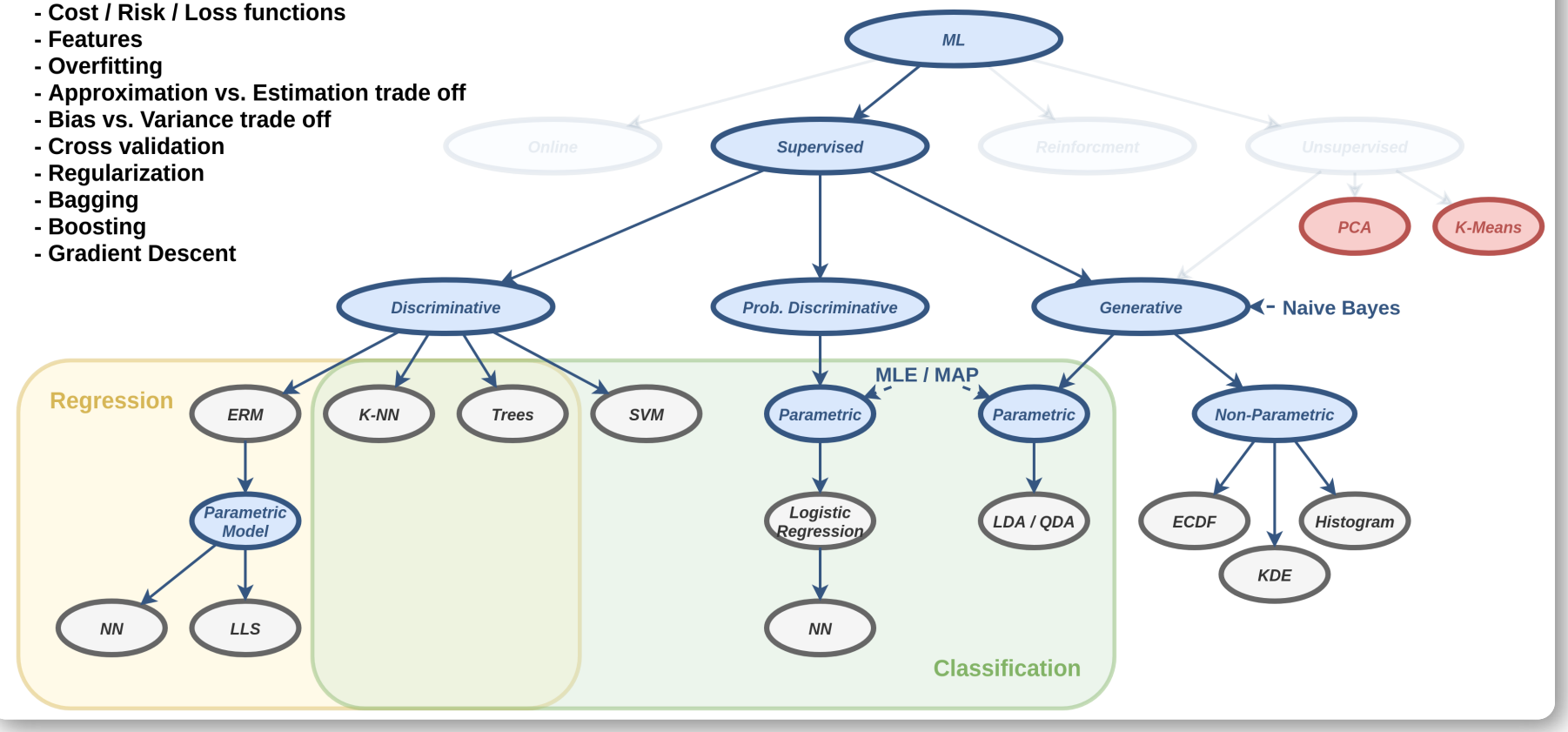
PCA and K- - 12 הרצאה
Means

PDF

מה נלמד היום

Subjects Covered in this Course

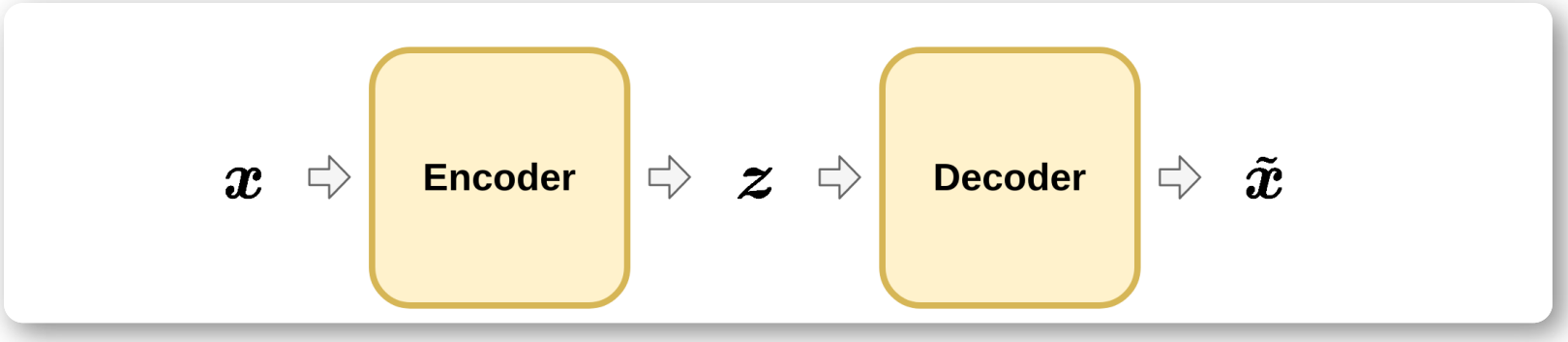
General concepts:
- Cost / Risk / Loss functions
- Features
- Overfitting
- Approximation vs. Estimation trade off
- Bias vs. Variance trade off
- Cross validation
- Regularization
- Bagging
- Boosting
- Gradient Descent



למידה לא מודרכת (Unsupervised Learning)

- שם כולל למגוון של בעיות בהם אנו מנסים בהינתן מדגם, ללמוד את התכונות של הדגימות או של המדגם כולו.
- המדגם יכיל אוסף של דגימות (x), ללא תווית (y).
- דוגמאות:
  - אשכול (חלוקה לקבוצות).
  - מציאת ייצוג "נוח" יותר של הדגימות.
  - דחיסה.
  - זיהוי אנומליות.
  - למידת הפילוג של הדגימות.

מערכת Encoder-Decoder



דוגמאות לשימושים במערכת encoder-decoder הינם:

- **דחיסה: נרצה ש z יהיה קטן ככל האפשר.**
- **תקשורת: נרצה ש z יהיה כמה שפחות רגיש לרעשים.**
- **הצפנה: נרצה שפעולת השחזור של x תהיה כמה שיותר קשה ללא ה decoder המתאים.**

$\tilde{x}$ נקרא השחזור של x. בחלק מהמערכות ניתן להגיע לשיחזור מושלם, $\tilde{x} = x$, ובחלק מהמערכות לא.

Principle Component Analysis ((PCA

ב PCA ננסה לבנות מערכת encoder-decoder שבה:

1. אנו מגבילים את האורך של הוקטור z.
2. אנו דורשים שה encoder וה decoder יהיו פונקציות אפיניות (affine = linear + offset).
3. התוחלת של שגיאת השחזור הריבועית $\mathbb{E}\left[\|\tilde{\mathbf{x}}-\mathbf{x}\|_2^2\right]$ היא מינימאלית.

נחליף את התוחלת בתוחלת אמפירית על מדגם.

Principle Component Analysis ((PCA

D האורך של x ו K האורך של הוקטור z כאשר מתקיים כי $K \leq D$.

