

תרגול 6 - SVM ופונקציות גרעין (Kernels)

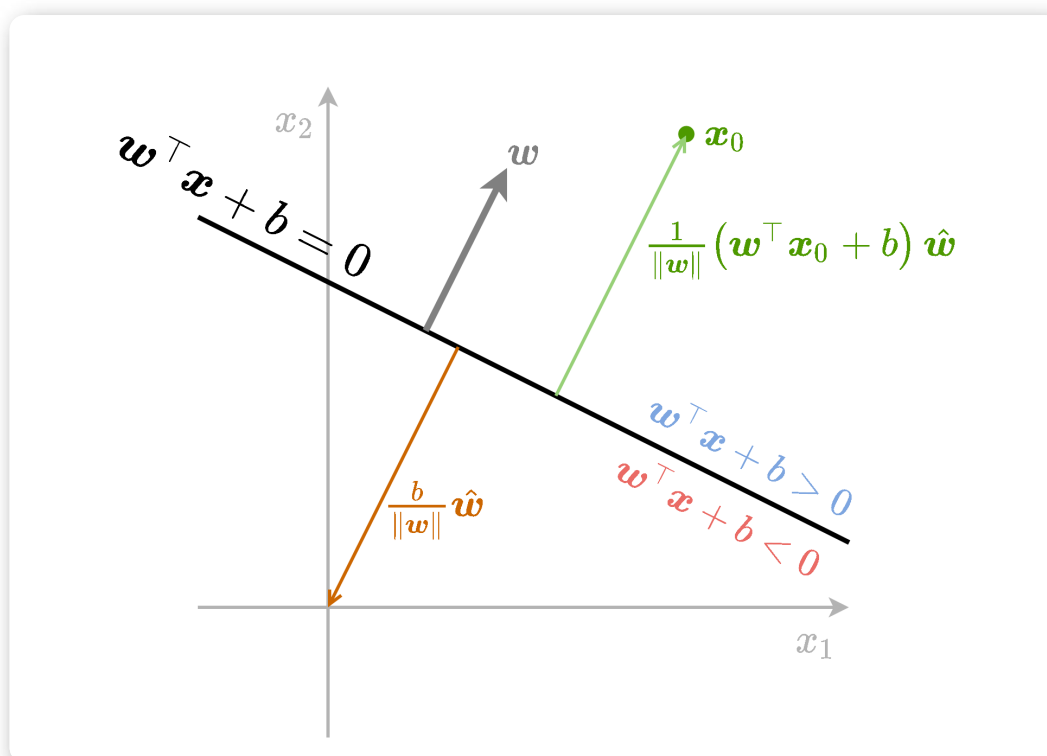
Slides

PDF

תקציר התיאוריה

הערה: בפרק זה נעסוק בסיווג בינארי ונסמן את התוויות של שתי המחלקות ב 1 ו -1 .

תזכורת - גאומטריה של המישור



דרך נוחה לתאר מישור במרחב (ממימד כלשהו) היא על ידי משוואה מהצורה $w^T x + b = 0$. מישור שכזה מחלק את המרחב לשניים ומגדיר צד חיובי ושלילי של המרחב. נתאר מספר תכונות של הצגה זו:

- w הוא הנורמל למישור אשר מגדיר את האוריינטציה שלו וגם את הצד החיובי.
- המרחק של המישור מהראשית הינו $\frac{b}{\|w\|}$. הסימן של גודל זה מצוין איזה צד של המישור נמצאת הראשית.
- המרחק של נקודה כל שהיא x_0 מהמישור הינה $\frac{1}{\|w\|} (w^T x_0 + b)$. הסימן של גודל זה מצוין את הצד של המישור בו נמצאת הנקודה.
- בעבור מישור נתון כל שהוא w ו b מוגדרים עד כדי קבוע. זאת אומרת שהמישור אינווריאנטי (לא משתנה) תחת שינוי של פרמטרים מהצורה של: $w \rightarrow \alpha w, b \rightarrow \alpha b$.

מסווג לינארי

מסווג לינארי הוא מסווג מהצורה של

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b) = \begin{cases} 1 & \mathbf{w}^\top \mathbf{x} + b > 0 \\ -1 & \text{else} \end{cases}$$

עם \mathbf{w} ו b כלשהם.

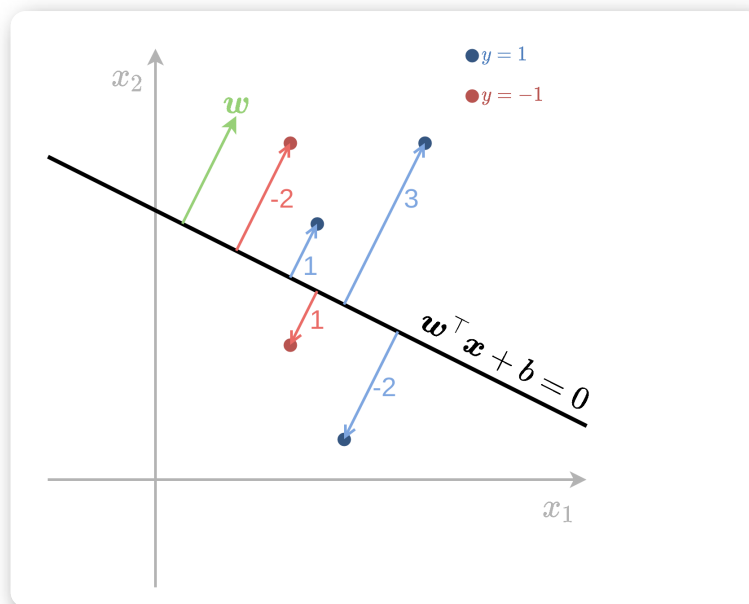
זאת אומרת שמסווג מחלק את המרחב לשני צידי של המישור $\mathbf{w}^\top \mathbf{x} + b = 0$ המכונה מישור ההפרדה.

