(20942) מבוא ללמידה חישובית | סיכום הרצאה 5

מנחה: שי מימון

סמסטר: 2022א'

נכתב על ידי: מתן כהן

1 רגרסיה לוגיסטית - חזרה וחידודים

בשיעור הקודם למדנו על רגרסיה לוגיסטית ככלי סיווג אשר כל דוגמה מתוייגת כ- $y_i \in \{-1, 1\}$ והחזאי מורכב מפונקצית סיגמויד מורכבת על פונקציה אפינית: $\hat{y}_n = \theta(\hat{\mathbf{w}}^T \mathbf{x}_n) \in \{0, 1\}$

כמו-כן ראינו כי:

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$$

כעת, נבחין כי ניתן לראות את המודל בצורה הבאה:

$$\log\left(\frac{P(y_n = +1 \mid \mathbf{x}_n)}{1 - P(y_n = +1 \mid \mathbf{x}_n)}\right) = \mathbf{w}^T \mathbf{x}_n$$

במילים אחרות - לוגריתם של יחס ההסתברויות לקבלת תיוג "1" ולקבלת תיוג "-1"

הפיתוח:

- נסמן $P(y_n = 1 \mid \mathbf{x}_n) = h(\mathbf{x}_n)$
- נקבל:

$$\begin{aligned}
&h(\mathbf{x}_n) = (1 - h(\mathbf{x}_n)) \cdot e^{\mathbf{w}^T \mathbf{x}_n} \\
\Rightarrow &h(\mathbf{x}_n) \cdot (1 + e^{\mathbf{w}^T \mathbf{x}_n}) = e^{\mathbf{w}^T \mathbf{x}_n} \\
\Rightarrow &h(\mathbf{x}_n) = \frac{e^{\mathbf{w}^T \mathbf{x}_n}}{1 + e^{\mathbf{w}^T \mathbf{x}_n}}
\end{aligned}$$

ככה קיבלנו את $E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \log(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$ משיעור קודם

הערה. גזירה פעמיים של פונקציה E_{in} תניב את העובדה שהיא קונבקסית - דבר נוח מ-2 סיבות עיקריות:

(1) לפונקציה קונבקסית יש אופטימום והוא גלובלי

(2) ישנם אלגוריתמים יעלים למצוא פתרון אופטימלי

מסקנה. נוח לעבוד עם פונקצית הסיגמויד

2 סיכום קצר

עד כה ראינו כמה אלגוריתמים

2.1 פרספטרון

(1) דוגמה מתויגת $y_n \in \{-1, 1\}$ וההיפותזה מקיימת $h(\mathbf{x}) = sign(\hat{\mathbf{w}}^T\mathbf{x})$

(2) אלגוריתמי PLA ו- $Pocket$

(א) ניסינו להביא למינימום את:

$$E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N} I(\hat{y}_n \neq y_n) = \frac{1}{N}\sum_{n=1}^{N} I(sign(\hat{\mathbf{w}}^T\mathbf{x}) \neq y_n)$$

הערה. יכלנו גם להשתמש להגדיר פונקצית מטרה $E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N} max(0, -y_n\mathbf{w}^T\mathbf{x}_n)$ ולהשתמש ב-Gradient Descent על מנת להגיע לפונקציה שכבר ראינו שמתקנת את קו החלוקה של הפרספטרון

2.2 רגרסיה לינארית

(1) כל דוגמה $y_n \in \mathbb{R}$ וההיפותזה מקיימת $h(\mathbf{x}_n) = \mathbf{w}^T\mathbf{x}_n$

(2) דיברנו על כמה פונקציות מחיר שהעיקרית מביניהן הייתה **ממוצע ריבועי השגיאות**:

$$E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{\mathbf{w}}^T\mathbf{x}_n\right)^2 = \frac{1}{N}\left\|\mathbf{y} - X\hat{\mathbf{w}}\right\|^2$$

(א) כאשר יש פתרון יחיד אז X בעלת דרגה מלאה ו $\hat{\mathbf{w}} = (X^TX)^{-1}X^T\mathbf{y}$

