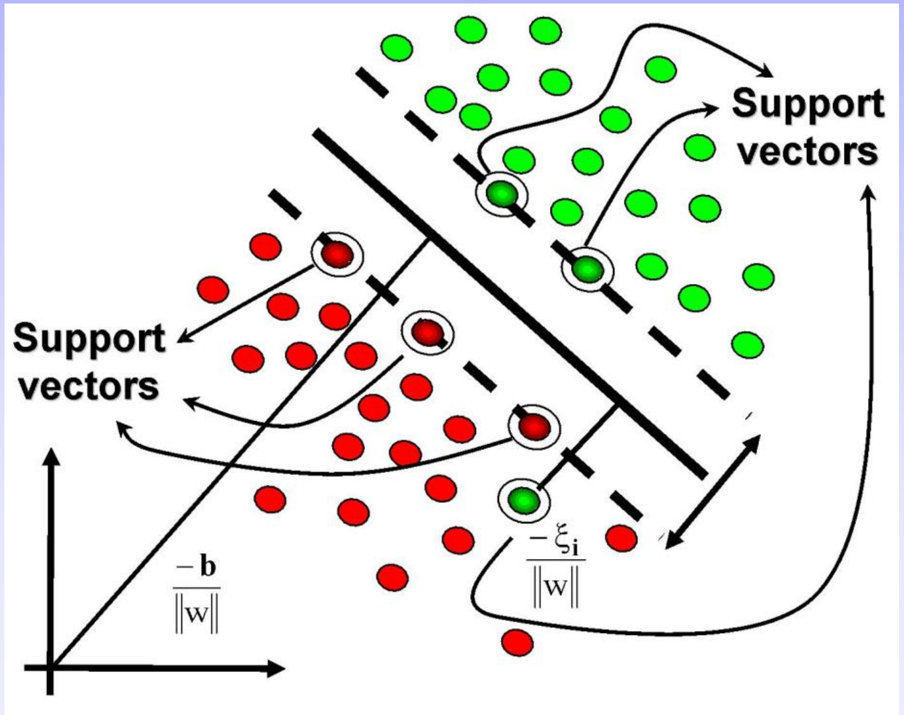
The Support Vector Machine

- עד עתה – הנחת הייחוס - קיים מסווג ליניארי בעל שוליים גאומטריים רחבים, כלומר שגבול ההחלטה שלו (על-מישור, hyperplane) מופרד היטב מכל דוגמאות האימון.
- נרצה להשתמש במסווג כזה.
- האם אפשר למצוא אותו באופן ישיר?

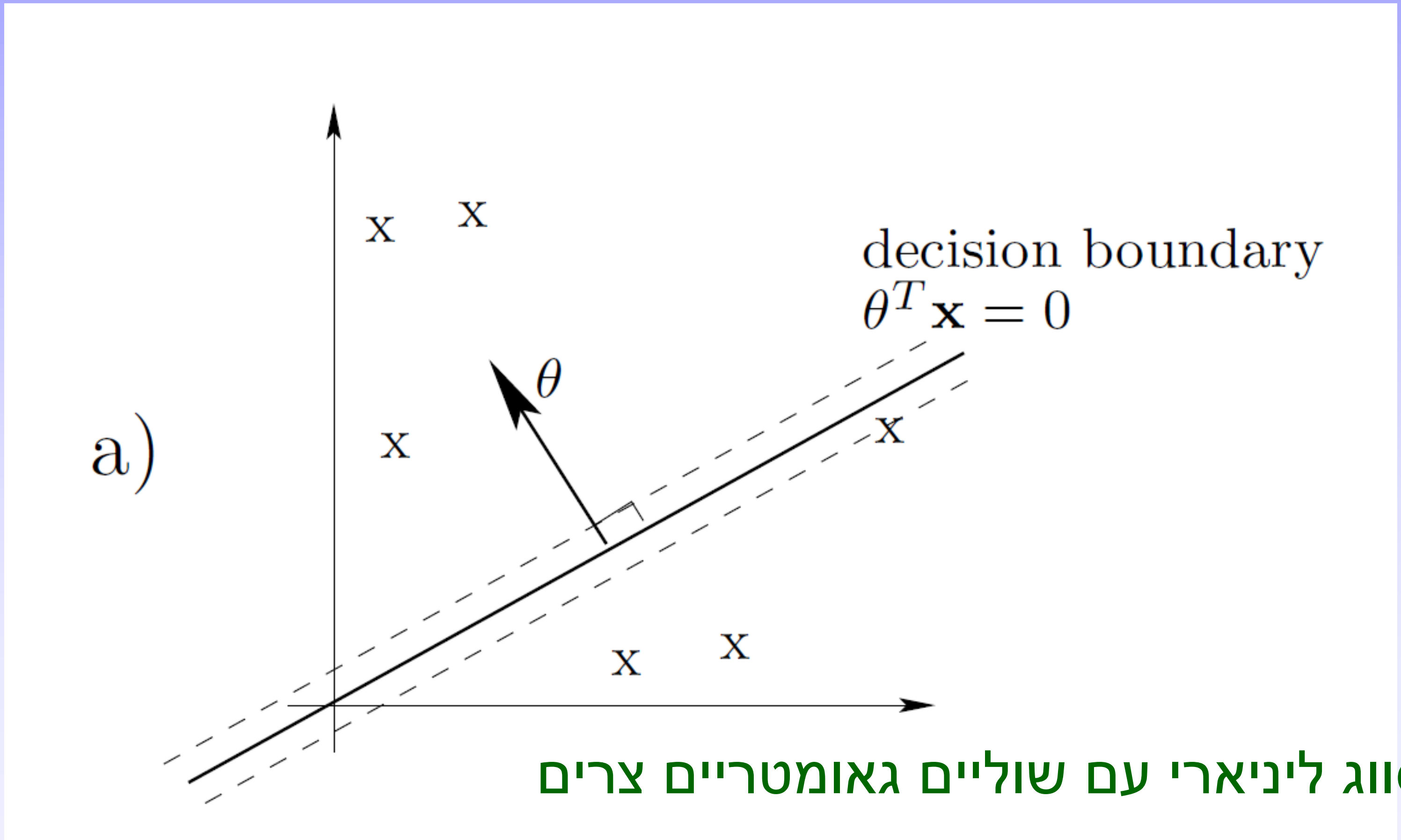
The Support Vector Machine

- כן אפשר!
- מסווג כזה נקרא Support Vector Machine



The Support Vector Machine

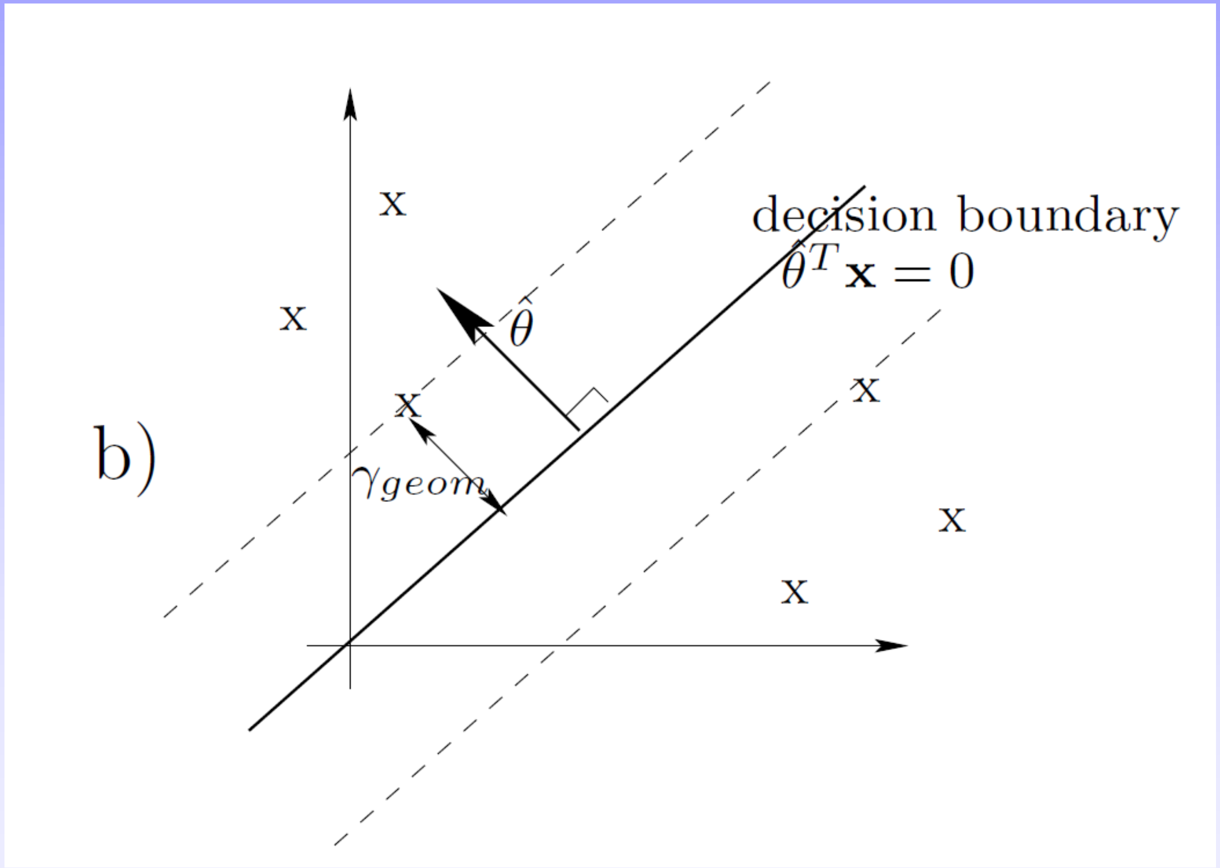
- נדמיין את מציאת המסווג הליניארי עם השוליים המקסימליים ראשית על-ידי זיהוי מסווג כלשהו המסווג נכון את כל הדוגמאות:



- מסווג ליניארי עם שוליים גאומטריים צרים

The Support Vector Machine

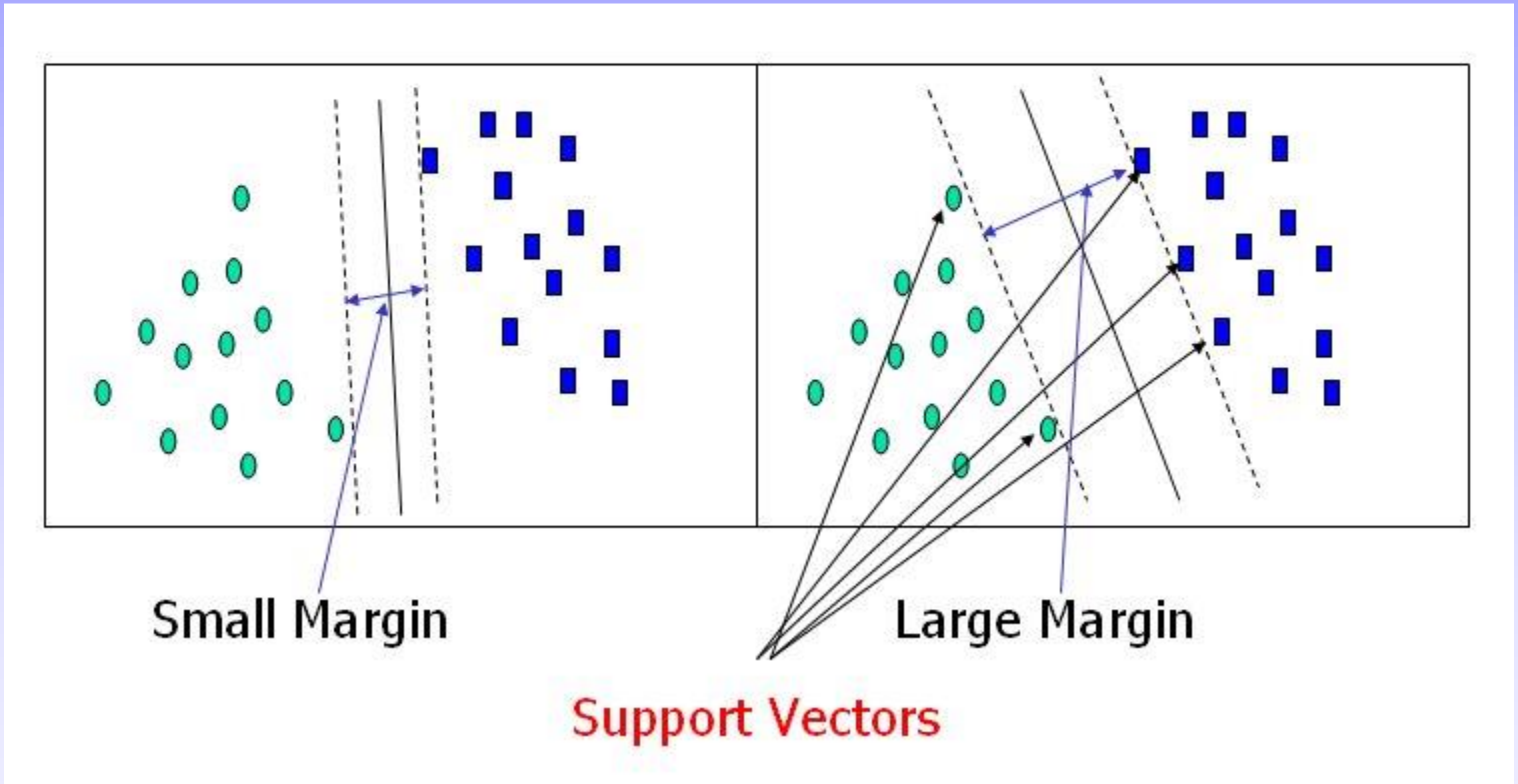
- ואז נגדיל את השוליים הגאומטריים (geometric margin): עד שהמסווג "יינעל" בנקודה בה אי-אפשר להגדיל את השוליים יותר מכך.



מסווג ליניארי עם שוליים גאומטריים רחבים (maximum margin linear classifier)

The Support Vector Machine

מסווג ליניארי עם שוליים גאומטריים רחבים (maximum margin linear classifier)



The Support Vector Machine

- באופן יותר פורמלי, אפשר להעמיד בעיית אופטימיזציה כדי למקסם את השוליים הגאומטריים.
- נדרוש שהמסווג יסווג נכונה את כל דוגמאות האימון:

$$y_t \cdot \theta^T \cdot \underline{x}_t \geq \gamma, \qquad \gamma > 0$$

$$t = 1, 2, \ldots, n \qquad \text{לכל}$$

תחת אילוצים אלה נרצה למקסם את

כלומר את השוליים הגאומטריים

$$\frac{\gamma}{\| \theta \|}$$

The Support Vector Machine

- באופן אלטרנטיבי אפשר למזער את ההפכי: $\dfrac{\|\theta\|}{\gamma}$
- או באופן שווה-ערך את ריבוע ההפכי:

$$\frac{1}{2}\cdot\left(\frac{\|\theta\|}{\gamma}\right)^2$$

תחת אותם אילוצים (ה- ½ כלול רק מטעמי נוחות)

The Support Vector Machine

- אז מתקבלת בעיית האופטימיזציה הבאה:

$$\text{minimize } \frac{1}{2}\cdot\left(\frac{\|\theta\|}{\gamma}\right)^2 \quad \text{subject to } y_t \cdot \theta^T \cdot \underline{x}_t \geq \gamma, \ \text{ for all } t = 1,2,\ldots,n$$

The Support Vector Machine

- מתקבלת בעיית האופטימיזציה הבאה:

$$\text{minimize } \frac{1}{2}\cdot\left(\frac{\|\theta\|}{\gamma}\right)^2 \quad \text{subject to } y_t\cdot\theta^T\cdot\underline{x}_t \geq \gamma, \text{ for all } t=1,2,\dots,n$$

אפשר לפשט את הבעייה על-ידי סילוק γ

נרשום מחדש את הבעייה באופן שידגיש את התלות ב- γ

$$\text{minimize } \frac{1}{2}\cdot\left(\left\|\frac{\theta}{\gamma}\right\|\right)^2 \quad \text{subject to } y_t\cdot\frac{\theta^T}{\gamma}\cdot\underline{x}_t \geq 1,$$

$$\text{for all } t=1,2,\dots,n$$

חילקנו את שני האגפים ב- γ

The Support Vector Machine

$$\text{minimize } \frac{1}{2}\cdot\left(\left\|\frac{\theta}{\gamma}\right\|\right)^2 \quad \text{subject to } y_t \cdot \frac{\theta}{\gamma}^T \cdot \underline{x}_t \geq 1,$$

$$\text{for all } t = 1, 2, \ldots, n$$

במלים אחרות, בעיית הסווג שלנו היא על היחס $\frac{\theta}{\gamma}$ ולא על כל אחד מהם לחוד.

