עניין, כאשר אנו מעוניינים להבין תכונות מסוימות של נתונים אלה, ולדלות (או להסיק) מהם מידע מועיל כלשהו. ניתוח מעין זה של נתונים הינו מוקד העניין בתחומי למידה בלתי-מודרכת (unsupervised learning) וכריית מידע.

בפרק זה נציג בקצרה שני נושאים בסיסיים מתחום זה – הורדת מימדיות של נתונים וקטוריים, ואישכול (חלוקת קבוצת הנתונים המידע למספר תת-קבוצות, או אשכולות). תיאור מעמיק יותר של נושאים אלה ניתן בקורס "עיבוד וניתוח מידע" ובקורסים מתקדמים בתחום עיבוד אותות.

3.1 מבוא: סוגי נתונים, פעולות בסיסיות, מאפיינים

כאמור, נתבונן בפרק זה באוסף נתונים, או פריטי מידע: $D_X = \{x_i\}_{i=1}^n$, כאשר $x_i \in \mathbb{X}$.

דוגמאות לסוגים נפוצים של פריטי מידע:

1. וקטור של גדלים מספריים, בעל מימד קבוע d : $x \in \mathbb{R}^d$. למשל:

$$x_1 = (1, 2, 7, 2), x_2 = (1, 1, 4, 5), x_3 = (8, 9, 3, 5), \dots$$
2. מידע סמנטי מחולק לרשומות: למשל $x = \{\text{שם, כתובת, עיסוק, מספר חשבון} \dots\}$ ואז:

$$x_1 = (\text{John, Librarian, 3032}), x_2 = (\text{Lisa, Doctor, 4315}), \dots$$
3. קובץ טקסט בעל אורך משתנה. למשל: דואר אלקטרוני.
4. סדרה זמנית. למשל: שערים יומיים של מניה לאורך שנה.
5. אות שמע דגום. למשל: דיבור, רעשי סביבה.
6. תמונה (בייצוג נומרי מטריצי), למשל:

$$x_1 = \text{[Siamese Cat]}, x_2 = \text{[Orange Cat]}, x_3 = \text{[Kitten]}, \dots$$

7. סרטון וידאו.

8. גרף קישוריות – בהקשר לרשתות חברתיות למשל.

דוגמאות לפעולות מקדימות של טיפול בנתונים:

1. ניקוי הנתונים (נתונים שגויים, חסרים, רועשים....).
2. אחסון הנתונים באופן נגיש (בסיסי נתונים).
3. איחוד נתונים ממקורות שונים.
4. דגימה ודילול: בחירת רכיבים מייצגים מתוך הקבוצה המלאה $D_X = \{x_i\}_{i=1}^n$.
5. נרמול נתונים מספריים.
6. התמרות (למשל התמרת פורייה של אות זמני, דחיסת תמונה).
7. הורדת מימדיות.
8. הגדרת מאפיינים.

אפשר כמובן לשלב פעולות, וצריך לבחור אותן בהתאם להקשר וסוג הנתונים.

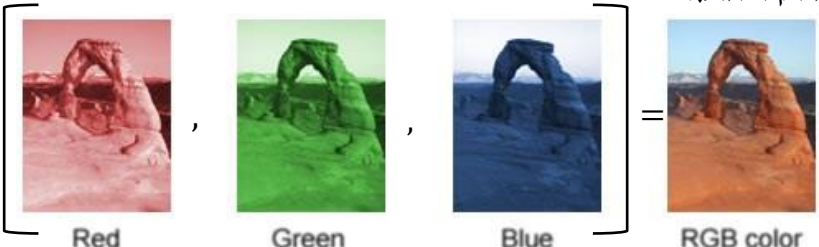
בקורס הנוכחי נתייחס בעיקר לנתונים מהסוג הראשון שהוזכר, כלומר וקטור מספרי בעל מימד קבוע, $x \in \mathbb{R}^d$. כלומר, כל דוגמא היא וקטור עמודה

$$x_i = (x_i(1), \dots, x_i(d))^T = (x_i(j))_{j=1}^d \in \mathbb{R}^d$$

שימו לב: בקורס זה, אלא אם כן נאמר אחרת, האינדקס בסוגריים (j) יסמן את הרכיב ה- j בדוגמא, בעוד שהאינדקס החיצוני i מציינ את מספר הדוגמא.

חשוב לציין כי גם סוגי נתונים מורכבים יותר ניתן לייצג (לצרכי סיווג למשל) באמצעות וקטור כזה, על ידי התמרה או ייצוג מתאים. למשל:

(1) ייצוג תמונה. תמונה ללא צבע ניתנת לייצוג ע"י מטריצה של מספרים המייצגים את הבהירות של כל פיקסל. תמונה עם צבע ניתנת לייצוג ע"י מספר מטריצות כאלה – בד"כ שלוש, אחת לכל צבע יסוד (אדום, ירוק וכחול):

$$x_1 = \left[\begin{array}{c} \text{Red} \\ \text{Green} \\ \text{Blue} \end{array} \right], \quad x_2 = \dots$$


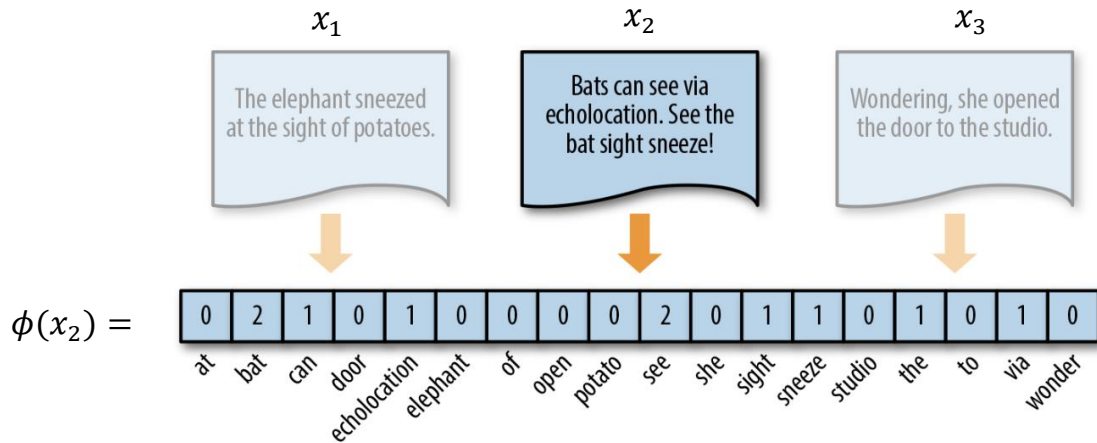
אם נשרשר את כל הערכים המספריים במטריצות אלה נוכל לכתוב כל דוגמא בתור וקטור

$$x_i = (x_i(1), \dots, x_i(d))^T \in \mathbb{R}^d,$$

כך שכל רכיב בוקטור, $x_i(j)$, הוא ערך מספרי המייצג את הבהירות של הפיקסל באחת המטריצות (אדום, ירוק או כחול) של הדוגמא ה- i .