נרצה למצוא encoder:

$$\boldsymbol{z} = T_1\boldsymbol{x} + \boldsymbol{b}_1$$

ו decoder מהצורה של:

$$\tilde{\boldsymbol{x}} = T_2\boldsymbol{z} + \boldsymbol{b}_2$$

אשר ממזערים את התוחלת האמפירית של שגיאת השחזור הריבועית:

$$\underset{T_1,T_2,\boldsymbol{b}_1,\boldsymbol{b}_2}{\arg\min}\frac{1}{N}\sum_{i=1}^{N}\|\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_2^2$$

שימושים

דוגמאות למקרים שבהם נרצה לבצע הורדת מימד (dimensionality reduction):

1. **בחירת מאפיינים לבעיות supervised learning**
2. **ויזואליזציה**
3. **דחיסה**

הפתרון לבעיית האופטימיזציה

מסתבר שיש מספר רב של פתרונות. ניתן לבחור את הפרמטרים כך שיקיימו את האילוצים:

$$\begin{aligned} \boldsymbol{b}_1 &= -T_1\boldsymbol{\mu} \\ \boldsymbol{b}_2 &= \boldsymbol{\mu} \\ T_1 &= T_2^\top = T^\top \\ T^\top T &= I \end{aligned}$$

כאשר $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}^{(i)}$**.**

הערה: שימו לב ש-$T \in \mathbb{R}^{D\times K}$ **כך שמתקיים כי** $T^\top T \in \mathbb{R}^{K\times K} = I_K$ **כאשר** I_K **היא מטריצת היחידה. בנוסף, מתקיים** $TT^\top \in \mathbb{R}^{D\times D}$ **והיא לא שווה בהכרח ל-**I_D**.**

הפתרון לבעיית האופטימיזציה

לדוגמה, עבור המיפוי הבא

$$\boldsymbol{z} = T_1\boldsymbol{x} + \boldsymbol{b}_1 \quad \tilde{\boldsymbol{x}} = T_2 z + \boldsymbol{b}_2 \quad T_1 \in \mathbb{R}^{K\times D},\, T_2 \in \mathbb{R}^{D\times K}$$

איברי ההטיה b_1 ו-b_2 יכולים להיקבע ע"י הדרישות

$$E[\boldsymbol{z}] = 0 \quad \Rightarrow \quad b_1 = -T_1\boldsymbol{\mu}$$

ו-

$$E\left[\tilde{\boldsymbol{x}}\right] = E\left[\boldsymbol{x}\right] \quad \Rightarrow \quad b_2 = E\left[x\right] = \boldsymbol{\mu}$$

הפתרון לבעיית האופטימיזציה

הטרנספורמציות במקרה זה הופכות להיות:

$$\boldsymbol{z} = T^\top(\boldsymbol{x} - \boldsymbol{\mu})$$
$$\tilde{\boldsymbol{x}} = T\boldsymbol{z} + \boldsymbol{\mu}$$

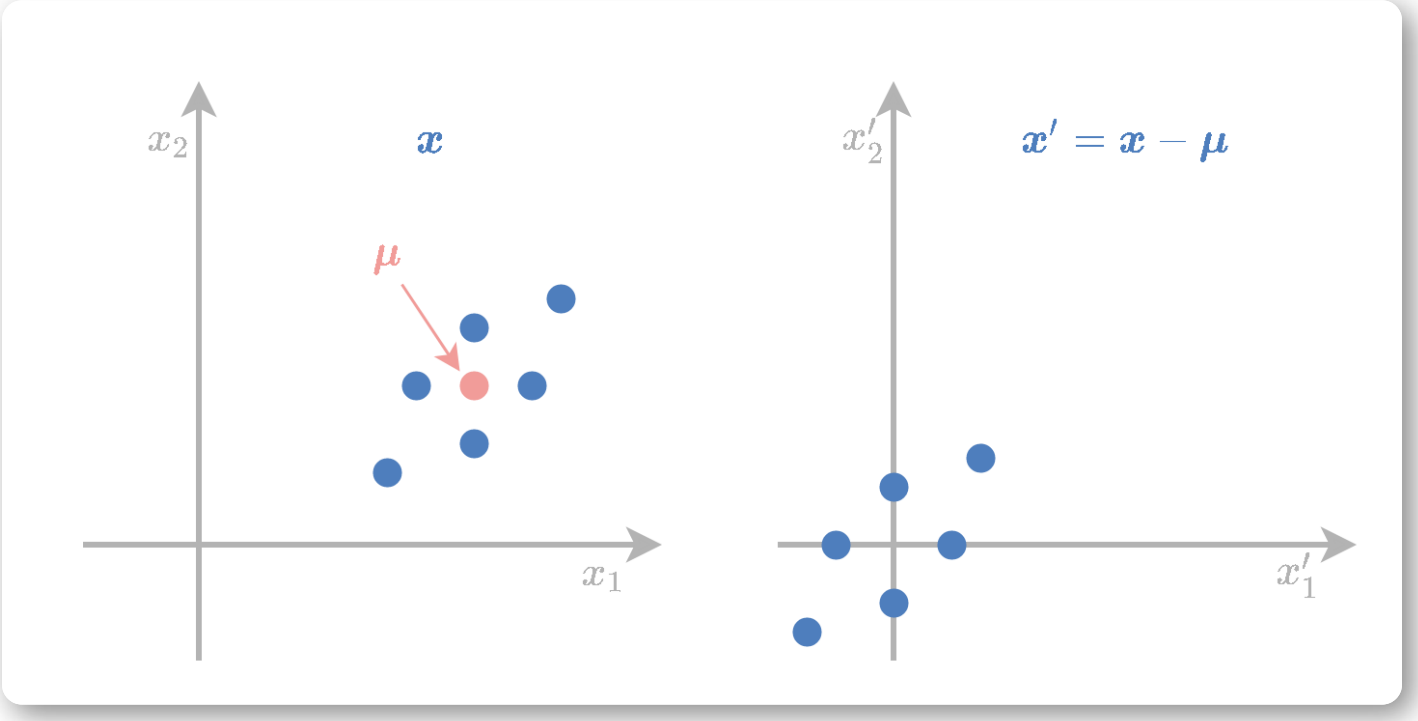
ובעיית האופטימיזציה הינה:

$$T^* = \arg\min_T \quad \frac{1}{N}\sum_{i=1}^{N}\|\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_2^2$$
$$\text{s.t.} \quad T^\top T = I$$

$$T^* = \arg\min_T \quad \frac{1}{N}\sum_{i=1}^{N}\|(TT^\top - I)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu})\|_2^2$$
$$\text{s.t.} \quad T^\top T = I$$

פרשנות גיאומטרית

- ה encoder מחסר את הממוצע של x וה decoder מוסיף אותו בחזרה.
- נניח מעתה שהנתונים ממורכזים סביב האפס.



הפתרון לבעיית האופטימיזציה

הטרנספורמציות המתקבלות הינן:

$$\boldsymbol{z} = T^\top \boldsymbol{x}$$
$$\tilde{\boldsymbol{x}} = T\boldsymbol{z} = TT^\top \boldsymbol{x}$$