2.3 רגריסה לוגיסטית

(1) כל דוגמה מתויגת $y_n \in \{-1, 1\}$ וההיפותזה $h(\mathbf{x}_n) = \theta(\mathbf{w}^T\mathbf{x}_n)$

(א) השתמשנו בפונקצית שגיאה cross-entropy:

$$E_{in}(\hat{\mathbf{w}}) = \frac{1}{N}\sum_{n=1}^{N}\log(1 + e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n})$$

(i) וכדי למצוא את הפתרון השתמשנו ב- Gradient Descent

2.4 הערה חשובה

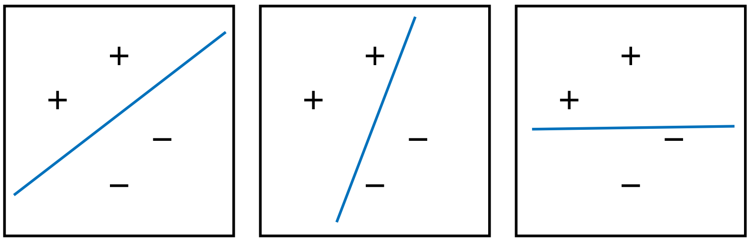
דבר שהמעטנו לדבר עליו הוא האפשרות להשתמש במודלים כמו רגרסיה לינארית לשם סיווג.
נזכור תמיד כי אפשרי להשתמש בחזאי כמו רגרסיה לינארית עבור מאורעות בינאריים ולהתאים פונקציה מהצורה

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

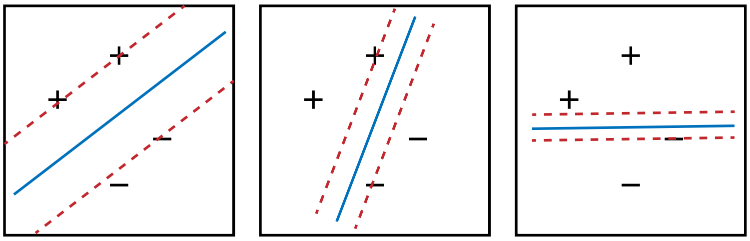
לפלט אשר שייך ל- $\{-1, 1\}$.
במידה ונעשה זאת נקבל חזאי שיניב ערכים שקרובים מאוד ל- -1 או קרובים מאוד ל- 1 (כי הרי נקבל ערכים רציפים).
כיצד נבצע זאת? נוסיף threshold ונרכיב פונקצית סימן על הפונקציה h.

3 שניה לפני שנמשיך - מוטיבציה ל SVM

בואו ניזכר באלגוריתם הפרספטרון ובעובדה שעבור סט דוגמאות יחיד נוכל לקבל כמה פתרונות **מושלמים**



משמע כל הפתרונות הניבו: $E_{in}(\hat{\mathbf{w}}) = 0$ (השוני נבע כידוע מאתחול האלגוריתם וסדר הצגת הדוגמאות לאלגוריתם).
עם זאת, האינטואיציה אומרת שהמפריד השמאלי ביותר בתמונה מעלה הוא המפריד הלינארי הטוב ביותר כיוון שהמפריד הכי רחוק מהדוגמאות שהכי קרובות אליו.
לכן על אף העובדה שאמנם כל פתרון מהפתרונות הוא מושלם (על פי הפרספטרון) - אנחנו צריכים תמיד לחשוב על מצב של מידע שעדיין לא נצפה ואיך המפריד שלנו יעבוד עליו.
לכן נרצה ליצור מעין שוליים לקו המפריד כך שיהיו כמה שיותר גדולים ומרחקם מהנקודות הכי קרובות יהיה הגדול ביותר שנוכל להשיג.



זוהי המוטיבציה שלנו ל - SVM.