(2) ייצוג קובץ טקסט. נניח כי $x = x_i$ הוא קובץ טקסט של הודעת דוא"ל, אותה יש לסווג למספר קטגוריות (למשל: דחוף \ רגיל \ אישי \ ספאם).

ייצוג בסיסי של הקובץ לצורך זה הוא בצורה של "שק מילים" – bag of words. בייצוג זה קובעים מראש מספר נתון של מילים (החל מכמה מעשרות וכלה במיליון השלם), ומייצגים את הקובץ על ידי התדירות היחסית של מילים אלה בתוכו. לפיכך, x מיוצג על ידי וקטור הסתברות $\phi(x)$ במימד קבוע (כמספר המילים במילון שבחרנו). למשל



נסו לחשוב – למה שווים $\phi(x_1)$ ו- $\phi(x_3)$?

מאפיינים:

הוקטור $\phi(x)$ בדוגמא הקודמת נקרא **וקטור מאפיינים** (feature vector) עבור פריט המידע המקורי x . ניתן לראות כי ייצוג הקובץ x בעזרת המאפיינים $\phi(x)$ בלבד כרוך באובדן מידע – אך זהו ייצוג נוח שעשוי להספיק לצרכינו.

כללית, מאפיין הינו גודל (מספרי לרוב) הנגזר מפריט המידע המקורי, ואשר עשוי להועיל בפעולות המשך כגון זיהוי וסיווג. בחירת מאפיינים מתאימים לייצוג והעשרת מידע הקלט הינה בעלת חשיבות קריטית במשימות של למידה המודרכת. על כך נרחיב בהמשך הקורס.

בהמשך הקורס, הסימון x (או x_i לפריט i) עשוי להתייחס הן לנתון הגולמי, והן לוקטור מאפיינים של הנתון הגולמי. כאשר נרצה להבליט כי מדובר במאפיינים נשתמש בסימון $\phi(x)$.

מירכוז ונירמול:

בהינתן אוסף נתונים וקטוריים $D_X = \{x_i\}_{i=1}^n$, $x_i = (x_i(j))_{j=1}^d \in \mathbb{R}^d$, לצרכים מסוימים (כגון שימוש כקלט לרשת נוירונים) רצוי לבצע פעולות מקדימות כגון:

מירכוז:

$$x_i \rightarrow x_i - \bar{x}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

מירכוז ונירמול לפי שונות (וריאנס):

$$x_i(j) \rightarrow \frac{x_i(j) - \bar{x}(j)}{\sigma(j)}, \quad \sigma(j) = \frac{1}{\sqrt{n}} \|x_i - \bar{x}\| = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i(j) - \bar{x}(j))^2}$$

מירכוז ונירמול משרעת:

$$x_i(j) \rightarrow \frac{x_i(j) - \text{Min}_j}{\text{Max}_j - \text{Min}_j} \in [0,1], \quad \text{Max}_j = \max_{i=1 \dots n} (x_i(j)), \text{Min}_j = \min_{i=1 \dots n} (x_i(j))$$

3.2 הורדת מימד: ניתוח רכיבים עיקריים (PCA)

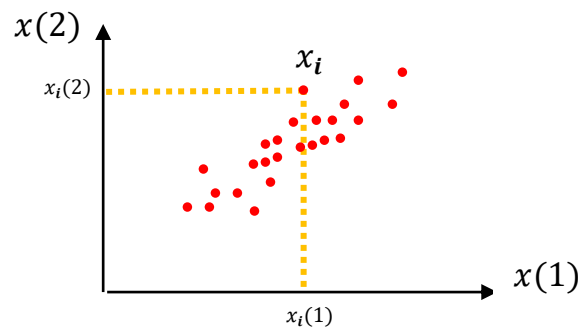
נניח כי $D_X = \{x_i\}_{i=1}^n$ הינה קבוצה של וקטורים רב מימדיים, $x_i \in \mathbb{R}^d$, במימד גבוה: $d \gg 1$. האם ניתן לייצג וקטורים אלה על ידי וקטורים במימד נמוך יותר, תוך שמירה על תכונות רצויות מסוימות של אוסף זה?

בעיה זו של הורדת מימדות הינה שימושית למספר צרכים, ביניהם:

1. ייצוג קומפקטי יותר של וקטורים גדולים (כגון תמונות).
2. הורדת מאפיינים (רכיבים) שאינם "אינפורמטיביים" (אינם תורמים למשל לסיווג המידע).
3. ויזואליזציה – ציור במימד נמוך של מאפיינים מרכזיים של קבוצת הנתונים.

קיים מספר רב של שיטות להורדת מימד, החל משיטות לינאריות שהן פשוטות יחסית, וכלה בשיטות לא-לינאריות שהן מורכבות יותר. פה נתאר בקצרה את גישת ניתוח הרכיבים העיקריים (PCA – Principle Component Analysis), שהיא בעיקרה שיטה של התמרה (או הטלה) לינארית של המידע למרחב במימד נמוך יותר. בבסיס הגישה ההנחה כי נקודות המידע מרוכזות על או קרוב לתת-מרחב לינארי כלשהו של המרחב הראשוני.

להדגמת הבעיה, נתבונן באוסף הנקודות הבאות במימד 2 ($d = 2$):



נניח כי ברצוננו לייצג נקודות אלו על ידי נקודות במימד 1, כלומר קואורדינטה אחת בלבד. כיצד נבחר אותה?