לדוגמא השוליים הגאומטריים מוגדרים רק על בסיס היחס הזה. ביצוע scaling של θ על-ידי קבוע לא משנה את משטח ההחלטה.

The Support Vector Machine

- לפיכך אפשר לקבוע כי $\gamma = 1$ ולפתור עבור θ באופן הבא:

$$\text{minimize } \frac{1}{2} \cdot \left(\|\theta\| \right)^2 \quad \text{subject to } y_t \cdot \theta^T \cdot \underline{x}_t \geq 1,$$

$$\text{for all } t = 1, 2, \ldots, n$$

זוהי בעיית האופטימיזציה הסטנדרטית של SVM והיא בעיית תכנות קוודרטית (המטרה היא ריבועית בפרמטרים עם אילוצים ליניאריים).

השוליים הגאומטריים הנוצרים הם $\frac{1}{\|\theta\|}$ כאשר θ הוא הפתרון היחיד לבעייה.

The Support Vector Machine

$$\text{minimize } \frac{1}{2}\cdot\left(\|\theta\|\right)^2 \quad \text{subject to } y_t\cdot\theta^T\cdot\underline{x}_t \geq 1,$$

$$\text{for all } t = 1, 2, \ldots, n$$

- העל-מישור המפריד (decision boundary) וכן השוליים הגאומטריים לא מושפעים מבחירת $\gamma = 1$

ניסוח כללי, פרמטרי היסט (offset)

- נשנה מעט את המסווג הליניארי – נוסיף **ביטוי היסט (offset).**
במלים אחרות, המסווג בו נתחשב הוא מהצורה:

$$f(x:\theta,\theta_0) = sign(\theta^T \cdot \underline{x} + \theta_0)$$

- עם פרמטרים θ (הנורמליים לעל-מישור המפריד) ועם פרמטר היסט θ_0 שהוא מספר ממשי.

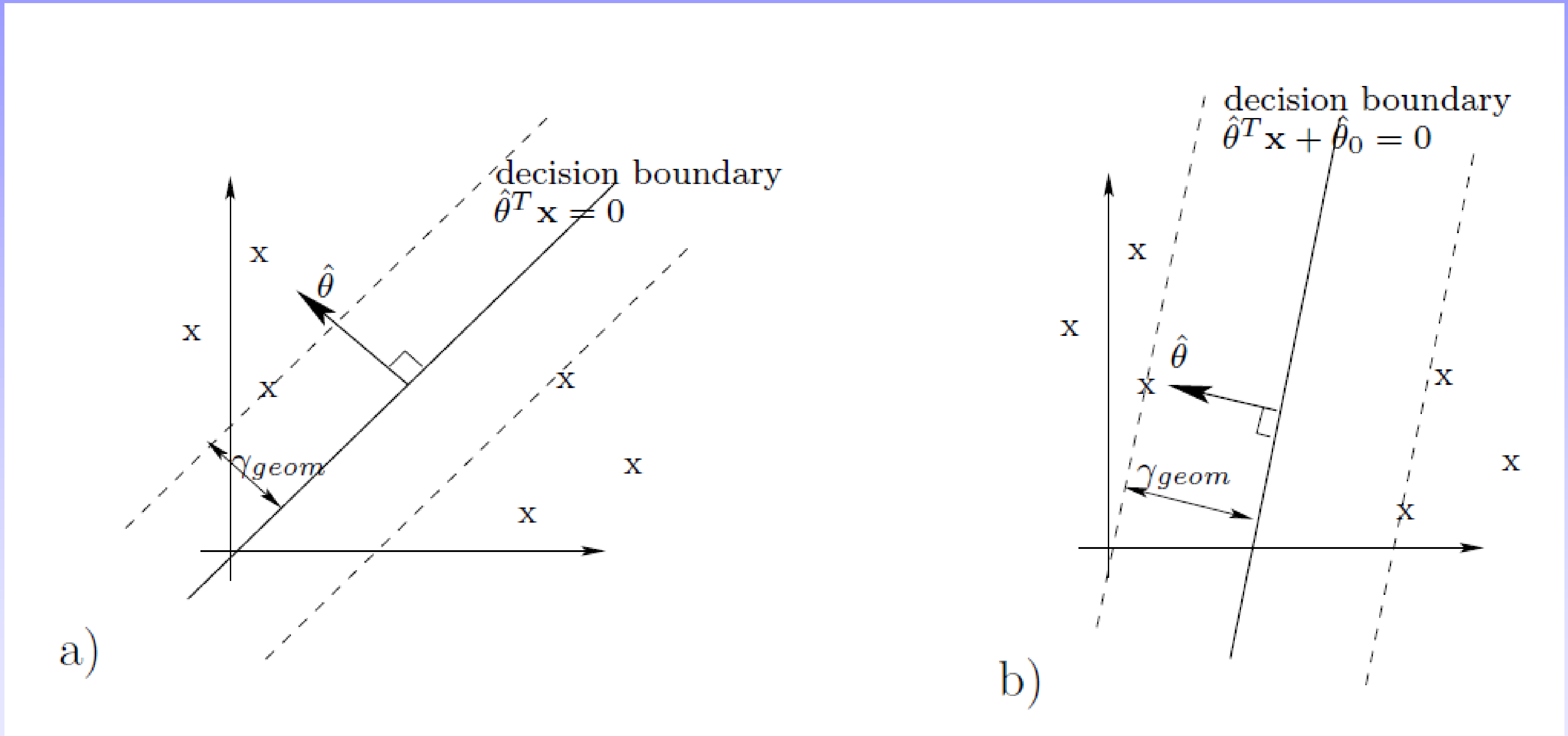
ניסוח כללי, פרמטרי היסט (offset)

- כמו קודם, המשוואה עבור העל-מישור המפריד היא על-ידי קביעת הארגומנט של פונקציית הסימן ל- 0, או:

$$\theta^T \cdot \underline{x} + \theta_0 = 0$$

- זוהי משוואה כללית לעל-מישור (hyper-plane)
- פרמטר ההיסט שהוספנו מוביל למסווג עם שוליים גאומטריים רחבים יותר:

ניסוח כללי, פרמטרי היסט (offset)



a) Maximum margin linear classifier through origin;
b) Maximum margin linear classifier with an offset parameter

ניסוח כללי, פרמטרי היסט (offset)

הערה: הפרמטר θ המתאים לפיתרון השוליים המקסימליים שונה בשני הציורים.

פרמטר ההיסט משנה מעט את בעיית האופטימיזציה:

$$\text{minimize } \frac{1}{2}\cdot\left(\|\theta\|\right)^2 \quad \text{subject to } y_t\cdot(\theta^T\cdot \underline{x}_t+\theta_0)\geq 1,$$

$$\text{for all } t=1,2,\ldots,n$$

(הערה: פרמטר ההיסט מופיע רק באילוצים.