Signed distance (מרחק מסומן)

נסתכל על בעיית סיווג בינארית עם תוויות $y = \pm 1$. בעבור משטח הפרדה כלשהו נגדיר את ה signed distance של דגימה כלשהי ממשטח ההפרדה באופן הבא:

$$d = \frac{1}{\|\mathbf{w}\|} (\mathbf{w}^\top \mathbf{x} + b)y$$

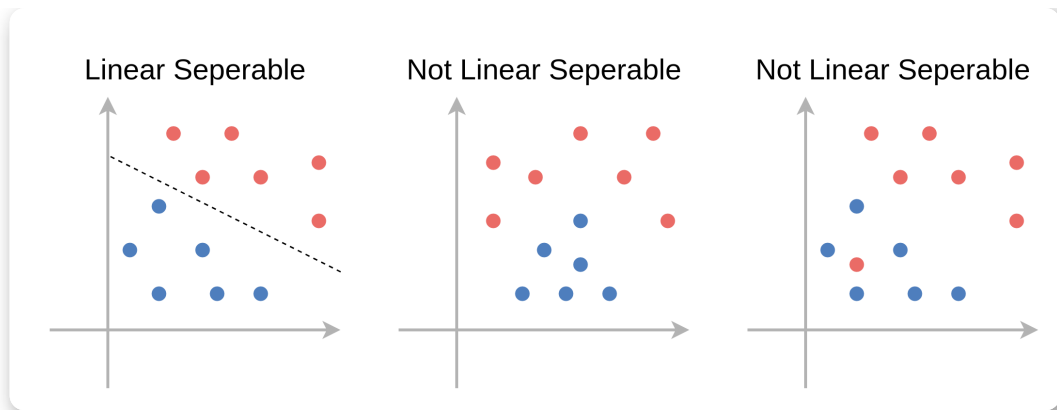
זהו המרחק של נקודה מהמישור כאשר המרחק של נקודות עם תווית $y = 1$ הוא חיובי כאשר הן בצד החיובי של המישור ושלילי אחרת והפוך לגבי נקודות עם תווית $y = -1$.



פרידות לינארית (linear separability)

בבעיות של סיווג בינארי, אנו נאמר על מדגם שהוא פריד לינארית אם קיים מסווג לינארי אשר מסווג את המדגם בצורה מושלמת (בלי טעויות סיווג).

כאשר המדגם פריד לינארית יהיה יותר ממסווג לינארי אחד אשר יכול לסווג בצורה מושלמת את המדגם.



(Support Vector Machine (SVM

SVM הוא אלגוריתם דיסקרימינטיבי לסיווג בינארי אשר מחפש מסווג לינארי אשר יסווג בצורה טובה את המדגם.

Hard SVM

Hard SVM מתייחס למקרה שבו המדגם הוא פריד לינארית. באלגוריתם זה נחפש את המסווג הלינארי אשר בעבורו ה signed distance המינימאלי על ה train set הוא המקסימאלי:

$$\mathbf{w}^*, b^* = \arg \max_{\mathbf{w}, b} \min_i \left\{ \frac{1}{\|\mathbf{w}\|} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) y^{(i)} \right\}$$

ניתן להראות כי במקרה שבו המדגם פריד לינארי, בעיה זו שקולה לבעיית האופטימיזציה הבאה:

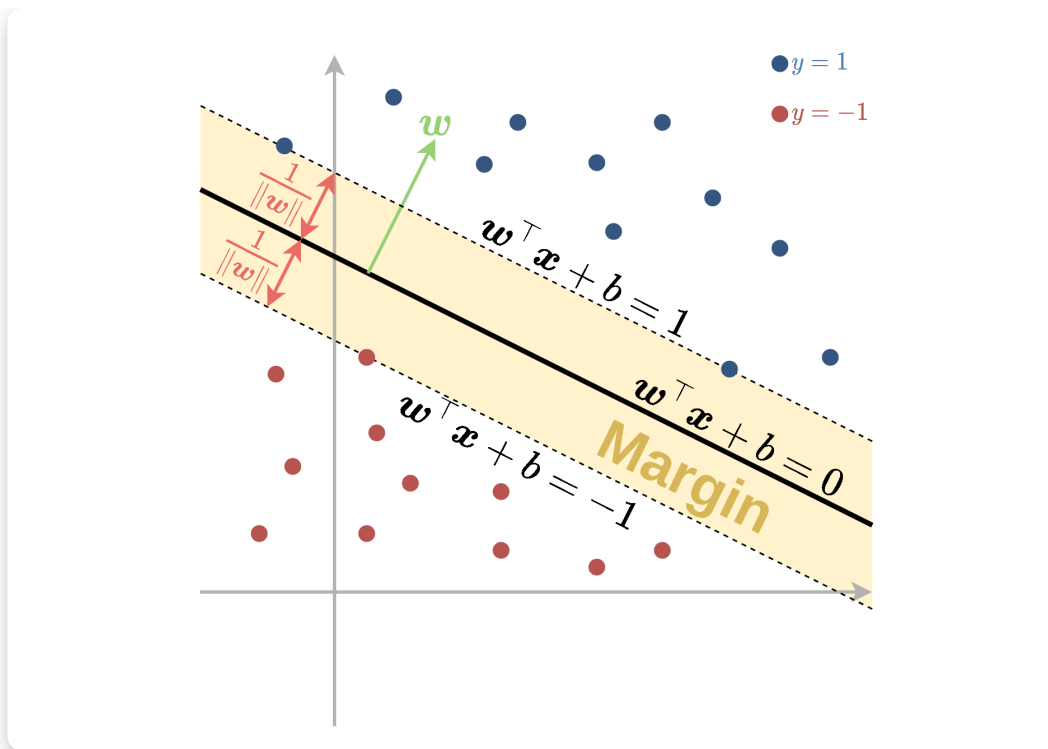
$$\begin{aligned} \mathbf{w}^*, b^* = \arg \min_{\mathbf{w}, b} & \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} & \quad y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \quad \forall i \end{aligned}$$

בעיית אופטימיזציה זו מכונה הבעיה הפרימאליית.

Margin

בכדי להבין את המשמעות של בעיית האופטימיזציה שקיבלנו נגדיר את המושג שוליים (margin) של מסווג. האיזור של ה margin מוגדר כאיזור סביב משטח ההפרדה אשר נמצא בתחום:

$$1 \geq \mathbf{w}^\top \mathbf{x} + b \geq -1$$



הרוחב של איזור זה נקבע על פי הגודל של הוקטור w ושווה ל $\frac{2}{\|w\|}$.

האילוץ $y^{(i)} (w^T x^{(i)} + b) \geq 1$, אשר מופיע בבעיית האופטימיזציה הפרימאלית, דורש למעשה שכל הנקודות יסווגו נכונה וימצאו מחוץ ל margin. בעיית האופטימיזציה מחפשת את הפרמטרים של המישור בעל ה margin הגדול ביותר אשר מקיים תנאי זה.

Support Vectors

בעבור פתרון מסוים של בעיית האופטימיזציה, ה support vectors מוגדרים כנקודות אשר יושבות על השפה של ה margin, נקודות אלו מקיימות $y^{(i)} (w^T x^{(i)} + b) = 1$. אלו הנקודות אשר ישפיעו על הפתרון של בעיית האופטימיזציה, זאת אומרת שהסרה או הזזה קטנה של נקודות שאינן support vectors לא תשנה את הפתרון.

הבעיה הדואלית

דרך שקולה נוספת לרישום בעיית האופטימיזציה הינה על ידי הגדרת N משתני עזר נוספים $\{\alpha_i\}_{i=1}^N$. בעזרת משתנים אלו ניתן לרשום את בעיית האופטימיזציה באופן הבא:

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j x^{(i)T} x^{(j)} \\ \text{s.t. } &\alpha_i \geq 0 \quad \forall i \\ &\sum_i \alpha_i y^{(i)} = 0 \end{aligned}$$

מתוך המשתנים $\{\alpha_i\}_{i=1}^N$ ניתן לשחזר את w אופן הבא:

$$w = \sum_i \alpha_i y^{(i)} x^{(i)}$$

תכונות:

•	•	•
$\alpha_i = 0$	$y^{(i)} (w^T x^{(i)} + b) > 1$	נקודות רחוקות מה margin

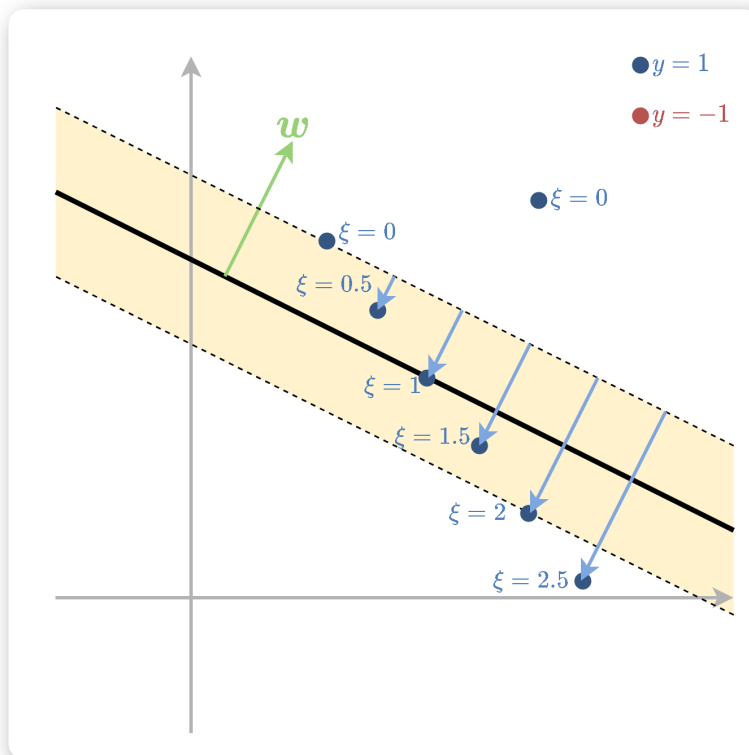
•	•	•
$\alpha_i \geq 0$	$y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$	נקודות על ה margin (שהם support vectors)