הבחירה תלויה כמובן בתכונה אותה מעוניינים לשמר. נציין שתי תכונות רצויות:

1. שונות מירבית: נבחר כיוון במרחב x אשר לאורכו המרחק בין נקודות המידע הוא מכסימלי.
2. שגיאת שחזור מינימאלית: נבחר ייצוג אשר יאפשר לשחזר את הנקודות המקוריות עם שגיאה מינימאלית.

שני קריטריונים אלה מובילים לפתרון זהה שהוא אלגוריתם ה-PCA. זהו אלגוריתם נפוץ, פשוט ושימושי להורדת המימד. ניתן לראות [כאן](#) כמה דוגמאות לשימושים של אלגוריתם זה, כמו ניתוח מאפייני צריכת האוכל בבריטניה (האלגוריתם מוצא שבצפון אירלנד יש הבדל משמעותי – שאוכלים יותר תפוחי אדמה).

בפרק זה, נתאר ראשית פתרון זה, ולאחר מכן נגדיר ביתר דיוק את הקריטריונים שהוזכרו.

א. הגדרות ותזכורות

עבור אוסף נקודות $\{x_i\}_{i=1}^n$, כאשר $x_i \in \mathbb{R}^d$, נגדיר את מטריצת שונות המדגם (sample- covariance matrix) באופן הבא:

$$P_n \triangleq \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \in \mathbb{R}^{d \times d}$$

פה $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ הוא כרגיל וקטור הממוצעים.

הערות להגדרת P_n :

1. המקדם $\frac{1}{n}$ מוחלף לעתים ב- $\frac{1}{n-1}$ או פשוט ב- 1. הדבר אינו משנה את התוצאות להלן, כיוון שאנו מעוניינים רק בערכים העצמיים והוקטורים העצמיים של המטריצה.

2. בחישובי PCA מקובל ראשית למרכז את הנתונים, כלומר להציב $x_i \leftarrow x_i - \bar{x}$, ולאחר מכן להמשיך בחישוב. כאשר הנתונים ממורכזים, מטריצת שונות המדגם היא פשוט

$$P_n = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$$

3. מטריצה שנפגוש בהמשך הקורס היא מטריצת הנתונים: $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$.

במונחים של מטריצה זו, נקבל (עבור נתונים ממורכזים): $P_n = \frac{1}{n} X^T X$.

המטריצה P_n הינה ממשית, סימטרית, ואי-שלילית מוגדרת (מדוע?). לפיכך, כפי שידוע מתוצאות בסיסיות באלגברה לינארית:

- א. P_n היא בעלת d ערכים עצמיים ממשיים, שיסומנו: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$.
- ב. P_n היא בעלת d וקטורים עצמיים אורתונורמליים, שיסומנו בהתאמה v_1, v_2, \dots, v_d .
דהיינו: $P_n v_k = \lambda_k v_k$, $v_k^T v_j = \delta_{kj}$.

חשוב לזכור: סדר הוקטורים העצמיים הוא לפי סדר הערכים העצמיים.

כלומר, v_1 הוא הוקטור העצמי המתאים לערך העצמי הגדול ביותר, v_2 מתאים לערך העצמי השני הגדול ביותר, וכך הלאה.

נזכיר גם כי מתקיים (לכל מטריצה סימטרית):

$$P_n = V \Lambda V^T = \sum_{k=1}^d \lambda_k v_k v_k^T,$$

כאשר $V = [v_1, \dots, v_d]$, $\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}$, וכן $V^{-1} = V^T$.

ב. כיוונים ורכיבים העיקריים

הווקטורים העצמיים v_1, v_2, \dots, v_d נקראים גם **הכיוונים העיקריים** של P_n . v_1 הוא הכיוון העיקרי הראשון, v_2 השני, וכולי.

וקטורים אלה מהווים בסיס למרחב \mathbb{R}^d . ניתן לפיכך להציג את כל וקטור $x \in \mathbb{R}^d$ (ובפרט את הוקטורים $\{x_i\}_{i=1}^n$) בעזרת בסיס זה. ברישום מטריצי:

$$x_i = Vz_i, \quad z_i = V^T x_i$$

או, ברישום לפי רכיבים:

$$x_i = \sum_{k=1}^d z_{ik} v_k, \quad z_i(k) \triangleq v_k^T x_i = \sum_{j=1}^d x_i(j) v_k(j)$$

הוקטור z_i הוא ייצוג של x_i בעזרת הבסיס החדש (שינוי בסיס).

המקדמים $z_i(k)$ הם ה**רכיבים העיקריים** של הוקטור x_i , כאשר $z_i(1)$ הוא הרכיב העיקרי הראשון, $z_i(2)$ הוא הרכיב העיקרי השני, וכך הלאה.

הגדרה: יהי $1 \leq m \leq d$. ייצוג רכיבים-עיקריים (PCA) ממימד m של הנקודות $\{x_i\}_{i=1}^n$ מתקבל על ידי לקיחת m הרכיבים העיקריים הראשונים עבור כל נקודה x_i (והשמטת יתר הרכיבים). כלומר, ברישום מטריצי, אם נסמן $V_m = [v_1, \dots, v_m]$, אז

$$z_i^{(m)} \triangleq V_m^T x_i$$

או, בכתיבה לפי רכיבים,

$$z_i^{(m)} = \begin{bmatrix} z_i(1) \\ z_i(2) \\ \vdots \\ z_i(m) \end{bmatrix}, \quad z_i(k) = v_k^T x_i$$

ג. אלגוריתם ה-PCA להורדת המימד

נתון: אוסף וקטורי עמודה x_1, \dots, x_n ב- \mathbb{R}^d , מימד רצוי $m < d$.

$$1. \text{ מירכוז: } x_i \leftarrow x_i - \bar{x}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

2. נחשב את m הוקטורים העצמיים הראשונים v_1, \dots, v_m של מטריצת שונות המדגם:

$$P_n = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$$

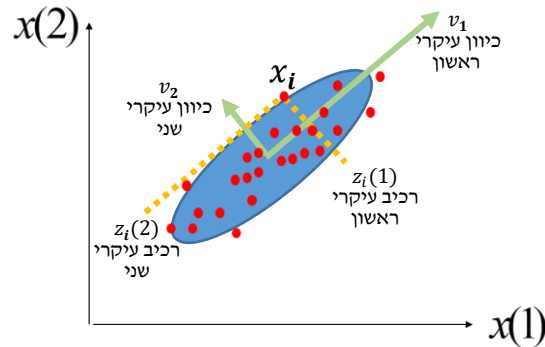
כזכור, סדר הוקטורים העצמיים נקבע לפי גודל הערכים העצמיים, ולכן אנחנו בוחרים את m הוקטורים העצמיים להם יש את הערכים העצמיים הכי גדולים.