זה שונה משינוי באופן פשוט של המסווג הליניארי דרך הראשית, כך שנזין אותו בדוגמאות שיש להן בנוסף רכיב קבוע, לדוגמא

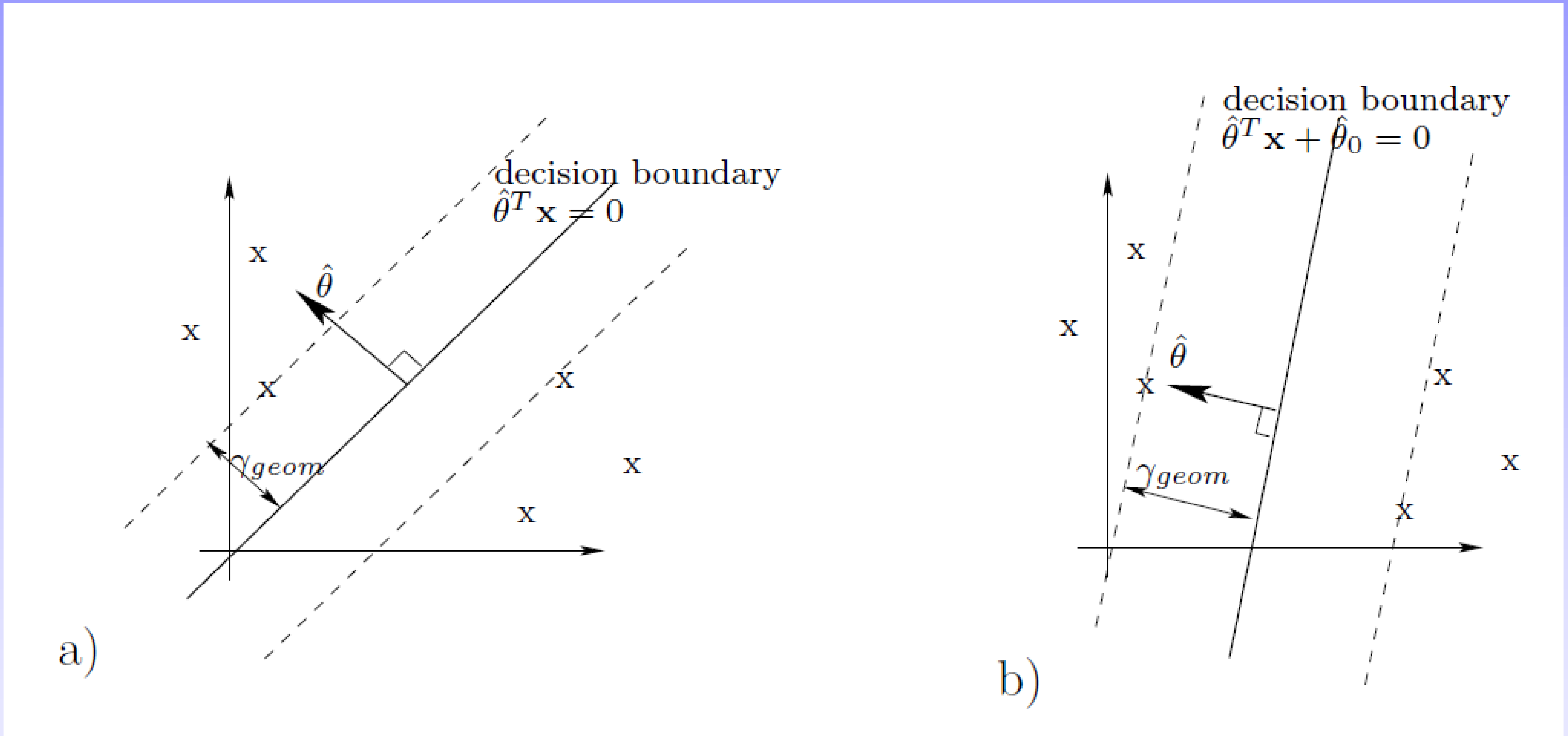
$$x' = [x; \quad 1]$$

תכונות המסווג הליניארי עם שוליים מקסימליים

יתרונות:

- פתרון ייחודי לכל קבוצת אימון פרידה ליניארית
- הצבת המשטח המפריד רחוק ככל האפשר מכל דוגמאות האימון – רובסטיות לדוגמאות רועשות (אך לא לתגיות רועשות).
- הפתרון תלוי רק בתת-קבוצה של דוגמאות האימון, אלה הנמצאות בדיוק על השוליים
- (ראו בציורים הבאים את הקוים המקווקוים, המקבילים לעל-מישור המפריד).

ניסוח כללי, פרמטרי היסט (offset)



a) Maximum margin linear classifier through origin;
b) Maximum margin linear classifier with an offset parameter

תכונות המסווג הליניארי עם שוליים מקסימליים

- דוגמאות אלה (התבניות, הוקטורים), הנמצאות בדיוק על השוליים נקראות Support Vectors.
- שאר הדוגמאות יכולות להיות בכל מקום בלי להשפיע על התוצאה (המסווג, העל-מישור המפריד).
- לפיכך אותה תוצאה מתקבלת גם אם רק ה- Support Vectors נמצאים בקבוצת האימון.

- האם זו תוצאה טובה? כדי לענות על שאלה זו צריך דרך יותר פורמלית למדידת כמה טוב כל מסווג.

Cross-Validation

- מדד אפשרי "הוגן" אחד המשוערך רק על בסיס דוגמאות האימון הוא אימות מצולב (cross-validation)
- זוהי פשוט שיטה של אימון מחדש של המסווג על תת-קבוצות של דוגמאות האימון, ועריכת מבחן עבור המסווג על הדוגמאות הנותרות בהן לא משתמשים באימון, ולכן זהו מדד "הוגן".

- גירסה מיוחדת של פרוצדורה זו: leave-one-out cross-validation

Leave-one-out cross-validation

הפרוצדורה מוגדרת באופן הבא:

1. בחרו אחת מדוגמאות האימון, והוציאו אותה מקבוצת האימון.
2. אמנו את המסווג על בסיס שאר הדוגמאות (דוגמאות האימון)
3. בחנו את המסווג שנוצר על הדוגמא שיצאה.
4. מנו את מספר השגיאות
5. חזרו על צעדים 1-4 מספר פעמים כמספר דוגמאות האימון.

Leave-one-out cross-validation

באופן מדוייק יותר:

- נסמן ב- i את הפרמטרים (θ) המושגים על-ידי מציאת ה- maximum margin linear separator ללא הדוגמא ה- i –ית.
- אזי:

$$Leave-one-out\ CV\ error = \frac{1}{m}\sum_{i=1}^{m} L(y^{(i)}, f(x^{(i)};\theta^i,\theta_0^i))$$

- כאשר פונקציית ההפסד: $L(y, y') \in \{0,1\}$

Leave-one-out cross-validation

- אנו מנסים למדוד באופן יעיל עד כמה יוכל המסווג להכליל עבור כל אחת מדוגמאות האימון, אם היא לא חלק מקבוצת דוגמאות האימון.

$$Leave-one-out\ CV\ error = \frac{1}{m}\sum_{i=1}^{m} L(y^{(i)}, f(x^{(i)}; \theta^{i}, \theta_0^{i}))$$

$$L(y, y') \in \{0,1\}$$

- מסווג בעל שגיאת leave-one-out CV error נמוכה סביר שיוכל לבצע הכללה היטב, (למרות שזה לא מובטח)

Leave-one-out cross-validation error

- עתה ננתח מהי שגיאת ה-‏‏Leave-one-out cross-validation עבור המסווג הליניארי

Maximal margin linear classifier

- דוגמאות הנמצאות מחוץ לשוליים יסווגו נכון ללא תלות אם הם חלק מקבוצת האימון.
- לא-כך עבור וקטורי התמך – support vectors

(אלה מהווים מפתח להגדרת המפריד הליניארי, ולכן אם נוציא אותם מקבוצת האימון, הם עלולים כפי הנראה להיות מסווגים לא נכון).