4 Support Vector Machine

4.1 כמה נקודות חשובות

- SVM מאוד רובסטי (חסין) לרעשים
- ניתן להוכיח בצורה תאורטית שמובטחת לנו הכללה
- מציאת ה-SVM יכולה להיעשות בצורה יעילה
- המודל מאוד נפוץ ומשתמשים בו המון
- ניתן להרחיבו לסיווג לא לינארי (כמו במקרה של רגרסיה לינארית)

4.2 בידוד ה-bias

עד כה דיברנו על מצב בו מרחב הקלט שלנו היה מהצורה $\mathbf{x} \in \{1\} \times \mathbb{R}^d$ מרחב הפלט היה מהצורה $y \in \{-1, 1\}$ וסט המשקלים שלנו היה מהצורה $\mathbf{w} \in \mathbb{R}^{d+1}$:

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$$

במודל ה-SVM נרצה להפריד את ה-bias משאר המשקולות ולכן יהיה $b \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$ ו- $\mathbf{w} \in \mathbb{R}^d$ אשר יקיימו:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

ההיפותזה שלנו תהיה מהצורה:

$$h(\mathbf{x}) = sign\left(\mathbf{w}^T\mathbf{x} + b\right)$$

4.3 הקדמה - Spearating Hyperplane

הערה. בשלב הראשון נניח שסט הדוגמאות שלנו $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ ספרבילי באופן לינארי - ניתן להפרידו בעזרת hyperplane; במילים אחרות קיים היפרפליין $\mathbf{w}^T\mathbf{x} + b = 0$ עבורו לכל $n = 1, ..., N$ מתקיים $sign\left(\mathbf{w}^T\mathbf{x}_n + b\right) = y_n$ וכמובן מכך נובע ש:

$$y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) > 0 \quad \forall n = 1, .., N$$

4.4 שולי המישור המפריד - Margin of a Hyperplane

חישוב שולי ה-Hyperplane דורש מאיתנו לחשב את המרחק מהמפריד לבין הנקודה הכי קרובה אליו בקבוצת האימון.
לצורך כך נרצה קודם לחשב מרחק של נקודה כלשהי מהמפריד.
נסמן את L בתור ה-Hyperplane
נתבונן ב-2 נקודות $\mathbf{x}'$ ו- $\mathbf{x}''$ שנמצאות על L, מההנחה שהן על L הן מקיימות:

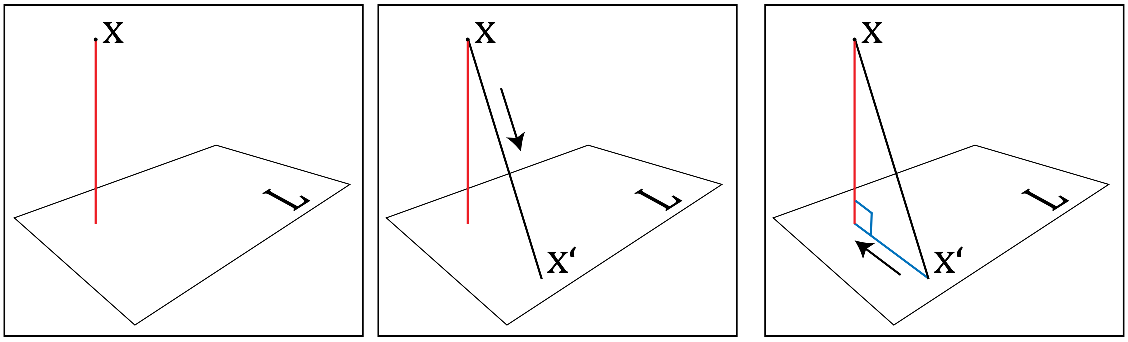
$$\left.\begin{array}{l} \mathbf{w}^T\mathbf{x}' + b = 0 \\ \mathbf{w}^T\mathbf{x}'' + b = 0 \end{array}\right\} \Rightarrow \mathbf{w}^T(\mathbf{x}' - \mathbf{x}'') = 0 \quad \forall \mathbf{x}', \mathbf{x}'' \in L$$

ולכן כיוון שהדבר נכון לכל 2 נקודות שכאלו הרי ש $\mathbf{w}$ **אורתוגונלי** לכל וקטור ב- L ובעזרת עובדה זו נוכל לחשב את המרחק בין דוגמה כלשהי $\mathbf{x}$ ל-L.

הערה חשובה - וקטור ב L הוא לא וקטור שמתחיל בראשית הצירים אלא **נמצא על** L ובעצם מורכב מהפרש בין 2 נקודות על L כפי שראינו עם $\mathbf{x}' - \mathbf{x}''$

הוקטור $\mathbf{x}'$ **שונה** מהוקטור $\mathbf{x}$ שמתחיל בראשית הצירים!