3. נחשב את וקטורי הרכיבים העיקריים במימד m ,

$$z_i^{(m)} = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_m^T \end{bmatrix} x_i \equiv V_m^T x_i, \quad i = 1, \dots, n$$

יש להדגיש שחישוב מפורש של מטריצת הקווריאנס P_n הוא יקר או אף אינו מעשי כאשר המימד n גדול, וקיימים אלגוריתמים יעילים הנמנעים מחישוב זה.

אינטואיציה גיאומטרית: PCA מתקבל על ידי התאמת אליפסואיד במימד d סביב הדוגמאות, כאשר הכיוונים העיקריים הם כיווני הצירים הראשיים של האליפסואיד (המאונכים זה לזה), בסדר יורד של אורכם (ראו ציור מטה). הרכיבים הראשיים מתקבלים כהטלת הנקודות על צירים אלה.



למשל, בציור למעלה אם $m=1$, עבור כל נקודה x_i נשמור רק את $z_i(1)$, ההטלה של x_i על הכיוון העיקרי הראשון v_1 . כך נוריד את המימדיות של הנתונים מ-2 ל-1.

הסבר מפורט לאינטואיציה הגיאומטרית (*למתעניינים): מדוע הוקטורים העצמיים של מטריצת שונות המדגם הם הצירים הראשיים של האליפסואיד המתאים לנתונים? לשם הפשטות נניח שהנתונים ממורכזים, כך ש- $\bar{x} = 0$. כדי "להתאים אליפסואיד לנתונים" נניח מודל פרמטרי גאוס

$$p_X(x|\theta) = \frac{1}{\sqrt{2\pi}|\Sigma|} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

עבורו קווי הגובה בעלות צורה של אליפסואיד, כפי שנראה בהמשך. כדי להתאים את המודל, נניח שכל הפרמטרים אינם ידועים ונשתמש במשעך MLE. כפי שלמדנו בשיעור הקודם,

$$\hat{\mu}_{MLE} = \bar{x} = 0, \quad \hat{\Sigma}_{MLE} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T = P_n$$

לכן, לאחר שערוך הפרמטרים, המודל יהיה

$$p_X(x|\theta) = \frac{1}{\sqrt{2\pi}|P_n|} \exp\left(-\frac{1}{2}x^T P_n^{-1}x\right)$$

מכאן, אוסף כל הנקודות בקו גובה מסויים $\{x: p_X(x|\theta) = C\}$, יהיה זהה לאוסף הנקודות

שמגדירות אליפסואיד $\{x: x^T A x = C'\}$, עבור $A = P_n^{-1}$ וקבוע C' מסויים.

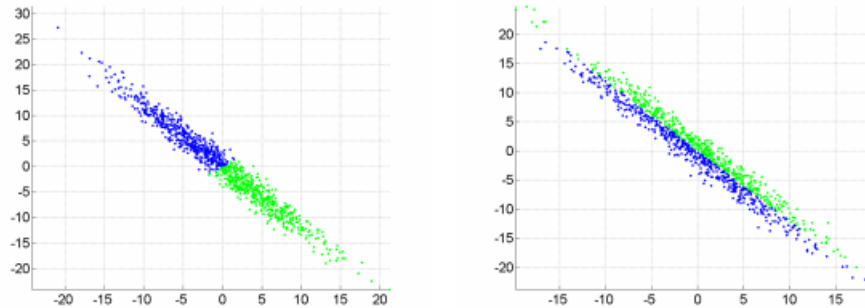
כזכור, $P_n = V\Lambda V^T$ כאשר Λ מטריצה אלכסונית ו- V מטריצה אורתוגונלית (מטריצת סיבוב).

לכן $P_n^{-1} = V\Lambda^{-1}V^T$, ואם נגדיר סיבוב של המרחב לפי V , כלומר $z = V^T x$, נקבל

$$C' = x^T P_n^{-1} x = x^T V \Lambda^{-1} V^T x = z^T \Lambda^{-1} z = \sum_{k=1}^d \lambda_k^{-1} z^2(k)$$

כלומר, אליפסואיד שציריו הראשיים מקבילים למערכת הצירים המסובבת, והאורך של כל ציר פרופורציונלי לשורש הערך העצמי המתאים.

האם הבחירה בצירים הראשיים של האליפסה המתאימה לנתונים בהכרח טובה לשימור המידע המשמעותי בנתונים? לא בהכרח. ההנחה המובלעת בשימוש באלגוריתם ה-PCA היא שלמידע שחשוב לנו יש מתאם גבוה עם השונות בנתונים, אבל זה לא תמיד המצב. למשל, נבחן את שני המקרים בתמונות מטה, בהן יש אוסף של דוגמאות המחולקות לשתי מחלקות – כחול וירוק.



במקרה השמאלי, הציר הראשי של האליפסה (כלומר, הרכיב העיקרי הראשון, שיתקבל מאלגוריתם PCA עם $m=1$), אכן מכיל את כל המידע הנדרש לגבי סיווג הדוגמאות למחלקה הנכונה. לעומת זאת, במקרה הימני, כל המידע לגבי הסיווג הנכון מוכל דווקא ברכיב העיקרי השני. במקרה זה אלגוריתם PCA (עם $m=1$) יפגע ביכולת שלנו לסווג את הנתונים.

ד. שחזור לינארי עם שגיאה מינימאלית

אפיון מעניין של הרכיבים העיקריים מתקבל על ידי סכמת השחזור הבאה. תחילה נניח שהנתונים ממורכזים, כך ש- $\bar{x} = 0$.