Leave-one-out cross-validation error

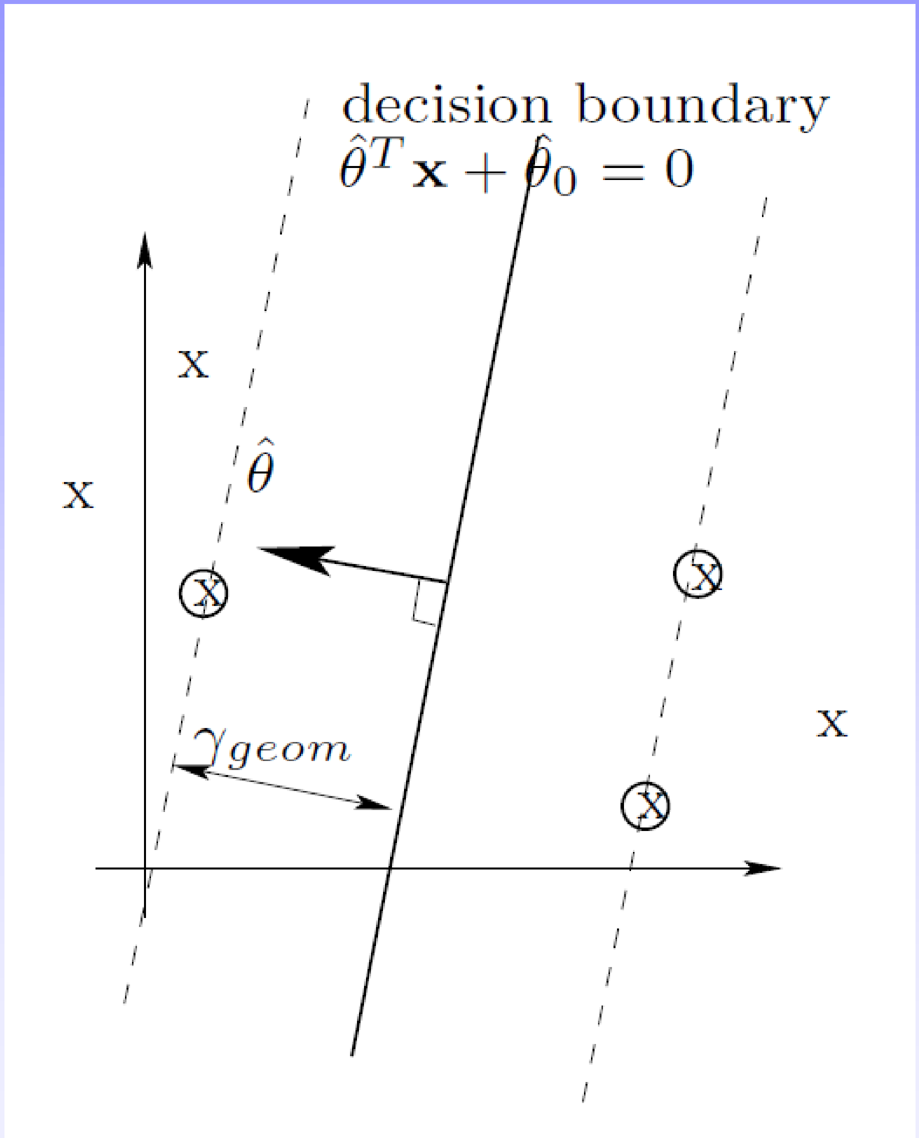
- אפשר לגזור חסם עליון פשוט עבור ה Leave-one-out cross-validation error:
-

$$Leave-one-out\ CV\ error \leq \frac{\#\ \text{number of support vector}}{n}$$

מספר קטן של וקטורי תמך – פתרון דליל – יש לו יתרון.

זהו טיעון נוסף לטובת ה- Maximal margin linear classifier

Problems



- בעיות:
- מספיק דוגמת תיוג אחת המתוייגת באופן שגוי, כדי לגרום לשינוי רדיקלי במסווג הליניארי בעל השוליים המקסימליים
- לדוגמא, מה יקרה אם נחליף את התיוג עבור וקטור התמך הימני העליון בציור?
- לפיכך נאפשר דוגמאות המסווגות לא נכון, relaxation

Relaxation

- שגיאות תיוג הן שכיחות מאוד בבעיות מעשיות רבות, ונרצה לנסות להפחית את ההשפעה שלהן.
- אי-אפשר לדעת מראש אם דוגמאות האימון הן קשות לסווג בגלל שגיאות תיוג או בגלל שהן בלתי פרידות ליניארית.
- בכל מקרה צריך לנסח את הפשרה בין סווג לא נכון של דוגמת אימון לבין פוטנציאל התועלת שיש לדוגמאות אחרות.

Slack Variables

- הדרך הפשוטה ביותר לאפשר שגיאות במסווג הליניארי בעל השוליים המקסימליים היא להציג משתני מרווח (slack variables) עבור אילוצי הסווג/שוליים (classification/margin) בבעיית האופטימיזציה
- במלים אחרות, מודדים את הדרגה בה מופר אילוץ השוליים ומקשרים עלות לכל הפרה כזאת

Slack Variables

- ממזערים את העלויות (costs) של אילוצי ההפרה יחד עם הנורמה של וקטור הפרמטרים:

$$\text{minimize } \frac{1}{2}\cdot\left(\|\theta\|\right)^2 + C\sum_{t=1}^{n}\varepsilon_t$$

$$\text{subject to } y_t\cdot(\theta^T\cdot \underline{x}_t+\theta_0)\geq 1-\varepsilon_t \quad and\ \varepsilon_t\geq 0,$$

$$\text{for all } t=1,2,\ldots,n$$

- כאשר ε_t הם משתני המרווח.

Slack Variables

- אילוץ השוליים מופר כאשר יש צורך לקבוע $\varepsilon_t > 0$ עבור דוגמא כלשהי.
- "העונש " עבור חריגה זו הוא $C \cdot \varepsilon_t$ והוא מהווה פשרה עם הרווח האפשרי במיזעור הנורמה הריבועית של וקטור הפרמטרים $\| \theta \|^2$
- אם מגדילים את C עבור חריגות או הפרות שוליים, אזי בנקודה מסויימת כל ה- $\varepsilon_t = 0$ ואז חוזרים ל- maximum margin linear separator (אם אפשרי).

Slack Variables

- אם מגדילים את C עבור חריגות או הפרות שוליים, אזי בנקודה מסויימת כל ה- $\varepsilon_t = 0$ ואז חוזרים ל- maximum margin linear separator (אם אפשרי).
- מצד שני עבור C קטן מדי, אילוצי שוליים רבים מדי עשויים להיות מופרים.
- הבעייה עתה נקראת בעיית האופטימיזציה המשוככת relaxed optimization problem והיא מציינת פשרה כמותית בין הנורמה הריבועית של וקטור הפרמטרים לבין הפרות או חריגות השוליים.

$$\text{minimize } \frac{1}{2}\cdot\left(\|\theta\|\right)^2 + C\sum_{t=1}^{n}\varepsilon_t$$

$$\text{subject to } y_t\cdot(\theta^T\cdot\underline{x}_t+\theta_0)\geq 1-\varepsilon_t \quad and\ \varepsilon_t\geq 0,$$

$$\text{for all } t=1,2,\ldots,n$$

Relaxation

- ננסה להבין את העמדת הבעייה:
- לדוגמא, מהם השוליים הנוצרים כאשר חלק מהאילוצים מופרים?
- עדיין אפשר להשתמש ב- $\frac{1}{\|\theta\|}$ בתור השוליים הגאומטריים.
- אלה אכן השוליים הגאומטריים המבוססים על הדוגמאות עבורן $\varepsilon_t^* = 0$ כאשר "*" מציין את הערך האופטימלי.

האם זהו המקרה שמקבלים את מסווג השוליים המקסימליים הליניארי עבור תת-הקבוצה של הדוגמאות עבורן $\varepsilon_t^* = 0$?

התשובה היא לא, הדוגמאות המפרות את אילוצי השוליים, כולל אלה שמסווגים באופן לא נכון (הפרה גדולה יותר), משפיעות על הפתרון.

Relaxation

- התשובה היא לא, הדוגמאות המפרות את אילוצי השוליים, כולל אלה המסווגות באופן לא נכון (הפרה גדולה יותר), משפיעות על הפתרון.
- במלים אחרות, וקטור הפרמטרים האופטימלי מוגדר על בסיס דוגמאות הנמצאות על השוליים, כאלה החורגות מהשוליים אך לא מספיק כדי להיות מסווגות באופן שגוי, וכאלה המסווגות באופן שגוי. כל הדוגמאות האלה הן במובן זה supprot vector

$$\text{minimize} \quad \frac{1}{2}\cdot\left(\|\theta\|\right)^2 + C\sum_{t=1}^{n}\varepsilon_t$$

$$\text{subject to } y_t\cdot(\theta^T\cdot \underline{x}_t+\theta_0)\geq 1-\varepsilon_t \quad \mathit{and}\ \varepsilon_t\geq 0,$$

$$\text{for all } t=1,2,\ldots,n$$

המודל הליניארי לעתים מוגבל

- במה המודל הוא ליניארי?