לשם כך נבחר וקטור כלשהו $\mathbf{x}' \in L$, אותו נפחית מהוקטור $\mathbf{x}$ ונטיל על $\mathbf{w}$ - כך נקבל את המרחק של $\mathbf{x}$ מ- L:



- יהי $\mathbf{x}' \in L$
- נחשב את ההיטל האורתוגונלי של $(\mathbf{x} - \mathbf{x}')$ על $\mathbf{w}$:

$$\left| \frac{\mathbf{w}^T}{\|\mathbf{w}\|} \cdot (\mathbf{x} - \mathbf{x}') \right| = \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{x}' \right|$$

$$\mathbf{w}^T \mathbf{x}' + b = 0 \quad = \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x} + b \right|$$

- ולכן:

$$d(\mathbf{x}, \mathbf{w}, b) = \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x} + b \right|$$

- כעת נגדיר את בעיית האופטימיזציה שלנו:
  - כעת באופן יותר ספציפי, נרצה למצוא את הנקודה $\mathbf{x}_n$ מסט האימון שמרחקה מה-Hyperplane הוא מינימלי:

$$\min_{n=1,..,N} \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x}_n + b \right|$$

  - על כך נרצה למצוא $\mathbf{w}$ ו- b (אשר מגדירים את ה-Hyperplane) אשר יניבו לנו מרחק מקסימלי שכזה מה-Hyperplane:

$$\max_{\mathbf{w},b} \min_{n=1,..,N} \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x}_n + b \right|$$

  - לבסוף נוסיף את ההכרח שאותו Hyperplane אכן יפריד את הדוגמאות בצורה מושלמת ונקבל את בעית האופטימיזציה:

$$\boxed{\max_{\mathbf{w},b} \left\{ \min_{n=1,..,N} \frac{1}{\|\mathbf{w}\|} \cdot \left| \mathbf{w}^T \mathbf{x}_n + b \right| \right\}, \quad s.t \;\; y_n \left(\mathbf{w}^T \mathbf{x}_n + b \right) > 0 \;\; , \forall n = 1, .., N}$$

הערה. דרך נוספת לחישוב המרחק בין $\mathbf{x}$ ל-L היא מציאת נקודה $\mathbf{x}' \in L$ כלשהי עבורה:

$$\min_{\mathbf{x}'} \left\| \mathbf{x} - \mathbf{x}' \right\|^2 \quad s.t \;\; \mathbf{w}^T \mathbf{x}' + b = 0$$

4.5 The Maximum-Margin Separating Hyperplane

הערה. במידה וקיים Hyperplane אשר מוגדר בתור $\mathbf{w}^T\mathbf{x}+b=0$ ונחלקו בערך $\rho>0$ כלשהו - נשמור על אותו Hyperplane.
כיוון שעבור $\rho>0$, $\tilde{\mathbf{w}}=\frac{\mathbf{w}}{\rho}$ ו-$\tilde{b}=\frac{b}{\rho}$:

$$\mathbf{w}^T\mathbf{x}+b>0\Rightarrow\tilde{\mathbf{w}}^T\mathbf{x}+\tilde{b}>0$$

כמו-כן: במידה ונחלקו ב $\rho<0$ נקבל סיווג הפוך (כל מה שחיובי יתוייג כשלילי ולהיפך)

נבחר ρ כך ש:

$$\rho=\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)>0$$

נבחין כי $\rho>0$ כי בהינתן Hyperplane שאכן מפריד לנו את הדוגמאות בצורה מושלמת נקבל ש- $y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)>0$ לכל $n=1,..,N$
נגדיר כעת:

$$\tilde{\mathbf{w}}=\frac{\mathbf{w}}{\rho},\quad\tilde{b}=\frac{b}{\rho}$$

ולכן:

$$\begin{aligned}\min_{n=1,..,N}y_n\cdot\left(\tilde{\mathbf{w}}^T\mathbf{x}_n+\tilde{b}\right)&=\min_{n=1,..,N}\frac{y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)}{\rho}\\&=\frac{\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)}{\rho}\\&=\frac{\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)}{\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)}=1\end{aligned}$$