עבור $m < d$, תהינה $A \in \mathbb{R}^{m \times d}$, $B \in \mathbb{R}^{d \times m}$ מטריצת הפחתת-מימד ומטריצת שחזור, בהתאמה. נגדיר

$$u_i = Ax_i \in \mathbb{R}^m, \quad \hat{x}_i = Bu_i \in \mathbb{R}^d$$

שגיאת השחזור של x_i הינה $e_i = x_i - \hat{x}_i = (I - BA)x_i$

שגיאת השחזור הריבועית תוגדר על ידי $E(\hat{x}) = \sum_{i=1}^n \|e_i\|^2$, כאשר הנורמה כאן היא הנורמה הרגילה (האוקלידית).

משפט: הערך המינימאלי האפשרי של שגיאת השחזור הריבועית מתקבל עבור

$$A = V_m^T, \quad B = V_m \equiv [v_1, \dots, v_m]$$

כלומר, $\hat{x}_i = V_m V_m^T x_i$, והווקטור u_i כולל את m הרכיבים העיקריים הראשונים ושווה ל-

$$z_i^{(m)}. \text{ בנוסף, הערך המינימאלי של שגיאת השחזור הינו } E_{\min} = \lambda_{m+1} + \dots + \lambda_d$$

הוכחה: ראו למה 23.1 בעמוד 324 של

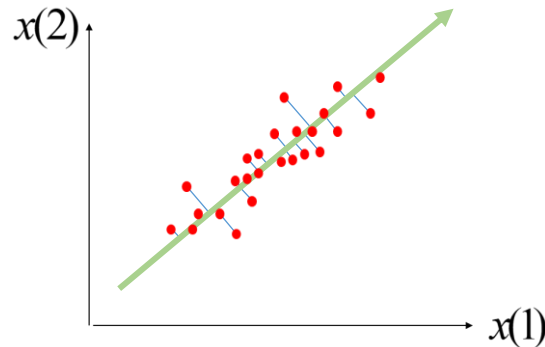
[S. Shalev-Shwartz and S. Ben David, Understanding Machine Learning, 2014.](#)

כאשר הנתונים אינם ממורכזים ($\bar{x} \neq 0$), ניתן להחסיר את הממוצע בהתחלה, ולהוסיף אותו בסוף, $\hat{x}_i = \bar{x} + V_m V_m^T (x_i - \bar{x})$ ונקבל את אותה שגיאת השחזור. באופן דומה ניתן להראות שסכמה זו מביאה למינימום את שגיאת השחזור עבור סכמת הפחתת-מימד ושחזור הלינארית באה (הכוללת הפעם גם איברי היסט)

$$u_i = A(x_i - a) \in \mathbb{R}^m, \quad \hat{x}_i = b + Bu_i \in \mathbb{R}^d$$

כאשר $b \in \mathbb{R}^d$, $a \in \mathbb{R}^d$, $B \in \mathbb{R}^{d \times m}$, $A \in \mathbb{R}^{m \times d}$

אינטואיציה גיאומטרית: הקווים הכחולים בציר למטה מראה את וקטורי שגיאת השחזור e_i (במקרה של $d=2, m=1$). שגיאת השחזור הריבועית היא סכום הנורמות הריבועיות שלהם – כלומר סכום המרחקים בריבוע של הנקודות מכיוון ההטלה (החץ הירוק). מהציר קל לראות שבחירה טובה של הכיוון היא בכיוון העיקרי, בו השונות בנתונים היא מקסימלית, כפי שנראה בסעיף הבא.



ה. תכונת השונות המירבית

עבור אוסף כלשהו של וקטורים $\{q_i\}_{i=1}^n$, נגדיר את **שונות המדגם** (sample variability) כממוצע ריבועי המרחקים מהערך ממוצע:

$$\text{Var}(q_1, \dots, q_n) = \frac{1}{n} \sum_{i=1}^n \|q_i - \bar{q}\|^2$$

כאשר הנורמה כאן היא הנורמה הרגילה (האוקלידית).

יהיו u_1, \dots, u_m וקטורים אורתונורמליים כלשהם ב- \mathbb{R}^d , ונגדיר הטלה לתת-המרחב הנפרש על ידי וקטורים אלה על ידי:

$$z_i^{(m)} = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_m^T \end{bmatrix} x_i \triangleq U_m^T x_i$$

משפט:

שונות המדגם של הוקטורים $\{z_1^{(m)}, \dots, z_n^{(m)}\}$ היא מקסימאלית כאשר כיווני ההטלה הינם m הכיוונים העיקריים:

$$u_1 = v_1, \dots, u_m = v_m$$

עבור בחירה זו, שונות המדגם המתקבלת היא

$$\text{Var}(z_1^{(m)}, \dots, z_n^{(m)}) = \lambda_1 + \dots + \lambda_m$$

הוכחה * (סקיצה, למתעניינים):

נראה תוצאה זו למקרה של $m = 1$ בלבד.

עבור וקטור יחידה כלשהו $u \equiv u_1$ נסמן $z_i \equiv z_i^{(1)} = u^T x_i$. נניח לשם פשטות כי הנתונים ממורכזים: $\bar{x} = 0$, ולכן $\bar{z} = 0$ אזי

$$\text{Var}(z_1, \dots, z_n) = \frac{1}{n} \sum_{i=1}^n (z_i)^2 = \frac{1}{n} \sum_{i=1}^n z_i z_i^T = \frac{1}{n} \sum_{i=1}^n u^T x_i x_i^T u = u^T P_n u$$

מתוצאה סטנדרטית מאלגברה לינארית נקבל

$$\max_{\|u\|=1} u^T P_n u = \max_{\|u\|=1} (Vu)^T \Lambda (Vu) = \max_{\|z\|=1} z^T \Lambda z = \lambda_{\max}(P_n) = \lambda_1$$

כאשר מקסימום זה מתקבל עבור הווקטור העצמי המתאים, $u = v_1$. (ניתן לקבל תוצאה זו כפתרון בעיית אופטימיזציה מאולצת תוך שימוש בכופל לגראנז', או על ידי שימוש בפרוק האורתונורמלי $(P_n = \sum_{k=1}^d \lambda_k v_k v_k^T)$.