- אם $\alpha_i > 0$ אז $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$ (אבל לא להיפך)
- אם $\alpha_i = 0$ אז $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) > 1$ (אבל לא להיפך)

את b ניתן לחשב על ידי בחירת support vector אחד ולחלץ את b מתוך $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$.

Soft SVM

Soft SVM מתייחס למקרה שבו המדגם אינו פריד לינארית. במקרה זה עדיין מגדירים את ה margin בצורה דומה אך מאפשרים למשתנים להיכנס לתוך ה margin ואף לחצות אותו לצד הלא נכון של מישור ההפרדה. על כל חריגה כזו משלמים קנס ב objective שאותו מנסים למזער. את החריגה של הדגימה ה i נסמן ב ξ_i . לנקודות שהן בצד הנכון של המישור ומחוץ ל margin יהיה $\xi_i = 0$.



המשתנים ξ_i נקראים **slack variables** ובעיית האופטימיזציה הפרימאלית תהיה:

$$\mathbf{w}^*, b^*, \{\xi_i\}^* = \arg \min_{\mathbf{w}, b, \{\xi_i\}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i$$

$$\text{s.t. } y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

כאשר C הוא hyper-parameter אשר קובע את גודל הקנס שאותו ה objective נותן על כל חריגה.

הבעיה הדואלית הינה:

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \mathbf{x}^{(i)\top} \mathbf{x}^{(j)} \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i \\ & \sum_i \alpha_i y^{(i)} = 0 \end{aligned}$$

ה support vectors מוגדרות להיות הנקודות שמקיימות $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1 - \xi_i$

תכונות:

$\alpha_i = 0$	$y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) > 1$	נקודות שמוסוגות נכון ורחוקות מה margin
$0 \leq \alpha_i \leq C$	$y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$	נקודות על ה margin (שהם support vectors)
$\alpha_i = C$	$y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1 - \xi_i$	נקודות שחורגות מה margin (גם support vectors)

פונקציות גרעין

מאפיינים: תזכורת

נוכל תמיד להחליף את וקטור המשתנים \mathbf{x} שעליו פועל האלגוריתם בוקטור חדש $\mathbf{x}_{\text{new}} = \Phi(\mathbf{x})$, כאשר Φ היא פונקציה אשר נבחרה מראש ונקראת פונקציית המאפיינים שכן היא מחלצת מאפיינים רלוונטים מתוך \mathbf{x} שבהם נשתמש.

פונקציות גרעין

במקרים רבים החישוב של $\Phi(\mathbf{x})$ יכול להיות מסובך, אך קיימת דרך לחשב בצורה יעילה את הפונקציה $K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1)^\top \Phi(\mathbf{x}_2)$ אשר נקראת פונקציית גרעין. ישנם מקרים שבהם וקטור המאפיינים יהיה אין סופי.

ישנם קריטריונים תחתם פונקציה מסוימת $K(\mathbf{x}_1, \mathbf{x}_2)$ היא פונקציית גרעין בעבור וקטור מאפיינים מסויים. בקורס זה לא נכנס לתנאים אלו. נציג שתי פונקציות גרעין נפוצות:

- גרעין גאוס: $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{2\sigma^2}\right)$ כאשר σ פרמטר שיש לקבוע.
- גרעין פולינומיאלי: $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^p$ כאשר $p \geq 1$ פרמטר שיש לקבוע.

פונקציות המאפיינים שמתיימות לגרעינים אלו הן מסורבלות לכתיבה ולא נציג אותן כאן.

Kernel Trick in SVM

הרעיון ב kernel trick הינו להשתמש ב SVM עם מאפיינים מבלי לחשב את Φ באופן ישיר על ידי שימוש בפונקציית גרעין. בעבור פונקציית מאפיינים Φ עם פונקציית גרעין K הבעיה הדואלית של SVM הינה:

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \quad \forall i \\ & \sum_i \alpha_i y^{(i)} = 0 \end{aligned}$$

בעיית אופטימיזציה זו מגדירה את המשתנים $\{\alpha_i\}$ בלי צורך לחשב את Φ באופן מפורש בשום שלב.

הפרמטר \mathbf{w} נתון על ידי:

$$\mathbf{w} = \sum_i \alpha_i y^{(i)} \Phi(\mathbf{x}^{(i)})$$

כדי לחשב את \mathbf{w} באופן מפורש יש לחשב את Φ אך ניתן להמנע מכך אם מציבים את הנוסחה ל \mathbf{w} ישירות לתוך המסווג:

$$\begin{aligned} h(\mathbf{x}) &= \text{sign}(\mathbf{w}^\top \Phi(\mathbf{x}) + b) \\ &= \text{sign}\left(\sum_i \alpha_i y^{(i)} \Phi(\mathbf{x}^{(i)})^\top \Phi(\mathbf{x}) + b\right) \\ &= \text{sign}\left(\sum_i \alpha_i y^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}) + b\right) \end{aligned}$$

כך שגם בשלב החיזוי ניתן להשתמש בפונקציית הגרעין בלי לחשב את Φ באופן מפורש.