$$sign\left(\sum_{i=1}^{n}\theta_i x_i\right)$$

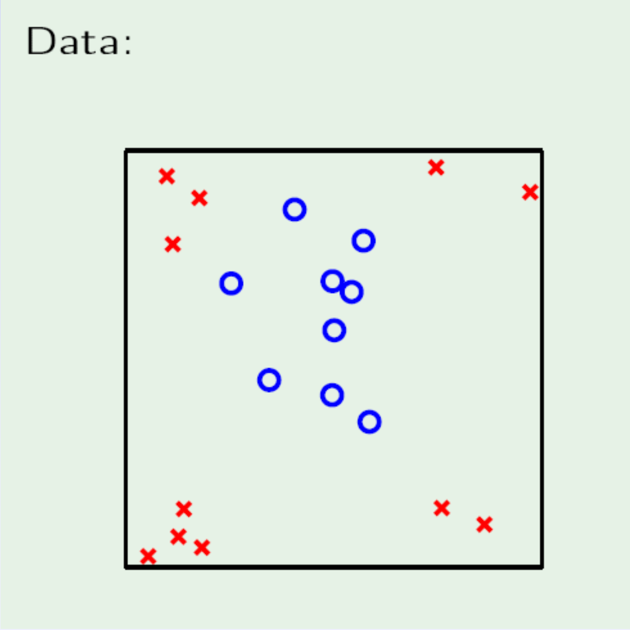
- ברור שהמסווג הוא ליניארי – ליניארי ב- xi, אבל יותר חשוב מכך, המסווג הליניארי הוא ליניארי במשקלות.
- כאשר מדובר על לימוד פונקציית הסווג – דוגמאות האימון הם הקלטים הידועים, כלומר אפשר להתייחס אליהם כאל פרמטרים, ולא כאל משתנים, ואילו הנעלמים הם ה- θ אותם צריך לקבוע.

$$sign\left(\sum_{i=1}^{n}\theta_i x_i\right)$$

המודל הליניארי לעתים מוגבל

- לעתים קרובות הנתונים לא פרידים ליניארית.

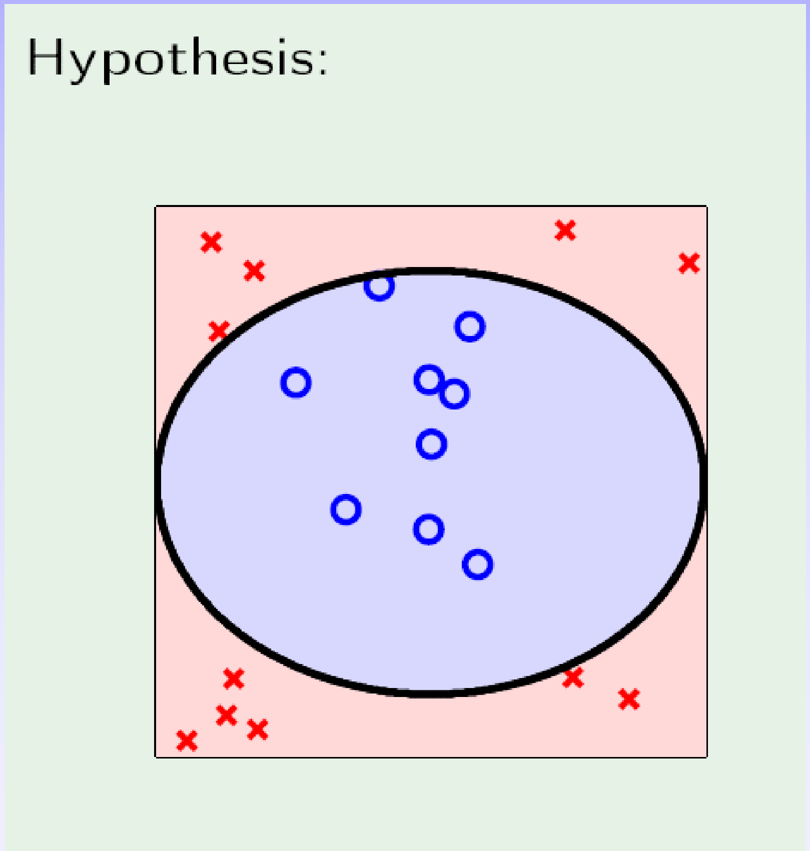
- האם אפשר להפריד את הנתונים באמצעות מסווג ליניארי?
- עדיין, ברור שהכחולים קרובים למרכז, (נניח כי ראשית הצירים נמצאת במרכז) והאדומים נמצאים בפריפריה.



היפותזה אפשרית

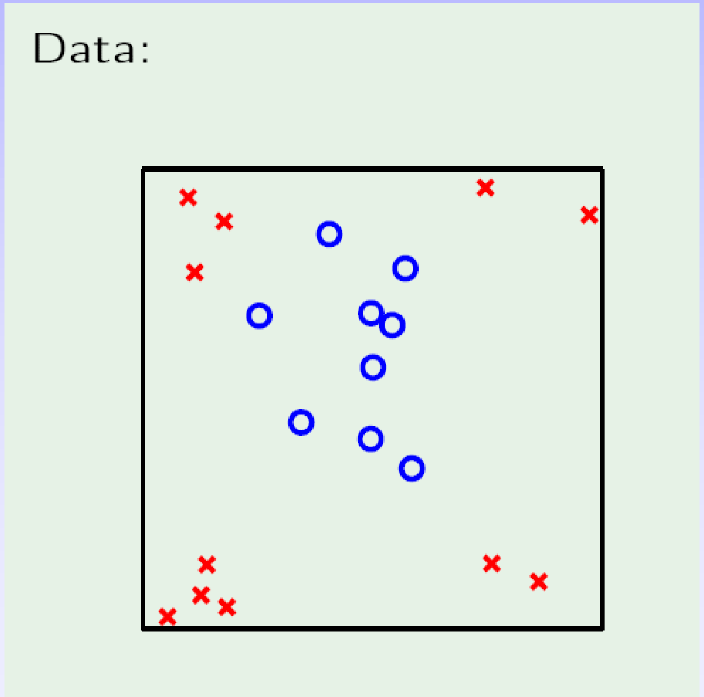
האם נוכל להשתמש בהיפותיזה הנראית כך :

הבעייה : זוהי היפותיזה לא ליניארית



היפותזה אפשרית

- האם אפשר להשתמש במודל ליניארי כדי להפריד בין הדוגמאות הכחולות (+1) והאדומות (-1)?

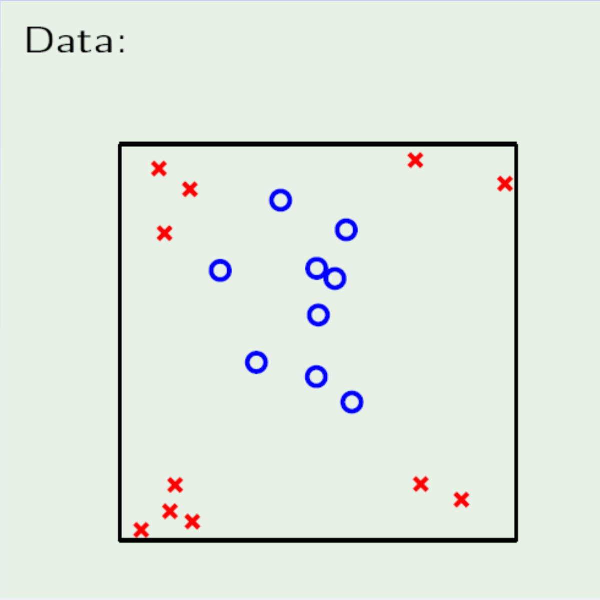


היפותזה אפשרית

- מה המשמעות של "המודל הוא ליניארי"