□

תרגיל: הראו כי מטריצת השונות האמפירית של הרכיבים העיקריים היא אלכסונית ("מולבנת"):

$$\frac{1}{n} \sum_{i=1}^n (z_i^{(m)} - \bar{z}^{(m)})(z_i^{(m)} - \bar{z}^{(m)})^T = \text{diag}\{\lambda_1, \dots, \lambda_m\}$$

פתרון: כזכור, $z_i^{(m)} = V_m^T x_i$, ולכן גם $\bar{z}^{(m)} = V_m^T \bar{x}$. מטריצה זו שווה ל-

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (V_m^T x_i - V_m^T \bar{x})(V_m^T x_i - V_m^T \bar{x})^T \\ &= V_m^T \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \right) V_m \\ &= V_m^T P_n V_m = V_m^T V \Lambda V^T V_m = \text{diag}\{\lambda_1, \dots, \lambda_m\} \end{aligned}$$

כאשר בשיויון האחרון השתמשנו בכך ש- $V^T V = I_{d \times d}$ ולכן $V_m^T V$ מכילה את m השורות הראשונות של $I_{d \times d}$.

יחס השונות: מהתרגיל לעיל, נובע שעבור בחירה של כל d הרכיבים העיקריים ($m = d$), נסמן בקיצור $z_i = z_i^{(d)}$ ונקבל

$$\text{Var}(z_1, \dots, z_n) = \lambda_1 + \dots + \lambda_d$$

כי שונות המדגם שווה לעקבה (סכום האלכסון) של מטריצת השונות האמפירית. ניתן לראות שזו גם השונות של סדרת המידע המקורי:

עקב $z_i = V^T x_i$ כאשר $V V^T = I$, נקבל $\|z_i\|^2 = z_i^T z_i = x_i^T V V^T x_i = \|x_i\|^2$. באופן דומה,

$$\text{Var}(z_1, \dots, z_n) = \frac{1}{n} \sum_{i=1}^n \|z_i - \bar{z}\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i - \bar{x}\|^2 = \text{Var}(x_1, \dots, x_n)$$

לפיכך: היחס בין שונות המדגם של הווקטורים $\{z_1^{(m)}, \dots, z_n^{(m)}\}$ במימד המופחת $m < d$ לשונות המדגם של הסדרה המקורית $\{x_1, \dots, x_n\}$ הינו

$$g_m \triangleq \frac{\lambda_1 + \dots + \lambda_m}{\lambda_1 + \dots + \lambda_d}$$

ככל שיחס זה קרוב ל-1, אנו מתקרבים לשונות המלאה של הסדרה המקורית.

1. בחירת המימד m :

לצורך בחינה גראפית של הנתונים, נוכל להסתפק במימד נמוך ($m \geq 2$). כאשר מבצעים הורדת מימד כפעולה מקדימה לצורך הורדת רכיבים מיותרים וכפילות במידע (קורלציות), אנו מעוניינים בשמירת מרבית השונות במידע המקורי. נזכור את יחס השונות:

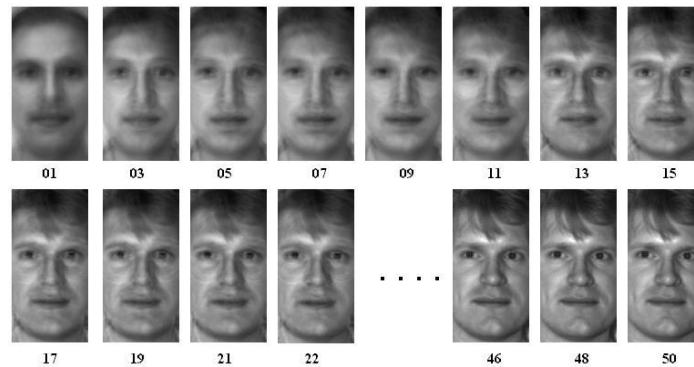
$$g_m = \frac{\lambda_1 + \dots + \lambda_m}{\lambda_1 + \dots + \lambda_d}$$

כזכור מחלק ד', שגיאת השחזור הריבועית של PCA עם m רכיבים PCA (כלומר, של המשערך $\hat{x}_i = \bar{x} + V_m V_m^T (x_i - \bar{x})$) היא $E_{\min} = \lambda_{m+1} + \dots + \lambda_d$, ולכן מתקיים גם הקשר

$$1 - g_m = \frac{E_{\min}}{\text{Var}(x_1, \dots, x_n)} = \frac{E_{\min}}{E(\bar{x})}$$

שזוהי שגיאת השחזור היחסית: היחס בין שגיאת השחזור של PCA ביחס לשונות המדגם המקורית (שזו גם שגיאת השחזור של המשערך הטריטוריאלי $\hat{x} = \bar{x}$).

לכן, כדאי לבחור את m כך ש- g_m זה עובר קרוב ל-1: למשל אם נבחר $g_m \geq 0.95$, אז נצפה לקבל 5% שגיאה יחסית בשחזור, במובן השגיאה הריבועית. הערך המדויק של הסף יהיה תלוי בכמה רכיבים אנחנו מוכנים לשמור בשביל לקבל שגיאת שחזור קטנה. לפעמים צריך לבחור את m לפי קריטריונים אחרים שאינם בהכרח שגיאה ריבועית. לדוגמא, נתון סט תמונות של פרצופים, ואנחנו רוצים להשתמש ב-PCA כדי לדחוס את הייצוג של כל תמונה. הגדלה של m תביא לשיפור באיכות:



אולם מעבר לשלב מסוים, לא רואים שיפור משמעותי באיכות. לכן, נבחר את m המינימלי עבור השחזור של כל תמונה נראה "טוב מספיק" בעין (וקריטריון זה לא קשור לשגיאה ריבועית, אך לא בהכרח זהה לה, עקב תכונות הראייה האנושית). למידע נוסף על דוגמא זו, ראו [קישור](#).