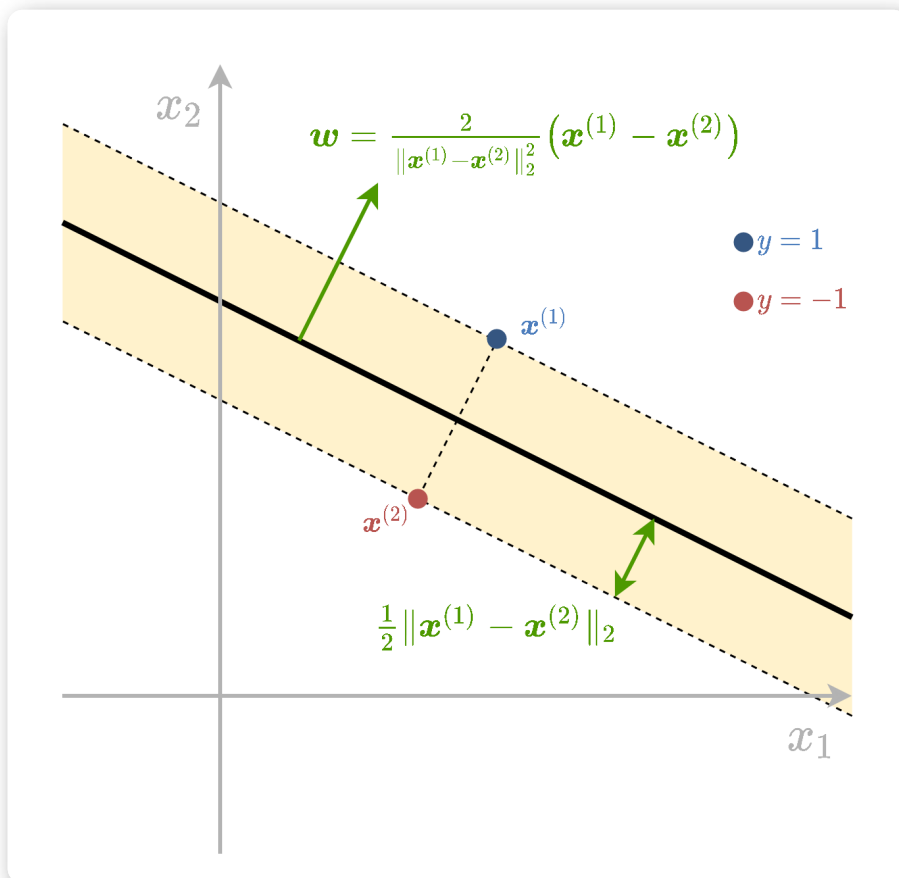
תרגיל 6.1 - Support Vectors 2

בשאלה זו נראה שבעבור מדגם המכיל 2 נקודות, אחת מכל מחלקה, הפתרון של בעיית hard SVM יהיה - מישור הפרדה שיעבור בדיוק במרכז בין 2 הנקודות, ויהיה ניצב לוקטור המחבר את שתי הנקודות.

בפועל נראה ש:

$$\mathbf{w} = \frac{2}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2} (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$$

$$b = -\frac{1}{2} (\mathbf{x}^{(1)} + \mathbf{x}^{(2)})^\top \mathbf{w}$$



1 הראו זאת על ידי פתרון הבעיה הדואלית.

פתרון 6.1

(1)

הבעיה הדואלית הינה:

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \mathbf{x}^{(i)\top} \mathbf{x}^{(j)} \\ \text{s.t. } & \alpha_i \geq 0 \quad \forall i \\ & \sum_i \alpha_i y^{(i)} = 0 \end{aligned}$$

נניח בלי הגבלת הכלליות ש $y^{(1)} = 1$ ו $y^{(2)} = -1$.

נסתכל על האילוץ בשורה האחרונה:

$$\begin{aligned} \sum_{i=1}^2 \alpha_i y^{(i)} &= 0 \\ \Leftrightarrow \alpha_1 - \alpha_2 &= 0 \\ \Leftrightarrow \alpha_1 = \alpha_2 &= \alpha \end{aligned}$$

מכאן שהמשתנה היחיד בבעיה הינו α . בעיית האופטימיזציה (ללא האילוץ) תהיה:

$$\begin{aligned} \alpha &= \arg \max_{\alpha} 2\alpha - \frac{\alpha^2}{2} \sum_{i,j} y^{(i)} y^{(j)} \mathbf{x}^{(i)\top} \mathbf{x}^{(j)} \\ \text{s.t. } & \alpha \geq 0 \\ &= \arg \max_{\alpha} 2\alpha - \frac{\alpha^2}{2} \left(\|\mathbf{x}^{(1)}\|_2^2 - \mathbf{x}^{(1)\top} \mathbf{x}^{(2)} - \mathbf{x}^{(2)\top} \mathbf{x}^{(1)} + \|\mathbf{x}^{(2)}\|_2^2 \right) \\ \text{s.t. } & \alpha \geq 0 \\ &= \arg \max_{\alpha} 2\alpha - \frac{\alpha^2}{2} \|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2 \\ \text{s.t. } & \alpha \geq 0 \end{aligned}$$