מהערה הפותחת וממה שהראינו נסיק שהחלוקה ב-ρ שווה ערך לכך שקיימים $\mathbf{w}$ ו-b שמייצגים Hyperplane המפריד את הדוגמאות בצורה מושלמת כך ש:

$$\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)=1$$

באופן אינטואיטיבי - מהחלוקה של $\mathbf{w}$ ו-b ב- $\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)$ מצאנו שאחת הדוגמאות לפחות תניב את הערך 1 וכל שאר הדוגמאות יהיו גדולות או שוות ל-1.

הגדרה 1. מעתה והלאה כשנרצה להגיד שה-Hyperplane שלנו מפריד בצורה לינארית מושלמת את הדוגמאות נתייחס לתנאי השקול ל- $y_n\left(\mathbf{w}^T\mathbf{x}_n+b\right)>0$:

$$\min_{n=1,..,N}y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)=1$$

נתבונן שוב בחלק מביטוי המרחק ונשתמש במה שמצאנו:

$$\begin{aligned}\left|\mathbf{w}^T\mathbf{x}_n+b\right|&\overset{y_n\in\{-1,1\}}{=}\left|y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)\right|\\(*)&=y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)\end{aligned}$$

$(*)$ ה-Hyperplane מפריד את הנקודות בצורה מושלמת ולכן $y_n\cdot\left(\mathbf{w}^T\mathbf{x}_n+b\right)>0$

נחזור כעת לבעיית האופטימיזציה:

$$\underset{\mathbf{w},b}{max}\left\{\underset{n=1,..,N}{min}\frac{1}{\|\mathbf{w}\|}\cdot\left|\mathbf{w}^{T}\mathbf{x}_{n}+b\right|\right\},\ \ s.t\ \ y_{n}\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)>0\ \ ,\forall n=1,..,N$$

מכך שמדובר ב- $y_{n}\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)>0\ \ ,\forall n=1,..,N$ הרי שכבר הוכחנו ש

$$\left|\mathbf{w}^{T}\mathbf{x}_{n}+b\right|=y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)$$

ולכן **מהגדרה 1**:

$$\underset{\mathbf{w},b}{max}\left\{\frac{1}{\|\mathbf{w}\|}\cdot\underset{n=1,..,N}{min}y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)\right\},\ \ s.t\ \ \underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)=1$$

וממה שהוכחנו לגבי $\underset{n=1,..,N}{min}y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)$ והשימוש ב-ρ מתאים נסיק:

$$\underset{\mathbf{w},b}{max}\left\{\frac{1}{\|\mathbf{w}\|}\right\},\ \ s.t\ \ \underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)=1$$

באותה מידה, המקסימום של $\frac{1}{\|\mathbf{w}\|}$ זה כמו המינימום של $\|\mathbf{w}\|$ ולכן:

$$\begin{aligned}&\underset{\mathbf{w},b}{min}\ \|\mathbf{w}\|,\ \ s.t\ \ \underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)=1\\ =&\underset{\mathbf{w},b}{min}\ \ \mathbf{w}^{T}\mathbf{w},\ \ s.t\ \ \underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)=1\\ =&\underset{\mathbf{w},b}{min}\ \ \frac{1}{2}\mathbf{w}^{T}\mathbf{w},\ \ s.t\ \ \underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)=1\end{aligned}$$

נוכיח כעת כי $y_{n}\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)\ngtr 1$

- נתבונן בבעיה יותר פשוטה:

$$\underset{\mathbf{w},b}{min}\ \frac{1}{2}\left\|\mathbf{w}\right\|^{2},\ \ s.t\ \ y_{n}\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)\geq1,\ \ \forall n=1,..,N$$

- נניח בשלילה שהפתרון שלה הוא $\mathbf{w}^{*},b^{*}$ וגם שמתקיים $y_{n}\left(\mathbf{w}^{*T}\mathbf{x}_{n}+b^{*}\right)>1$ לכל $n=1,..,N$
- נבחר $\rho^{*}=\underset{n=1,..,N}{min}\ y_{n}\cdot\left(\mathbf{w}^{*T}\mathbf{x}_{n}+b^{*}\right)>1$ ונגדיר:

$$\mathbf{w}=\frac{\mathbf{w}^{*}}{\rho^{*}},\quad b=\frac{b^{*}}{\rho^{*}}$$

- נקבל:

$$y_{n}\left(\mathbf{w}^{T}\mathbf{x}_{n}+b\right)\geq1$$

כאשר לפחות אחד מהדוגמאות מקיימת שוויון!