3.3 אישכול (Clustering) – אלגוריתם K-means

אישכול פירושו חלוקת אוסף נתונים $D_X = \{x_1, \dots, x_n\}$ לתת-קבוצות, כך שלחברים בכל תת-קבוצה יש קשר או קרבה. לטכניקה זו שימוש נרחב לצורך ניתוח והבנת נתונים. קיימות דוגמאות רבות בכל תחומי המדע וההנדסה: ביולוגים מעוניינים באישכול גנים לצורך זיהוי גנים בעלי מבנה או פעולה דומים, אסטרונומים מעוניינים באישכול כוכבים בעלי ספקטרום דומה, אנשי שיווק מעוניינים לזהות טיפוסים לקוחות, וכולי. אישכול שימושי גם לצורך בחירת מודל סטטיסטי מתאים לנתונים, וככלי לזיהוי מאפיינים חשובים של הנתונים אשר מפרידים בין פריטים או קבוצות שונות.

קיימות מספר רב של טכניקות וגישות שונות לאישכול. פה נתאר בקצרה מספר מאפיינים של הבעיה, נדגים גישה היררכית לאישכול, ונתאר את אלגוריתם K-means, שהוא שימושי במיוחד למידע מספרי-וקטורי.

א. מדדי קירבה וחלוקה

נקודת ההתחלה היא קיום מדד מרחק כלשהו בין פריטי הנתונים. נסמן ב- $d(x_i, x_j)$ את המרחק בין הפריטים x_i, x_j . המרחק הוא לרוב גודל אי-שלילי וסימטרי, אולם אינו חייב להיות נורמה או מטריקה. כתלות בסוג המידע הוא יכול להיות מורכב, כגון מרחק בין סדרות באורך שונה, בין אותות דיבור שונים, בין סרטי וידאו, וכולי.

עבור מידע וקטורי (וקטורים בגודל קבוע עם איברים מספריים), מדד מרחק נפוץ הוא המרחק הריבועי, שנחזור אליו בהמשך:

$$d(x_i, x_j) = \|x_i - x_j\|^2$$

כאשר הנורמה כאן היא הנורמה הרגילה (האוקלידית). נניח כי ברצוננו לחלק את הנתונים ל- $K > 1$ קבוצות או מחלקות (K מספר נתון, שעל בחירתו לא נתעכב פה). נמספר מחלקות אלו כ- $1, \dots, K$. עלינו לשייך כל פריט מידע x_i לאחת מהמחלקות, שתסומן $C(i)$. למשל: $C(4) = 2$. פירושו כי x_4 שויך למחלקה 2. $C = (C(i), i = 1, \dots, n)$ הם משתני ההחלטה שלנו.

מדד מרחק סביר עבור קבוצת פריטים הוא סכום המרחקים בין איברי הקבוצה, מנורמל במספר האיברים בקבוצה. לכל מחלקה $1 \leq k \leq K$ נגדיר

$$W_k(C) = \frac{1}{2n_k} \sum_{i,j:C(i)=C(j)=k} d(x_i, x_j)$$

אם נסכם על פני כל המחלקות, נקבל את ממד המרחק הכולל עבור שיוך נתון C :

$$W(C) = \sum_{k=1, \dots, K} W_k(C)$$

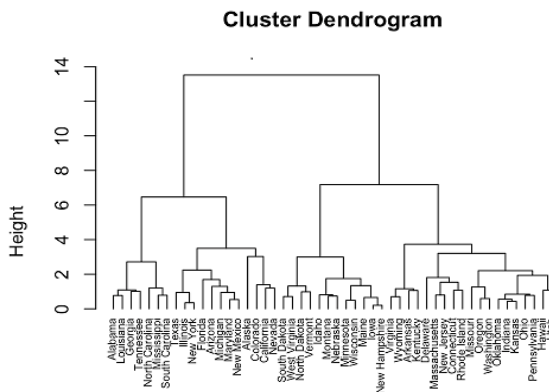
ניתן עתה להגדיר את בעיית האשכול כבעיית אופטימיזציה: עלינו למצוא שיוך C שמביא למינימום את $W(C)$.

לרוע המזל, לא קיים אלגוריתם יעיל לפתרון בעיה זו. בדיקת כל החלוקות בוודאי אינה אפשרית כיוון שמספרן מעריכי ב- n . לפיכך יש להסתפק בגישות ואלגוריתמים תת-אופטימליים.

ב. אישכול צובר (Agglomerative Clustering)

זו גישה נפוצה, שנביא פה כדוגמה לגישה היררכית. בגישה זו בונים את המחלקות בהדרגה (bottom up), כאשר בתחילה כל פריט נמצא במחלקה נפרדת משלו. בכל שלב מאחדים שתי מחלקות, עד שכל הפריטים אוחדו למחלקה אחת.

את התוצאה ניתן להמחיש ב"דיגרמת דנדוגרם", כדוגמת המוראית.



הקבוצות המאוחדות בכל שלב הן השתיים הקרובות ביותר. לשם כך יש להגדיר מדד מרחק מתאים בין קבוצות של פריטים. מספר אפשרויות לכך, המובילות לתוצאות שונות:

מרחק ממוצע (Group Average):

$$d(A, B) = \frac{1}{n(A)n(B)} \sum_{i \in A, j \in B} d(x_i, x_j)$$

השכן הקרוב (Single Linkage):

$$d(A, B) = \min_{i \in A, j \in B} d(x_i, x_j)$$

השכן הרחוק (Complete Linkage):

$$d(A, B) = \max_{i \in A, j \in B} d(x_i, x_j)$$

כאשר, כמקודם, $d(x_i, x_j)$ יכול להיות מדד מרחק כלשהו (כמו, למשל, המרחק הריבועי).