ניתן לפתור את הבעיה על ידי גזירה והשוואה ל:0:

$$\begin{aligned} \frac{d}{d\alpha} 2\alpha - \frac{\alpha^2}{2} \|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2 &= 0 \\ \Leftrightarrow 2 &= \alpha \|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2 \\ \Leftrightarrow \alpha &= \frac{2}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2} \end{aligned}$$

את \mathbf{w} נמצא על ידי:

$$\mathbf{w} = \sum_i \alpha_i y^{(i)} \mathbf{x}^{(i)} = \alpha (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) = \frac{2}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2} (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$$

את b מוצאים על ידי בחירת support vector אחד מתוך המשוואה $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$. נסתכל על הנקודה הראשונה:

$$\begin{aligned}
y^{(1)} (\mathbf{w}^\top \mathbf{x}^{(1)} + b) &= 1 \\
\Leftrightarrow \mathbf{w}^\top \mathbf{x}^{(1)} + b &= 1 \\
\Leftrightarrow b = 1 - \mathbf{w}^\top \mathbf{x}^{(1)} &= \frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|_2^2} - \mathbf{w}^\top \mathbf{x}^{(1)} = \mathbf{w}^\top \left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2^2} - \mathbf{x}^{(1)} \right) \\
\Leftrightarrow b = \mathbf{w}^\top \left(\frac{1}{2} (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) - \mathbf{x}^{(1)} \right) \\
\Leftrightarrow b = -\frac{1}{2} (\mathbf{x}^{(1)} + \mathbf{x}^{(2)})^\top \mathbf{w}
\end{aligned}$$

(2)

הבעיה הפרימלית הינה:

$$\begin{aligned}
\mathbf{w}^*, b^* &= \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\
\text{s.t. } & y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \quad \forall i
\end{aligned}$$

במקרה שבו יש רק שתי נקודות, אחת מכל מחלקה, שתיהן בהכרח יהיו support vectors ויקיימו $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) = 1$. ניח בלי הגבלת הכלליות ש $y^{(1)} = 1$ ו $y^{(2)} = -1$. שני האילוצים שנקודות אלו מגדירות הם:

$$\begin{aligned}
& \begin{cases} y^{(1)} (\mathbf{w}^\top \mathbf{x}^{(1)} + b) = 1 \\ y^{(2)} (\mathbf{w}^\top \mathbf{x}^{(2)} + b) = 1 \end{cases} \\
\Leftrightarrow & \begin{cases} (\mathbf{w}^\top \mathbf{x}^{(1)} + b) = 1 \\ -(\mathbf{w}^\top \mathbf{x}^{(2)} + b) = 1 \end{cases} \\
\Leftrightarrow & \begin{cases} \mathbf{w}^\top \mathbf{x}^{(1)} + b = 1 \\ \mathbf{w}^\top \mathbf{x}^{(2)} + b = -1 \end{cases} \\
\Leftrightarrow & \begin{cases} b = -\frac{1}{2} (\mathbf{x}^{(1)} + \mathbf{x}^{(2)})^\top \mathbf{w} \\ \mathbf{w}^\top (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) = 2 \end{cases}
\end{aligned}$$

נוכל להשתמש באילוץ השני ולכתוב בעזרתו את בעיית האופטימיזציה רק על \mathbf{w} :

$$\begin{aligned}
\mathbf{w}^* &= \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 \\
\text{s.t. } & \mathbf{w}^\top (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) = 2
\end{aligned}$$

בעיה זו מחפשת את ה \mathbf{w} בעל האורך המינימאלי כך שהמכפלה הסקלרית שלו עם $(\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$ היא 2. הוקטור הזה יהיה וקטור בכיוון של $(\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$ ובאורך של $2/\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2$ ולכן \mathbf{w} הינו:

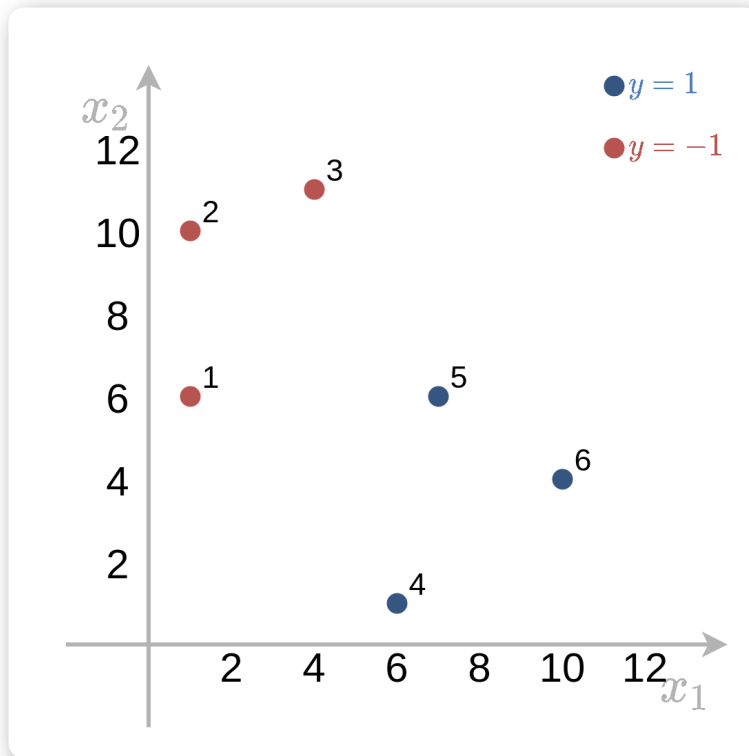
$$\mathbf{w} = \frac{2}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|_2^2} (\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$$