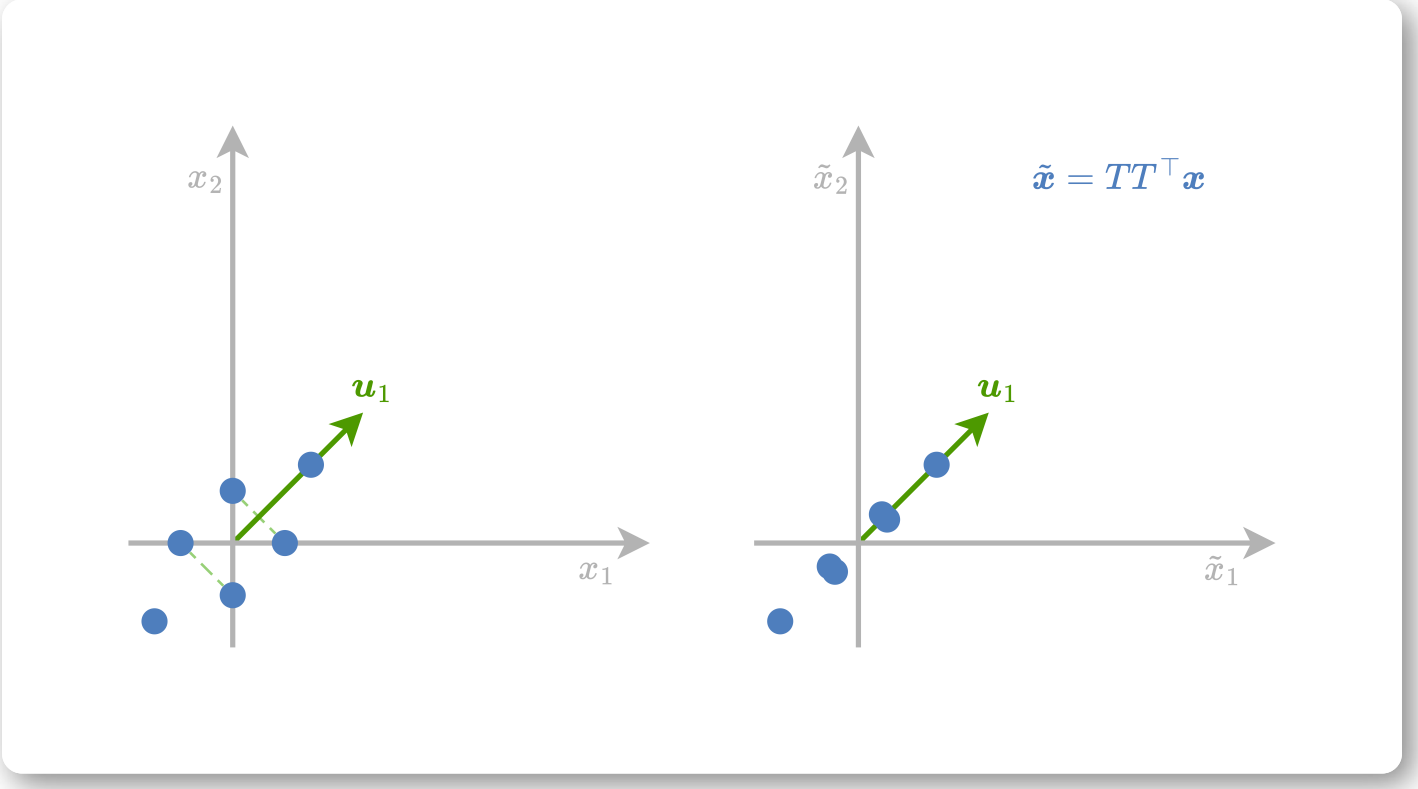
נתייחס כעת לאילוץ של $T^\top T = I$. אילוץ זה אומר שהעמודות של T צריכות להיות אורתו-נורמאליות.

נסמן את העמודות של T ב u_j:

$$T = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_K \\ | & | & & | \end{pmatrix}$$

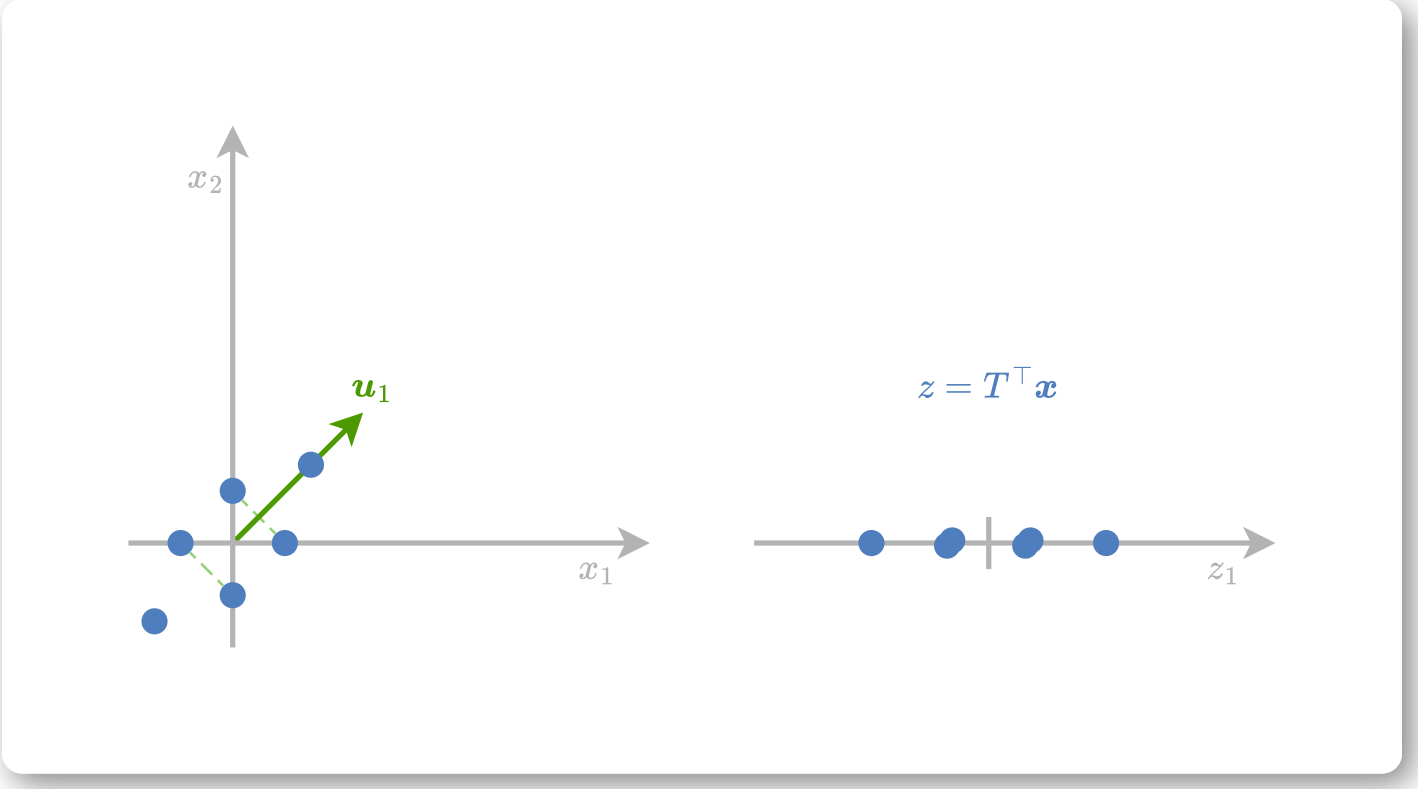
פרשנות גיאומטרית

הפעולה של $\tilde{x} = TT^\top x$ מטילה את הוקטור x על תת-המרחב הלינארי הנפרס על ידי הוקטורים u_j.



פרשנות גיאומטרית

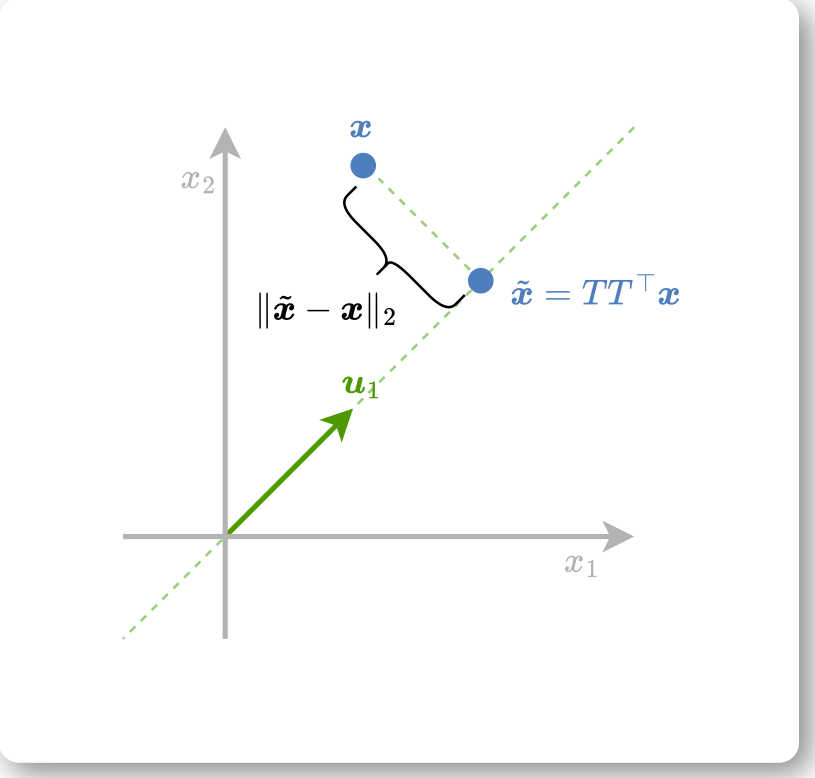
הפעולה של $z = T^\top x$ גם מטילה את x על אותו תת-מרחב, היא רק משאירה אותו במערכת הצירים של u_j:



פרשנות גיאומטרית

נסתכל כעת על המשמעות הגיאומטרית של שגיאת השחזור

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$$



בעיית האופטימיזציה היא הבעיה של מציאת תת-המרחב ממימד K אשר ההטלה של נקודות המדגם עליו הם הקרובות ביותר לנקודות המקוריות.

הבעיה השקולה

מתוך העובדה ש $T^\top T = I$ ניתן להראות ש:

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 = \|\boldsymbol{x}\|_2^2 - \|\tilde{\boldsymbol{x}}\|_2^2 = \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{z}\|_2^2$$

שכן, עבור $T^\top T = I$ מתקיים כי $\left(I - TT^\top\right)^2 = \left(I - TT^\top\right)$ לכן,

$$\begin{aligned}
\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2 &= \left\|\boldsymbol{x} - TT^\top x\right\|_2^2 \\
&= \left\|\left(I - TT^\top\right) x\right\|_2^2 \\
&= \boldsymbol{x}^\top \left(I - TT^\top\right) \boldsymbol{x} \\
&= \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{z}\|_2^2
\end{aligned}$$

ובנוסף

$$\|\tilde{\boldsymbol{x}}\|_2^2 = \|T\boldsymbol{z}\|_2^2 = \boldsymbol{z}^\top T^\top T\boldsymbol{z} = \|\boldsymbol{z}\|_2^2$$

הבעיה השקולה

מכאן שנוכל לרשום את בעיית האופטימיזציה באופן הבא:

$$T^* = \arg\min_T \quad \frac{1}{N}\sum_{i=1}^{N}\left(\|\boldsymbol{x}^{(i)}\|_2^2 - \|\boldsymbol{z}^{(i)}\|_2^2\right)$$
$$\text{s.t.} \quad T^\top T = I$$

נזכור ש $\|x\|_2^2$ והוא תכונה של הוקטורים במדגם והם אינם תלויים ב T ולכן:

$$T^* = \arg\min_T \quad -\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{z}^{(i)}\|_2^2$$
$$\text{s.t.} \quad T^\top T = I$$

הבעיה השקולה

$$T^* = \arg\min_T \quad -\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{z}^{(i)}\|_2^2$$
$$\text{s.t.} \quad T^\top T = I$$

הבעיה של מזעור שגיאת השחזור הריבועית שקולה לבעיה של מקסום הגודל $\sum_{i=1}^{N}\|\boldsymbol{z}^{(i)}\|_2^2$.

גדול זה מכונה ה variance של אוסף הוקטורים $\{\boldsymbol{z}^{(i)}\}_{i=1}^{N}$.

הפתרון

נגדיר:

$$X = \begin{pmatrix} - & \boldsymbol{x}'^{(1)} & - \\ - & \boldsymbol{x}'^{(2)} & - \\ & \vdots & \\ - & \boldsymbol{x}'^{(N)} & - \end{pmatrix} = \begin{pmatrix} - & (\boldsymbol{x}^{(1)} - \boldsymbol{\mu})^\top & - \\ - & (\boldsymbol{x}^{(2)} - \boldsymbol{\mu})^\top & - \\ & \vdots & \\ - & (\boldsymbol{x}^{(N)} - \boldsymbol{\mu})^\top & - \end{pmatrix}$$

ומטריצת ה covariance האמפירית של $\mathbf{x}$ תהיה:

$$P = X^\top X$$

הפתרון

$$P = X^\top X$$

P ממשית וסימטרית ולכן מובטח כי ניתן לפרק אותה באופן הבא:

$$P = U \Lambda U^\top$$

כאשר U היא מטריצה הוקטורים עצמיים:

$$U = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_D \\ | & | & & | \end{pmatrix}$$

ו Λ היא מטריצה הערכים העצמיים:

הפתרון

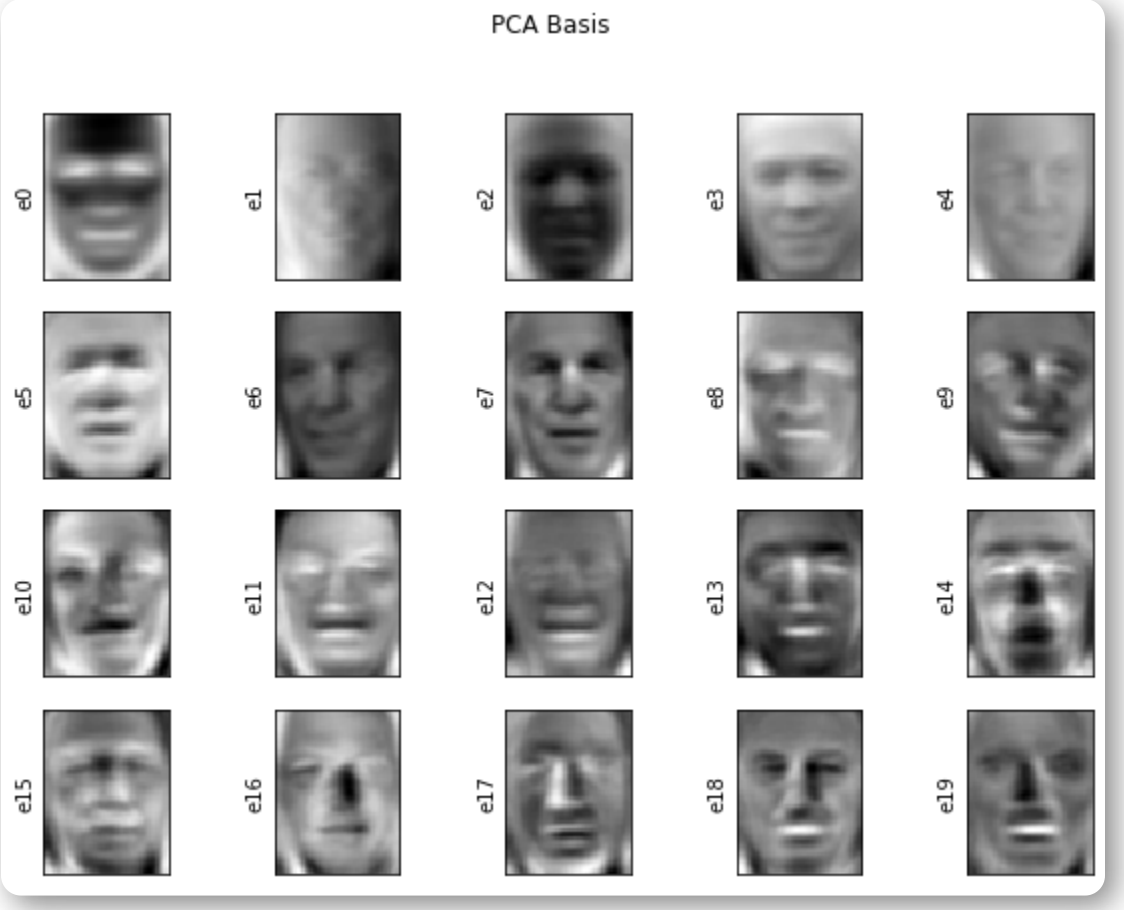
T תהיה מטריצה אשר העמודות שלה הם K העמודות הראשונות במטריצה U:

$$T = \begin{pmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_K \\ | & | & & | \end{pmatrix}$$