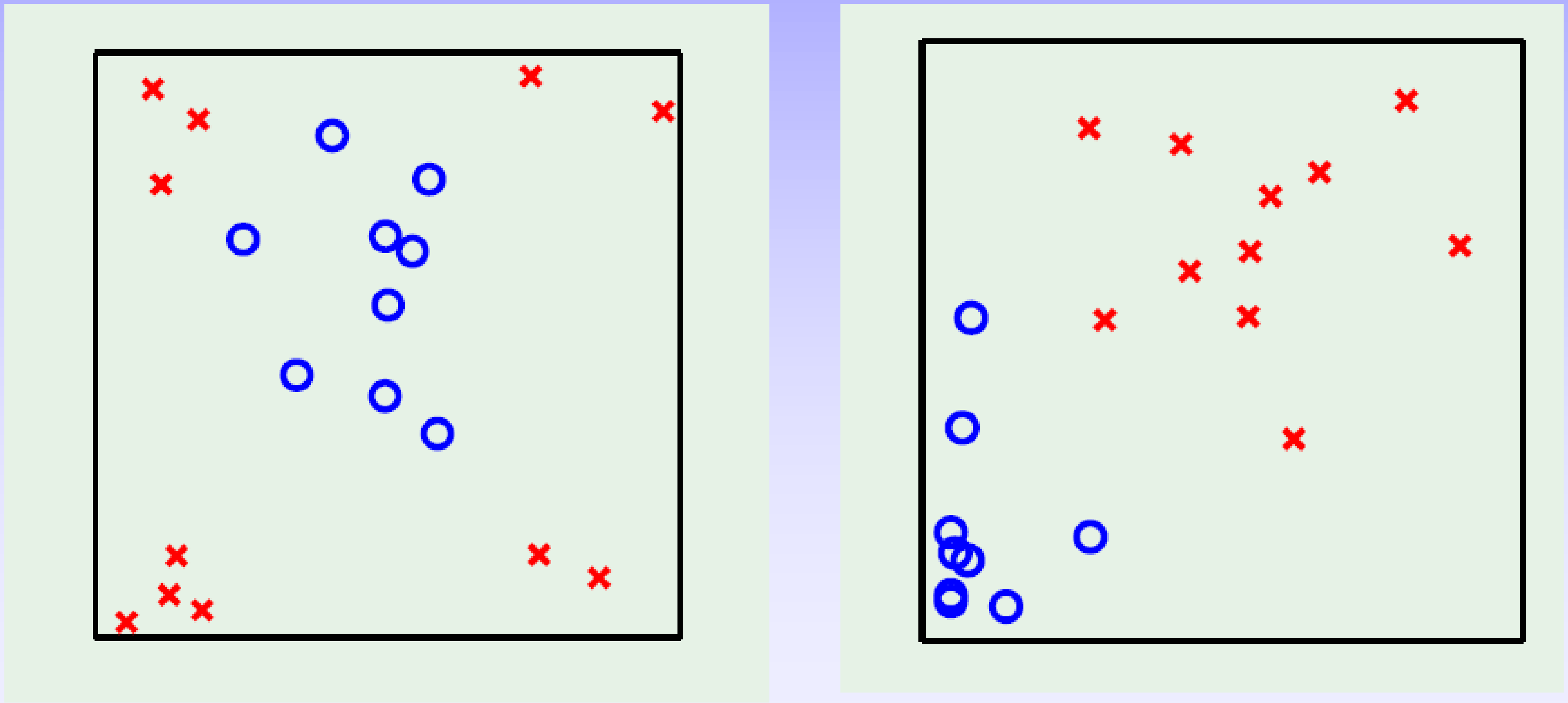
$$sign\left(\sum_{i=1}^{n}\theta_i x_i\right)$$

- הליניאריות היא במשתנה הבלתי תלוי x , אבל האלגוריתמים עובדים בגלל הליניאריות של המקדמים θ.



טרנספורמציה לא-ליניארית

$$(x_1, x_2) \xrightarrow{\Phi} (x_1^2, x_2^2)$$



טרנספורמציה לא-ליניארית

הביטוי $\theta^T x$ ליניארי ב- θ.

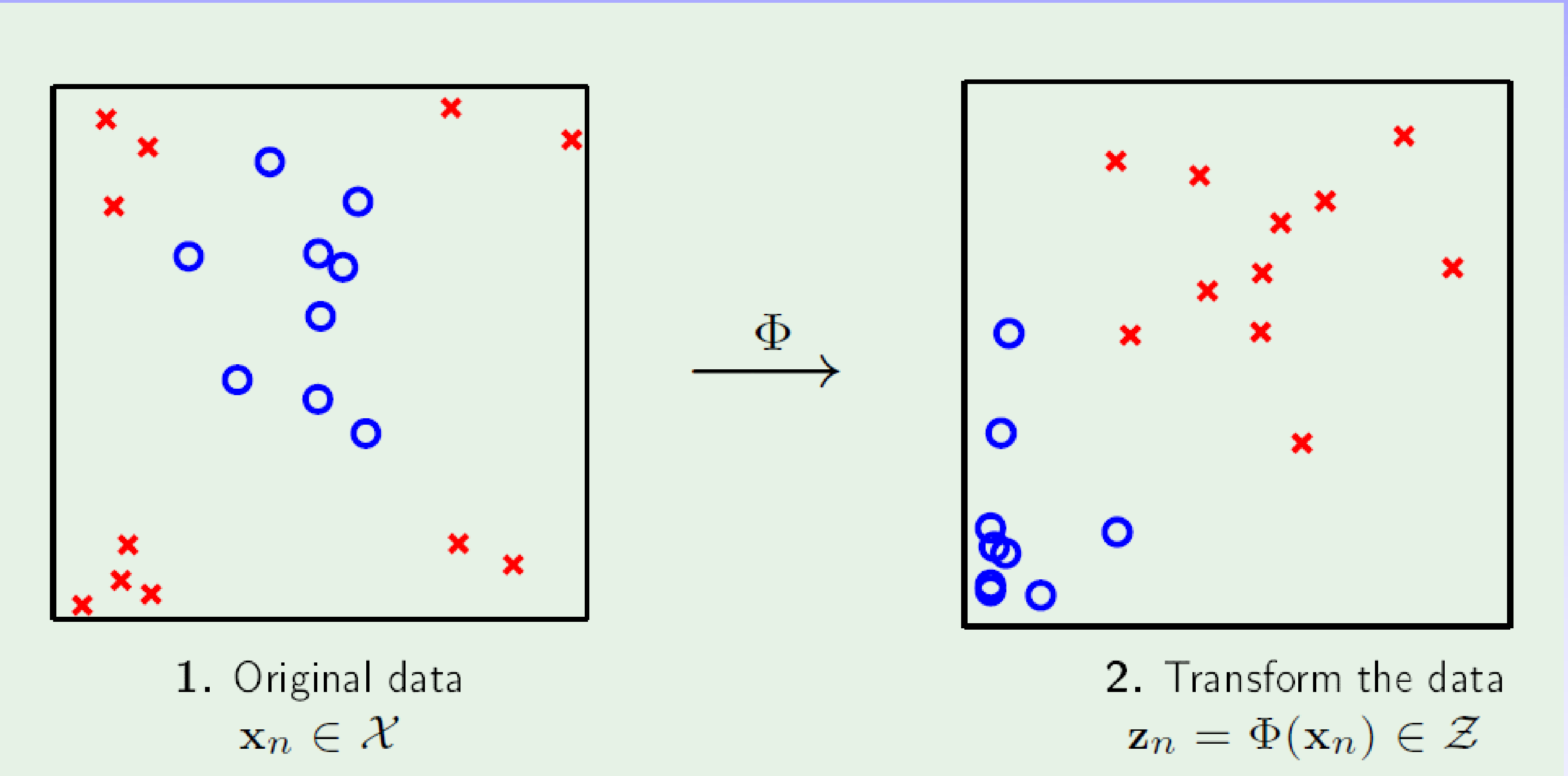
כל טרנספורמציה $X \xrightarrow{\Phi} Z$ משמרת תכונה זו.