תרגיל 6.2 - Hard SVM

נתון המדגם הבא:

.	1	2	3	4	5	6
y	-1	-1	-1	1	1	1
x ₁	1	1	4	6	7	10

.	1	2	3	4	5	6
x_2	6	10	11	1	6	4



- (1)** מצא את מסווג ה Hard SVM המתאים למדגם זה. מי הם וקטורי התמיכה?
(2) מבלי לפתור את הבעיה הדואלית. אלו ערכים של $\{\alpha_i\}$ בהכרח יתאפסו?
(3) מהו הרוחב של ה margin של הפתרון?

פתרון 6.2

(1)

בעיית האופטימיזציה הפרימאלית אותה נרצה לפתור הינה:

$$\mathbf{w}^*, b^* = \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{s.t. } y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 \quad \forall i$$

באופן כללי לפתור בעיות מסוג זה באופן ידני הוא קשה, אך בעבור מקרים פשוטים נוכל לפתור זאת תוך שימוש בעיקרון הבא:
 במידה ומצאנו את הפתרון האופטימאלי על תת מדגם מתוך המדגם המלא, ומתקיים שבעבור הפתרון שמצאנו, הנקודות שאינן חלק מתת המדגם נמצאות מחוץ ל margin, אז הפתרון הוא בהכרח הפתרון האופטימאלי בעבור המדגם כולו.
 (עקרון זה נכון בגלל העובדה שהוספה של אילוצים לבעיית מינימיזציה יכולה רק להגדיל את ה objective המינימאלי).

המשמעות של עיקרון זה הינו שנוכל לנחש מי הם ה support vectors ולפתור את הבעיה רק בעבור נקודות אלו, תוך התעלמות משאר הנקודות. לאחר שנפתור את הבעיה על הנקודות שניחשנו יהיה עלינו לבדוק אם שאר הנקודות במדגם מחוץ ל margin. אם אכן שאר הנקודות מחוץ ל margin אז פתרנו את הבעיה ואם לא, אז עלינו לנחש נקודות אחרות. בעבור ה support vectors האילוצים של $y^{(i)} (\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1$ הופכים להיות אילוצי שיוויון שאיתם הרבה יותר קל לעבוד.

נסה לנחש מהם ה support vectors. נתחיל ב 0 ונגדיל בהדרגה את כמות ה support vectors.

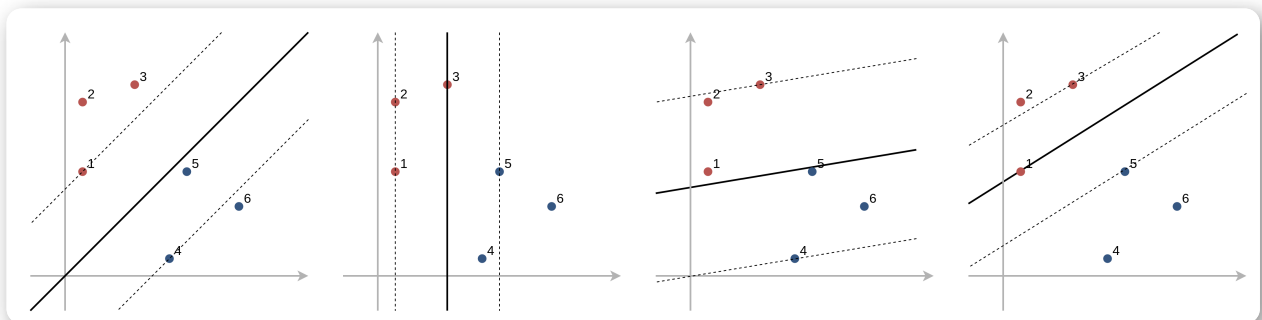
0 or 1 Support Vectors

באופן כללי בעבור מדגם "נורמלי" אשר מכיל נקודות משתי המחלקות שאותן יש לסווג, מקרים אלו לא יכולים להתקיים.

(באופן תיאורטי במקרה המנוון שבו המדגם מכיל רק תוויות מסוג אחד אז ניתן להגדיר בצורת שונות את כמות ה support vectors, אך מקרים אלו לא מאוד רלוונטיים)

2 Support Vectors

על פי התוצאה של התרגיל הקודם אנו יודעים שבעבור שני support vectors שמישור ההפרדה יעבור בדיוק במרכז בין 2 הנקודות, יהיה ניצב לוקטור המחבר את שתי הנקודות והשוליים של ה margin יעברו דרך הנקודות. אם נסתכל על המדגם ונסה לחפש 2 נקודות שיכולות להיות ה support vectors נראה שכל בחירה של זוג נקודות יצור איזור של margin שמכיל נקודה אחרת ולכן לא מקיים את האילוצים. לכן הפתרון לא יכול להכיל רק שני support vectors.