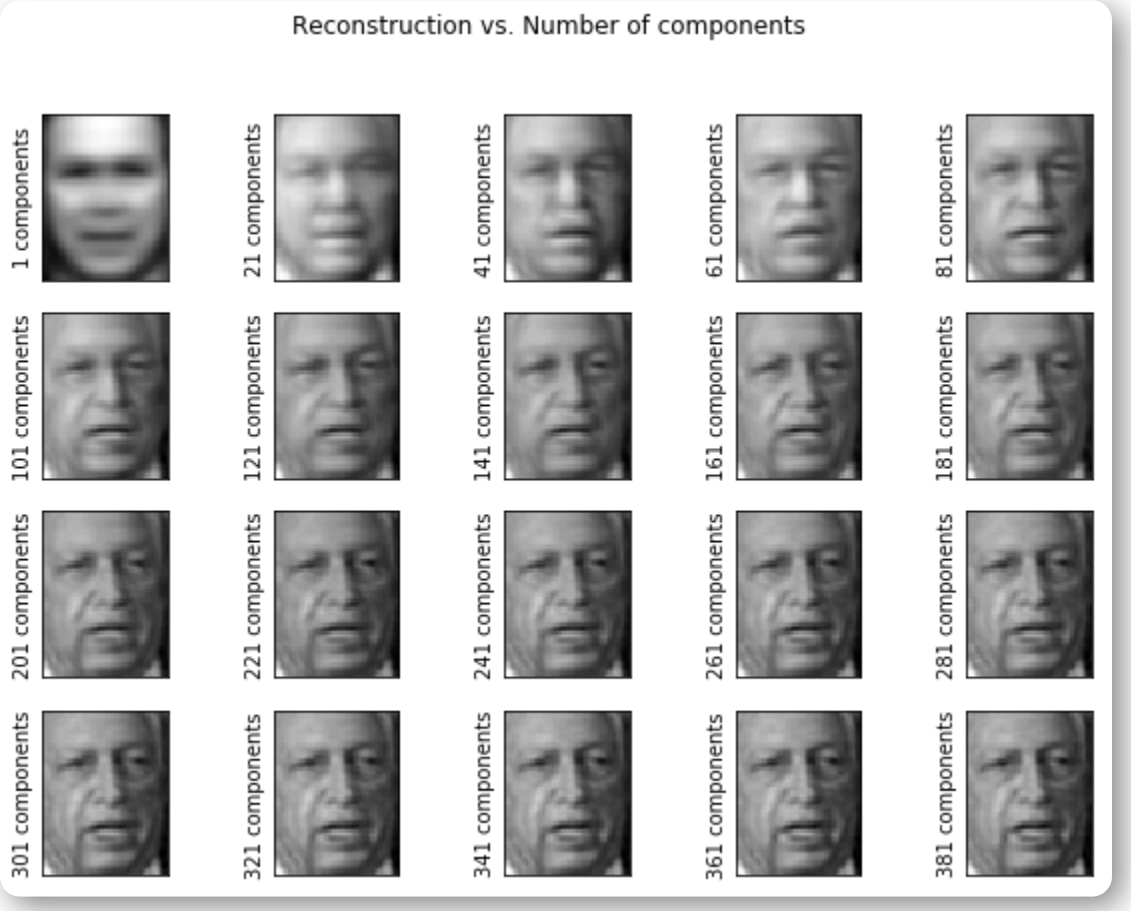
- הכיוונים $u^{(j)}$ מכונים ה**כיוונים העיקריים**
- הרכיבים של הוקטור z מכונים ה**רכיבים העיקריים (principal components)**.

דוגמא

פירוק תמונות של פנים לכיוונים העיקריים:



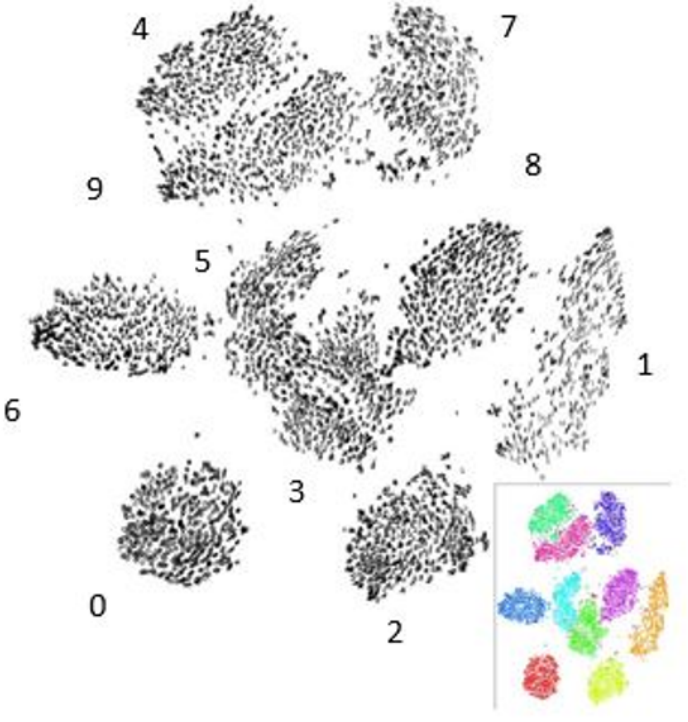
תמונה המשוחזרת עבור ערכים שונים של K:



הרחבות לא לינאריות

קיימות הרחבות לא לינאריות רבות ל-PCA. נידונות בקורס עיבוד וניתוח מידע (ענ"ם).

$D = 28 \times 28 = 784$

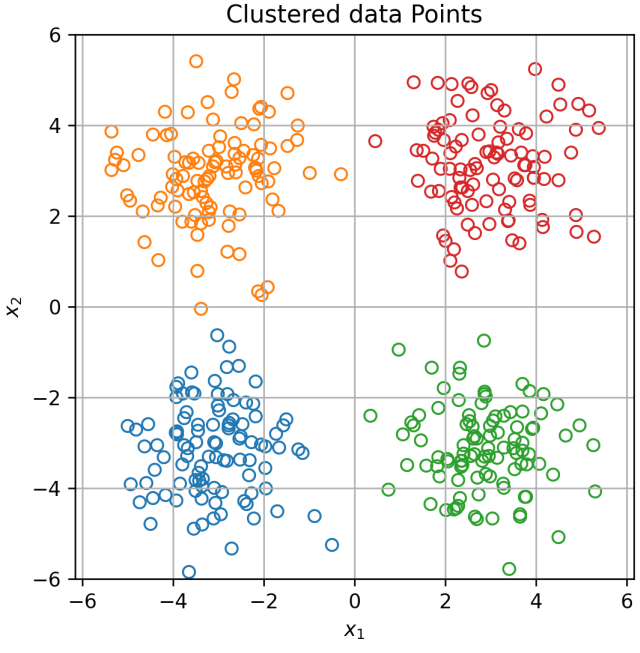


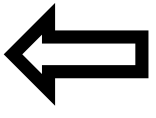
$K = 2$

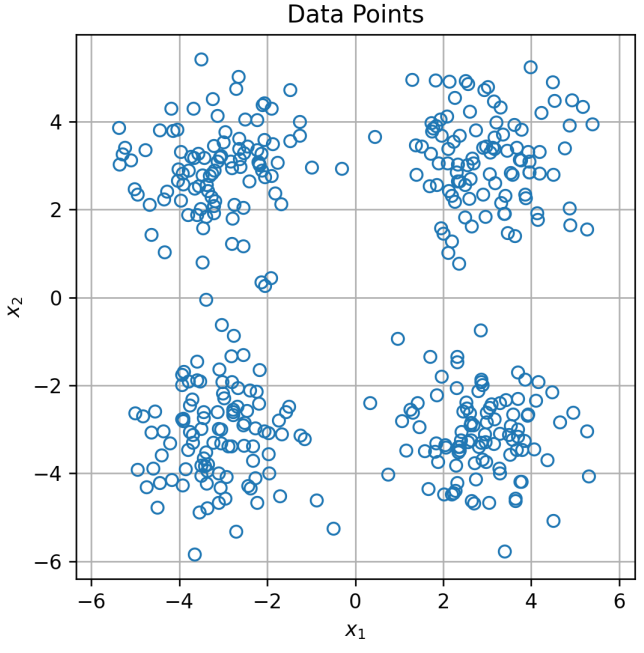
הפעלת אלגוריתם tSNE על MNIST.