ג. אלגוריתם K-means

אלגוריתם נפוץ זה מיועד למקרה של וקטורים נומריים ומדד מרחק ריבועי :

$$d(x_i, x_j) = \|x_i - x_j\|^2$$

תרגיל: הראו כי

$$W_k(C) \triangleq \frac{1}{2n_k} \sum_{i,j:C(i)=C(j)=k} d(x_i, x_j) = \sum_{i:C(i)=k} \|x_i - \mu_k\|^2$$

כאשר n_K מספר האיברים במחלקה k , ו- μ_k הוא הממוצע : $\mu_k = \frac{1}{n_k} \sum_{i:C(i)=k} x_i$

פתרון:

$$\begin{aligned} \sum_{i:C(i)=k} \|x_i - \mu_k\|^2 &= \sum_{i:C(i)=k} \left\| x_i - \frac{1}{n_k} \sum_{j:C(j)=k} x_j \right\|^2 = \\ \sum_{i:C(i)=k} \left\| \frac{1}{n_k} \sum_{j:C(j)=k} (x_i - x_j) \right\|^2 &= \frac{1}{n_k^2} \sum_{i,j,r:C(i)=C(j)=C(r)=k} (x_i - x_r)^T (x_i - x_j) \\ &= \frac{1}{n_k^2} \sum_{i,j,r:C(i)=C(j)=C(r)=k} (x_i^T x_i - x_i^T x_j - x_r^T x_i + x_r^T x_j) \\ &= \frac{1}{n_k} \sum_{i,j:C(i)=C(j)=k} (x_i^T x_i - x_i^T x_j - x_i^T x_j + x_j^T x_j) = \frac{1}{n_k} \sum_{i,j:C(i)=C(j)=k} (x_i^T x_i - x_i^T x_j) \\ &= \frac{1}{2n_k} \sum_{i,j:C(i)=C(j)=k} (x_i^T x_i - 2x_i^T x_j + x_j^T x_j) = \frac{1}{2n_k} \sum_{i,j:C(i)=C(j)=k} \|x_i - x_j\|^2 \end{aligned}$$

לפיכך, אנו מעוניינים להביא למינימום את מדד המרחק הכולל :

$$W(C) = \sum_{k=1}^K W_k(C) = \sum_{k=1}^K \sum_{i:C(i)=k} \|x_i - \mu_k\|^2$$

אלגוריתם K-means הוא אלגוריתם איטרטיבי שמטרתו להביא למינימום את $W(C)$.

האלגוריתם מתייחס לממוצעים μ_k כמשתנים נפרדים, ומבצע צעדים מתחלפים של :

1. מינימיזציה על פני השיוך C (כאשר הממוצעים μ_k קבועים) – שלב 1 באלגוריתם.

2. מינימיזציה על פני הממוצעים (כאשר C קבוע) – שלב 2 באלגוריתם

האלגוריתם (K-means clustering) :

• איתחול: בחירת מרכזים $k = 1, \dots, K, \mu_k \in \mathbb{R}^d$

• חזרו על הצעדים הבאים עד להתכנסות (אין עוד שינוי בשיוך) :

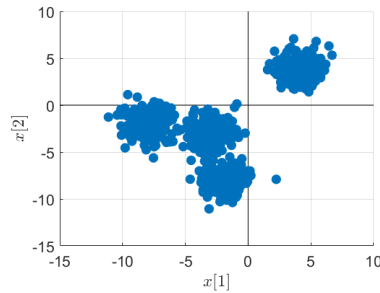
1. חשבו את השיוך $C(i)$ של כל פריט x_i , בהתאם לממוצע הקרוב ביותר :

$$C(i) = \operatorname{argmin}_{k=1,\dots,K} \|x_i - \mu_k\|^2$$

2. חשבו את הממוצעים בכל מחלקה לפי השיוך הקיים :

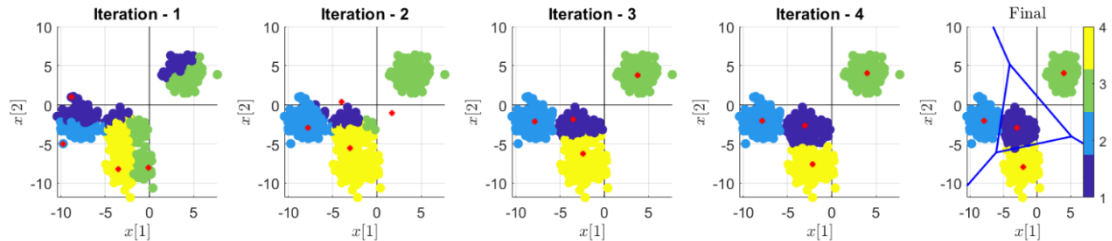
$$\mu_k = \frac{1}{n_k} \sum_{i:C(i)=k} x_i, \quad k = 1, \dots, K$$

התכנסות : ניתן לראות כי כל שלב באלגוריתם שבו מתבצע שינוי מקטין את מדד המרחק $W(C)$, ולפיכך מובטחת התכנסות במספר צעדים סופי. עם זאת כללית ההתכנסות תהיה למינימום מקומי ולא למינימום הגלובלי. כדי להדגים זאת, נפעיל את האלגוריתם על סט הנתונים הבא :

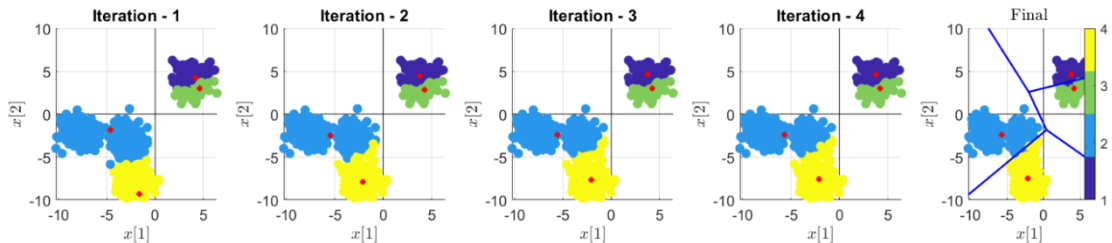


נריץ את האלגוריתם פעמיים עם אתחולים שונים, עבור בחירה של $K=4$. נסמן בצבעים שונים את השייך של כל נקודה (עם צבע שונה לכל מחלקה), בנקודה אדומה את המרכז של כל מחלקה, ופתרון הסופי נסמן גם את איזורי ההחלטה לשייך לכל מחלקה ע"י קווים כחולים.

עבור אתחול אחד נקבל פתרון מוצלח :



בעוד שעבור אתחול שני נקבל פתרון שנראה פחות מוצלח (כי, למשל, מה שנראה כמו אשכול אחד מימין למעלה פוצל לשתי אשכולות שלא לצורך) :



יש אלגוריתמים רבים לבחירת אתחול, כאשר אופציה פופולרית היא $K\text{-means}++$. בנוסף, בחירה שונה של K תשפיע בצורה משמעותית על הפתרון המתקבל.

למשל, עבור $K=2$, נקבל :