דוגמא :

$$(x_1, x_2) \xrightarrow{\Phi} (x_1^{\,2}, x_2^{\,2})$$

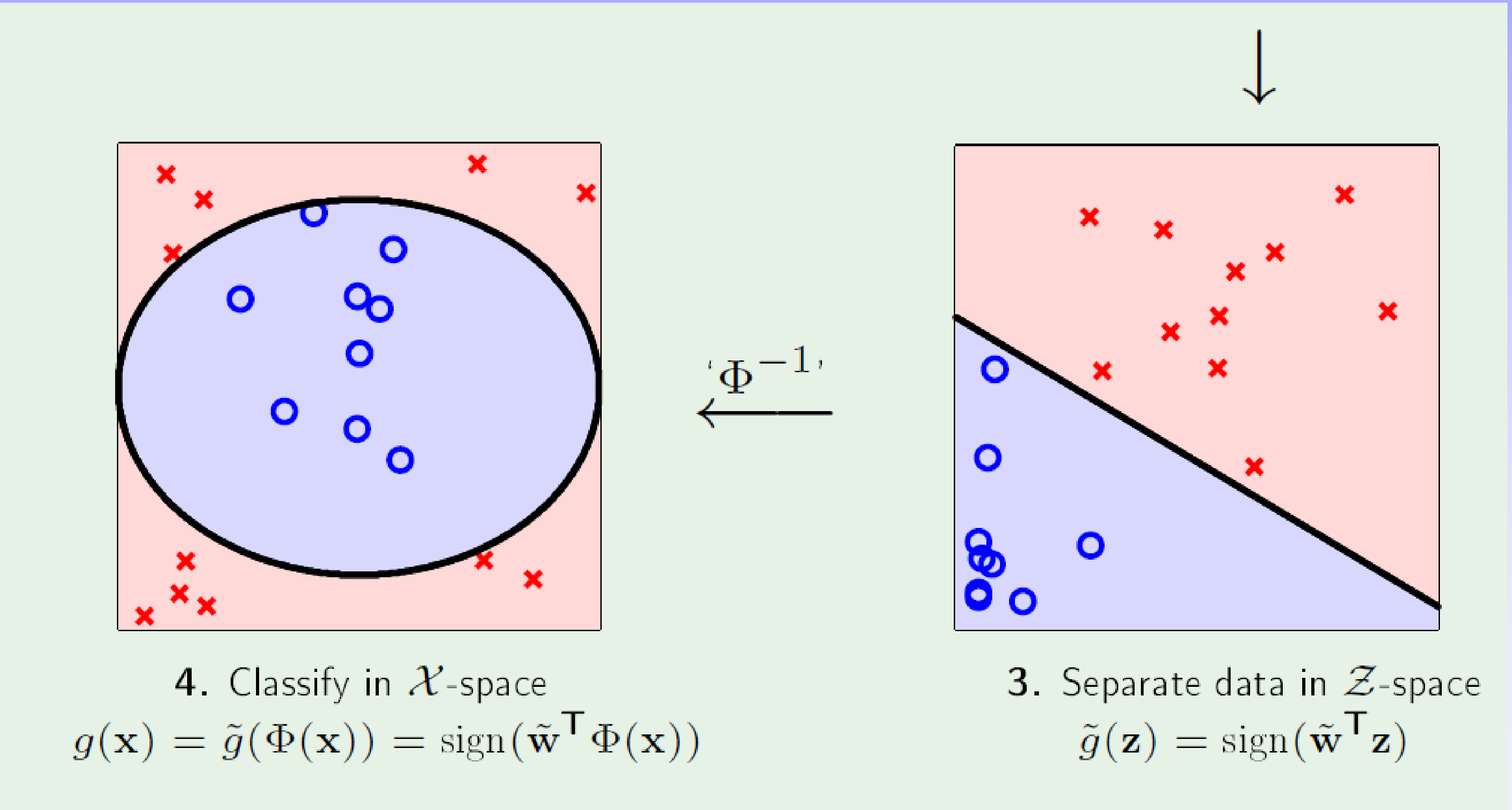
טרנספורמציה לא-ליניארית

המרחב המקורי, מרחב X, ראשית הצירים היא במרכז.
וכפי שראינו אי-אפשר להפריד את הנתונים באמצעות ישר (או מישור)
הרעיון : הפעלת טרנספורמציה לא ליניארית, אותה נסמן ב- ϕ

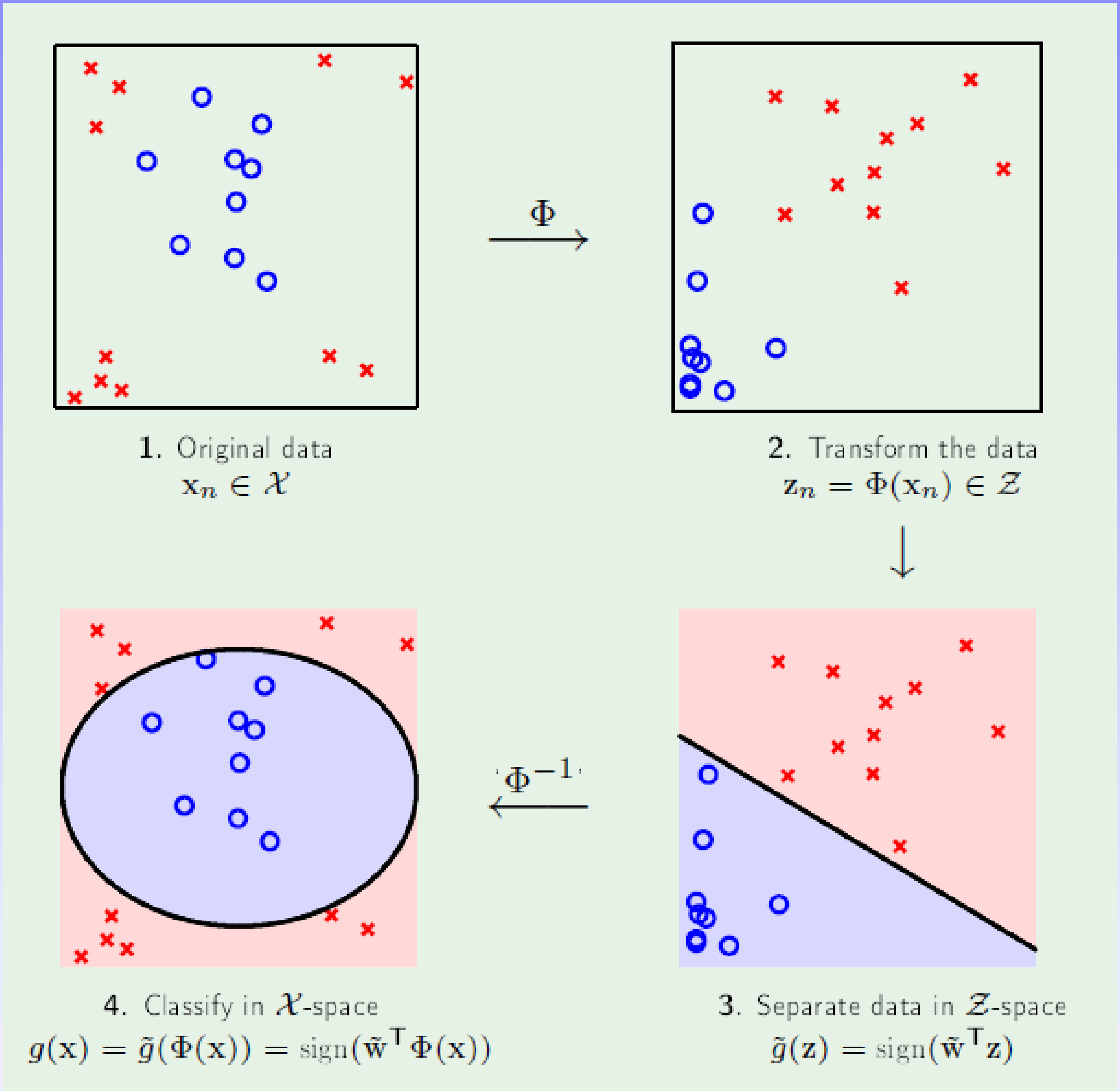


טרנספורמציה לא-ליניארית

במרחב Z הנתונים ניתנים להפרדה ליניארית – מקבלים גבול הפרדה.
אפשר להפריד את הנתונים הנמצאים באמצעות ישר



טרנספורמציה
לא-ליניארית



מה מועתק למה?

$$(x_0, x_1, \ldots, x_d) \xrightarrow{\Phi} (z_0, z_1, \ldots, z_{\tilde{d}})$$

$$X_1, X_2, \ldots, X_N \xrightarrow{\Phi} Z_1, Z_2, \ldots, Z_N$$

$$y_1, y_2, \ldots, y_N \xrightarrow{\Phi} y_1, y_2, \ldots, y_N$$

No weights in X

$$\tilde{\theta} = (\theta_0, \theta_1, \ldots, \theta_{\tilde{d}})$$

$$g(x) = sign(\tilde{\theta}^T \Phi(X))$$