אשכול

באלגוריתמי אשכול ננסה לחלק אוסף של פרטים לקבוצות המכונים אשכולות (clusters), כאשר לכל קבוצה איזשהן תכונות דומות. כמובן, בממדים גבוהים לא רואים זאת בעין.







אשכול

2 דוגמאות למקרים שבהם נרצה לאשכל:

1. **על מנת לבצע הנחות על אחד מהפרטים באשכול על סמך פרטים אחרים באשכול.**
   לדוגמא: להציע ללקוח מסויים בחנות אינטרנט מוצרים על סמך מוצרים שקנו לקוחות אחרים באשכול שלו.

2. **לתת טיפול שונה לכל אשכול.**
   לדוגמא: משרד ממשלתי שרוצה להפנות קבוצות שונות באוכלוסיה לערוצי מתן שירות שונים: אפליקציה, אתר אינטרנט, נציג טלפוני או הפניה פיסית למוקד שירות.

K-Means

K-Means הוא אלגוריתם אשכול אשר מנסה לחלק את הדגימות במדגם ל K קבוצות על סמך המרחק בין הדגימות.

סימונים

- K **- מספר האשכולות (גודל אשר נקבע מראש).**
- $\mathcal{I}_k$ **- אוסף האינדקסים של האשכול ה-k.**
  לדוגמא: $\mathcal{I}_5 = \{3, 6, 9, 13\}$
- $|\mathcal{I}_k|$ **- גודל האשכול ה-k (מספר הפרטים בקבוצה)**
- $\{\mathcal{I}_k\}_{k=1}^{K}$ **- חלוקה מסוימת לאשכולות**

בעיית האופטימיזציה

K-Means מנסה למצוא את החלוקה לאשכולות אשר תמזער את המרחק הריבועי הממוצע בין כל דגימה לכל שאר הדגימות שאיתו באותו האשכול:

$$\arg\min_{\{\mathcal{I}_j\}_{k=1}^{K}} \frac{1}{N} \sum_{k=1}^{K} \frac{1}{2|\mathcal{I}_k|} \sum_{i,j \in \mathcal{I}_k} \|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}\|_2^2$$

שאלה: האם פונקציית מרחק ריבועית תמיד מתאימה?

הבעיה השקולה

נגדיר את מרכז המסה:

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\boldsymbol{x}^{(i)}$$

ניתן להראות כי בעיית האופטימיזציה המקורית, שקולה לבעיה של מיזעור המרחק הממוצע של הדגימות ממרכז המסה של האשכול:

$$\arg\min_{\{\mathcal{I}_j\}_{k=1}^K}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\|_2^2$$

$$\sum_{i,j\in\mathcal{I}_k}^{K}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}\right\|_2^2=\sum_{i,j\in\mathcal{I}_k}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k+\boldsymbol{\mu}_k-\boldsymbol{x}^{(j)}\right\|_2^2$$

$$=\sum_{i,j\in\mathcal{I}_k}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right\|_2^2+\sum_{i,j\in\mathcal{I}_k}\left\|\boldsymbol{x}^{(j)}-\boldsymbol{\mu}_k\right\|_2^2-2\sum_{i,j\in\mathcal{I}_k}\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right)^\top\left(\boldsymbol{x}^{(j)}-$$

$$=2\left|\mathcal{I}_k\right|\sum_{i\in\mathcal{I}_k}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right\|_2^2-2\sum_{i\in\mathcal{I}_k}\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right)^\top\sum_{j\in\mathcal{I}_k}\left(\boldsymbol{x}^{(j)}-\boldsymbol{\mu}_k\right)$$

$$=2\left|\mathcal{I}_k\right|\sum_{i\in I_k}\left\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right\|_2^2$$

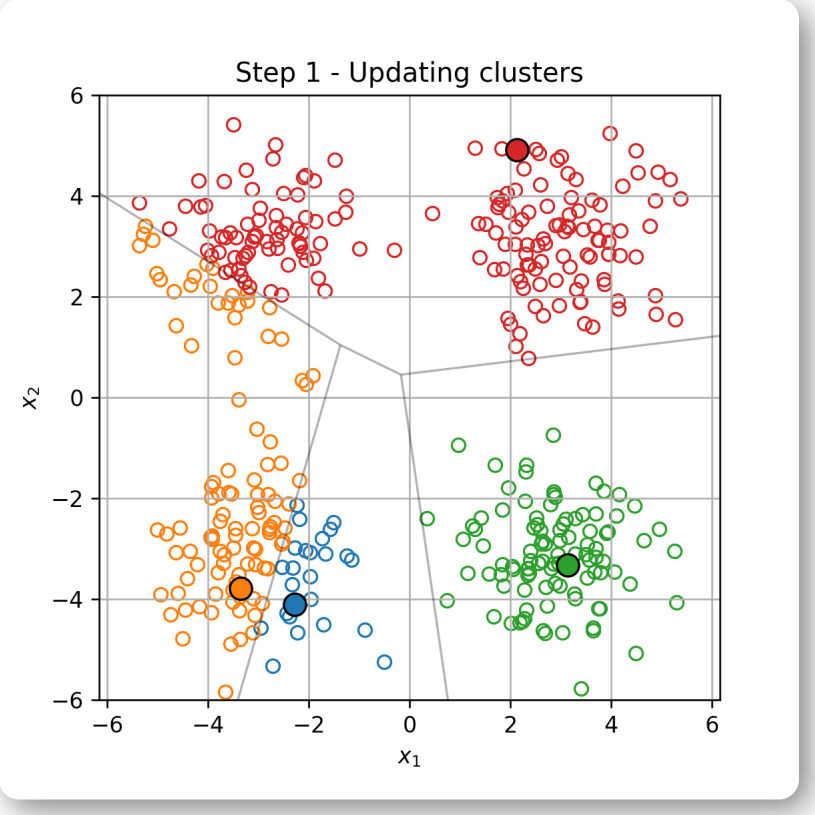
שכן:

$$\sum_{i\in I_k}\left(\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_k\right)=\left|\mathcal{I}_k\right|\cdot\frac{1}{\left|\mathcal{I}_k\right|}\sum_{i\in I_k}\boldsymbol{x}^{(i)}-\left|\mathcal{I}_k\right|\boldsymbol{\mu}_k=0$$