- מצאנו כבר שהחלוקה של Hyperplane בקבוע לא משנה אותו ולכן מצאנו סתירה להנחת השלילה.

הערה. מכך נסיק ש:

$$\|\mathbf{w}\|=\frac{1}{\rho^{*}}\cdot\|\mathbf{w}^{*}\|\underbrace{<}_{\rho^{*}>1}\|\mathbf{w}^{*}\|$$

הערה. $\mathbf{w}^{*}\neq0$ כיוון שאם $\mathbf{w}^{*}=0$ הרי שהחזאי היה רק b קבוע והיה קובע את כל הדוגמאות בתור אותו סיווג מה ששובר את ההנחה שיש לפחות 2 דוגמאות עם סיווג שונה.

ממה שהוכחנו עד מצאנו ש:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w},\ \ s.t\ \min_{n=1,..,N} y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right) = 1$$

$$\equiv$$

$$\min_{\mathbf{w},b} \frac{1}{2}\left\|\mathbf{w}\right\|^2,\ \ s.t\ \ y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1,\ \ \forall n = 1,..,N$$

כעת נוכל להתבונן בבעית אופטימיזציה יותר נוחה שמגדירה לנו את ה- Hard-Margin SVM:

HARD-MARGIN SVM

$$\min_{\mathbf{w},b} \frac{1}{2}\left\|\mathbf{w}\right\|^2,\ \ s.t\ \ y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1,\ \ \forall n = 1,..,N$$

דוגמה. יהיו $y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$, $X = \begin{array}{c} \begin{matrix} \mathbf{x}_1 & \mathbf{x}_2 \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \end{array}$, נרצה לפתור עבור $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$:

$$\min \frac{1}{2} \cdot \underbrace{\left(w_1^2 + w_2^2\right)}_{\|\mathbf{w}\|^2},\ \ s.t\ \ y_n \cdot \left(\mathbf{w}^T\mathbf{x}_n + b\right)$$

נרשום בצורה מפורשת את האילוצים:

$$-1 \cdot (\mathbf{w}^T \cdot \mathbf{0} + b) = -1 \cdot (b) \geq 1 \quad (1)$$
$$-1 \cdot (2w_1 + 2w_2 + b) \geq 1 \quad (2)$$
$$+1 \cdot (2w_1 + b) \geq 1 \quad (3)$$
$$+1 \cdot (3w_1 + b) \geq 1 \quad (4)$$

נתבונן במשוואות 2 ו- 3:

$$(2) + (3) \Rightarrow -2w_2 \geq 2 \Rightarrow \boxed{w_2 \leq -1}$$

ועל 1 ו- 3:

$$(1) + (3) \Rightarrow 2w_1 \geq 2 \Rightarrow \boxed{w_1 \geq 1}$$

מצאנו 2 אילוצים שיכולים להניב לנו את הפתרון:

$$w_1^* = 1,\ \ w_2^* = -1$$

נציב את ערכי $\mathbf{w}$ ונקבל שבהכרח $b^* = -1$ ולכן: $\boxed{w_1^* = 1,\ \ w_2^* = -1,\ \ b^* = -1}$

לבסוף הקו המפריד:

$$g(\mathbf{x}) = sign(x_1 - x_2 - 1) \Rightarrow x_1 - x_2 - 1 = 0 \Rightarrow x_2 = x_1 - 1$$

עבור $x_2 = 0$ נקבל $x_1 = 1$ ועבור $x_1 = 0$ נקבל $x_2 = -1$ שם יעבור הישר, כמו-כן אם נציב את התשובה נראה ש-3 הנקודות הראשונות יושבות על ה-margin.