תכונות

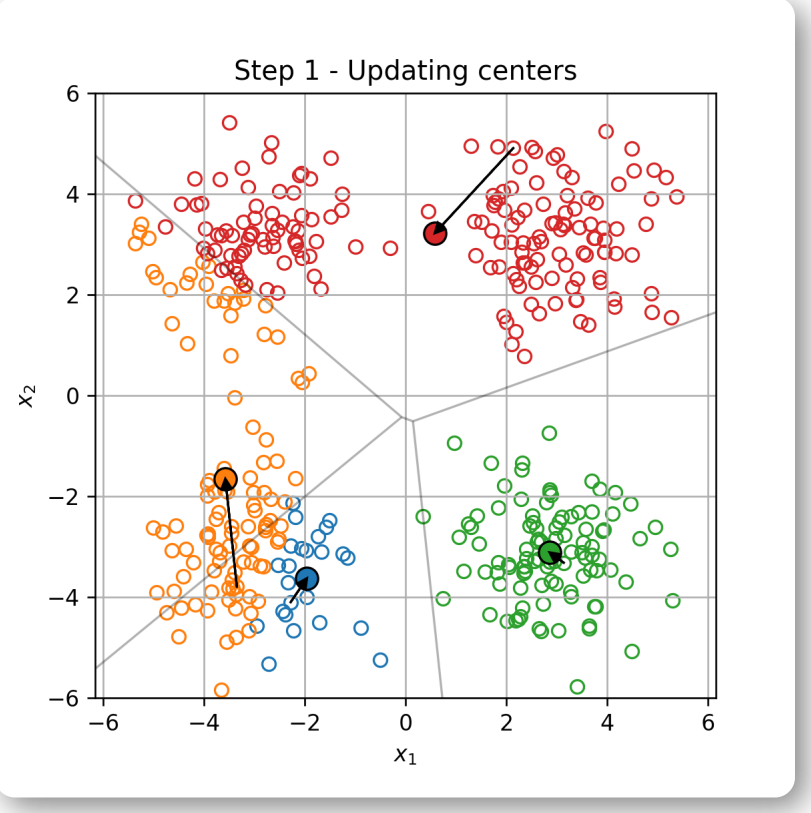
- מובטח כי פונקציית המטרה תקטן בכל צעד.
- מובטח כי האלגוריתם יעצר לאחר מספר סופי של צעדים.
- **לא** מובטח כי האלגוריתם יתכנס לפתרון האופטימאלי. בפועל במרבית מתכנס לפתרון קרוב מאד לאופטימאלי.
- אתחולים שונים יכולים להוביל לתוצאות שונות.

אתחול (וחלוקה ראשונית לאשכולות):

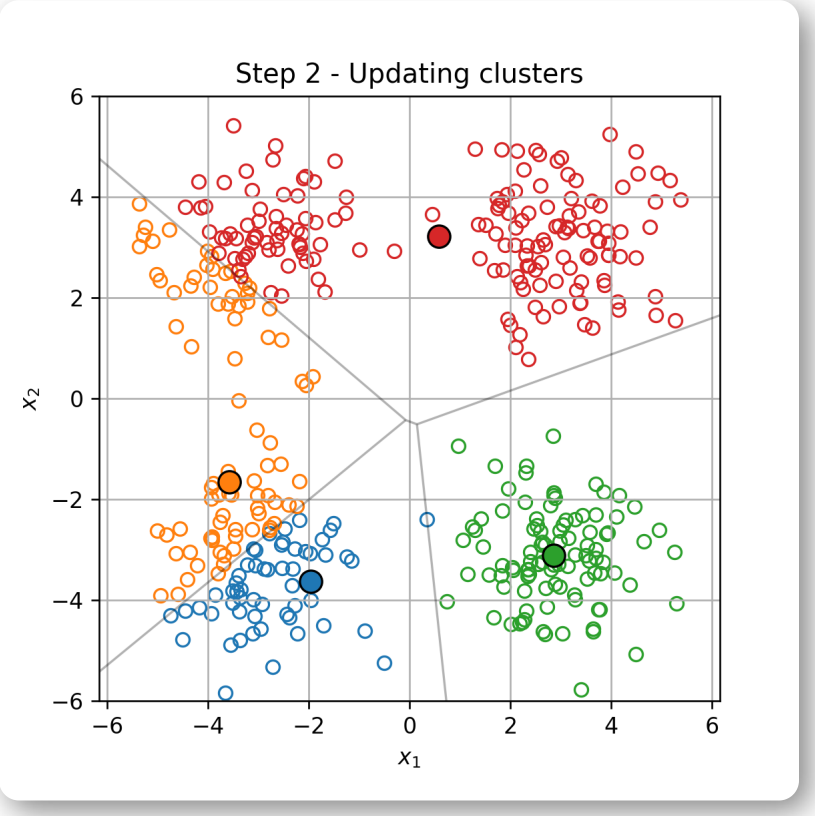


דוגמא

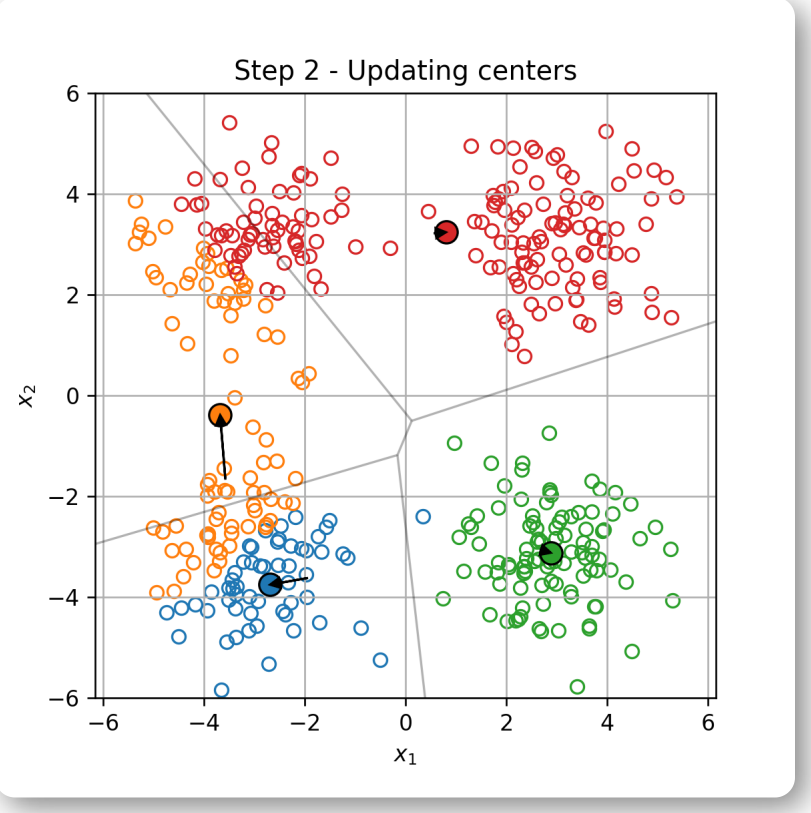
עדכון המרכזים:



עדכון האשכולות:

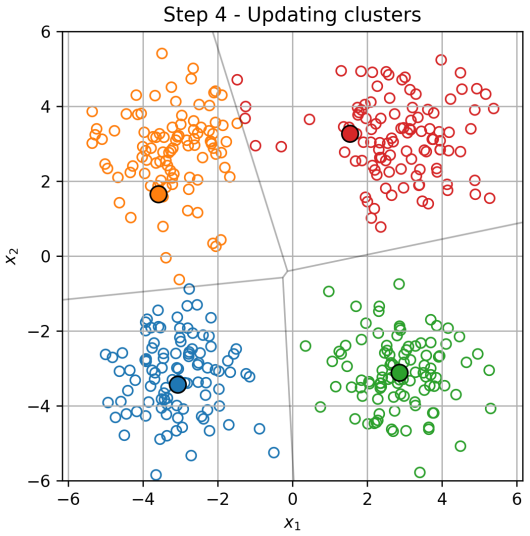


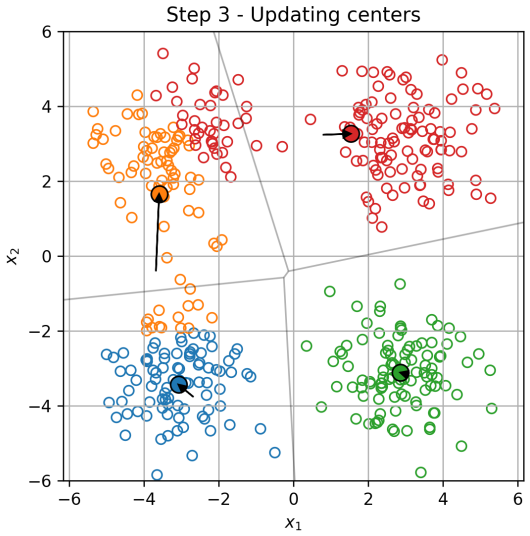
עדכון המרכזים:

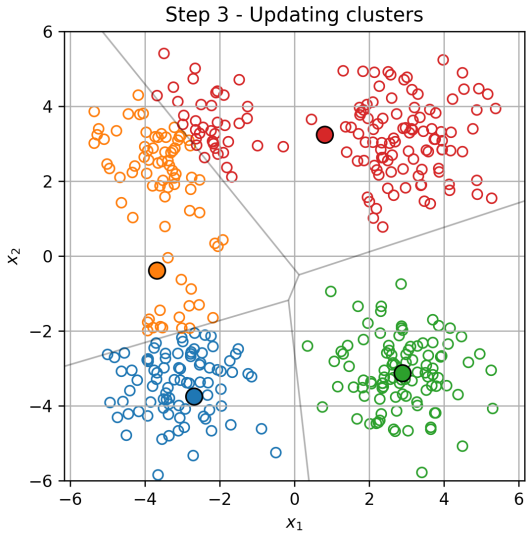


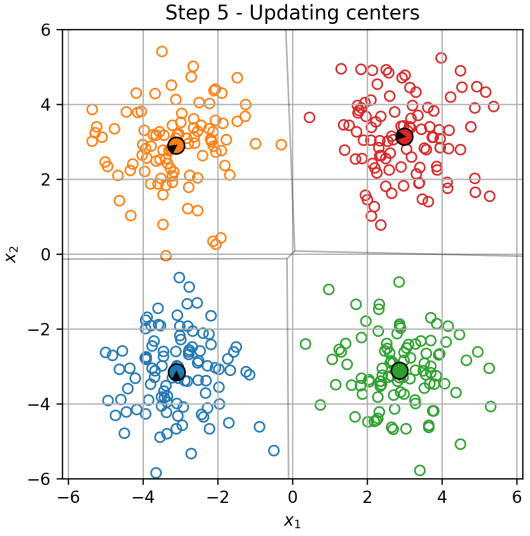
דוגמא

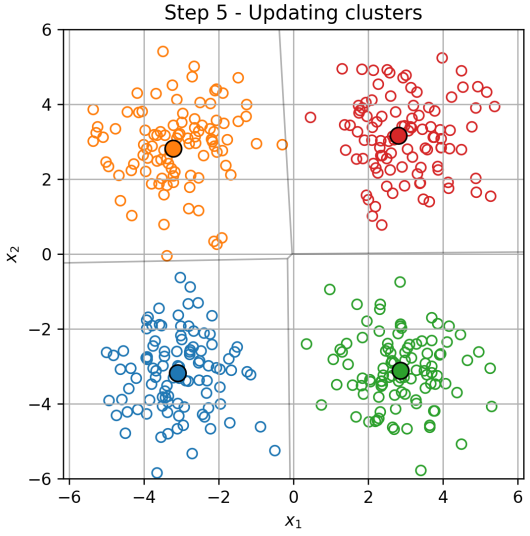
וחוזר חלילה (הסדר הוא מימין לשמאל):

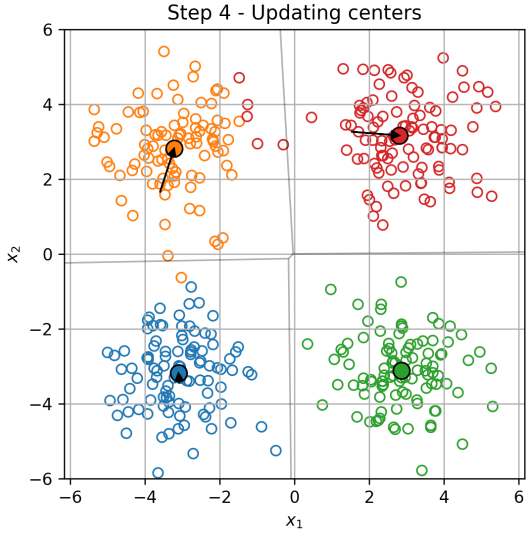




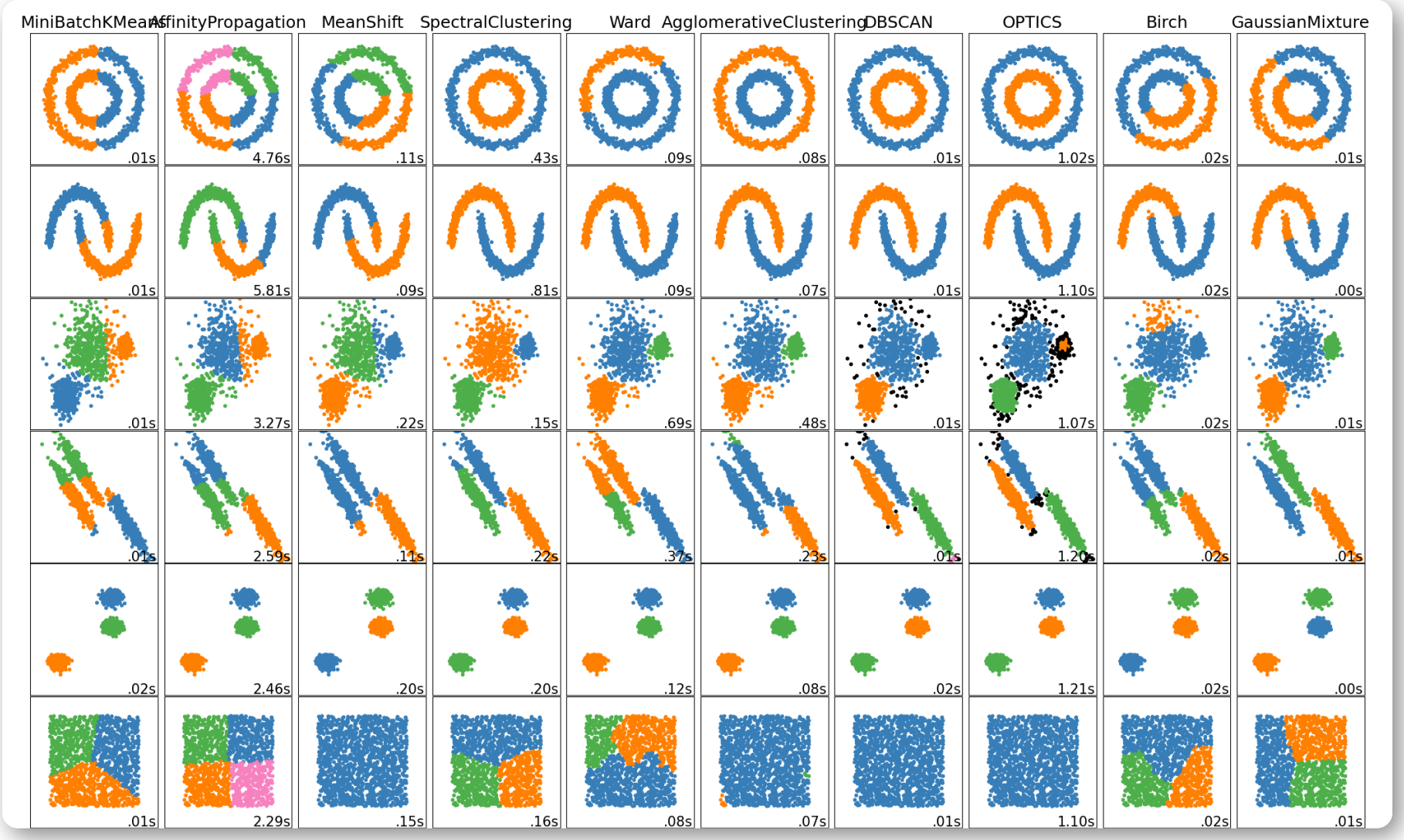








אלגוריתמי אשכול שונים



לרוב לא נוכל לצייר את האשכולות בשני ממדים.