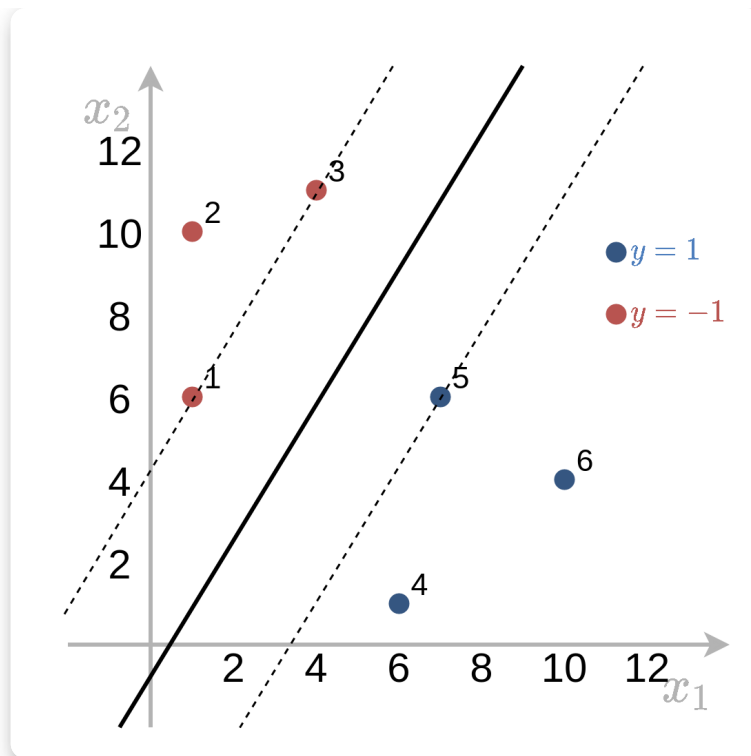


3 Support Vectors

ישנן שתי שלשות שיתכן שיהיו ה support vectors של הפתרון:

- $\{1, 3, 5\}$
- $\{3, 4, 5\}$

בעבור מדגם הכולל את $\{3, 4, 5\}$ ה support vectors יהיו הנקודות 3 ו 5 וכבר ראינו כי נקודות אלו יוצרות פתרון שלא מקיים את האילוצים. לעומת זאת השלשה של $\{1, 3, 5\}$ מגדירה איזור margin שאינו מכיל נקודות אחרות ולכן מקיים את האילוצים:



לכן נקודות אלו יהיו שלושת ה support vectors שיגדיר את הפתרון לבעיה. נחשב את הפתרון המתקבל משלושת הנקודות האלה.

בעבור שלוש support vectors נקבל מתוך האיילוצים 3 משוואות ב-3 נעלמים. בעבור הנקודות {1, 3, 5} משוואות אלו יהיו:

$$\begin{cases} -((1, 6)\mathbf{w} + b) = 1 \\ -((4, 11)\mathbf{w} + b) = 1 \\ ((7, 6)\mathbf{w} + b) = 1 \end{cases}$$

$$\Leftrightarrow \begin{cases} w_1 + 6w_2 + b = -1 \\ 4w_1 + 11w_2 + b = -1 \\ 7w_1 + 6w_2 + b = 1 \end{cases}$$

הפתרון של מערכת המשוואות הזו הינה $b = -\frac{2}{15}$ ו $\mathbf{w} = \frac{1}{15}(5, -3)^T$

(2)

אנו יודעים כי בעבור נקודת שאינן support vectors בהכרח יתאפס. בבעיה זו הנקודות שאינן support vectors הם {2, 4, 6} ולכן $\alpha_2 = \alpha_4 = \alpha_6 = 0$.

(3)

ה margin תלוי בגודל של הוקטור \mathbf{w} והוא שווה ל:

$$\frac{2}{\|\mathbf{w}\|} = \frac{2 \cdot 15}{\sqrt{5^2 + 3^2}} = 5.145$$

תרגיל 6.3 - גרעין גאומטרי

נתון מדגם המכיל 2 נקודות - אחת מכל מחלקה:

$$\begin{aligned} \mathbf{x}^{(1)} &= (1, 1)^\top, & \mathbf{y}^{(1)} &= +1 \\ \mathbf{x}^{(2)} &= (-1, -1)^\top, & \mathbf{y}^{(2)} &= -1 \end{aligned}$$

חשבו את המסווג המתקבל מ Hard SVM עם גרעין גאומי מהצורה $K(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$.

6.3 פתרון

הבעיה הדואלית עם הגרעין הגאומי הינה:

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \mathbf{y}^{(i)} \mathbf{y}^{(j)} \alpha_i \alpha_j K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \\ \text{s.t. } & \alpha_i \geq 0 \quad \forall i \\ & \sum_i \alpha_i \mathbf{y}^{(i)} = 0 \end{aligned}$$

בדומה לתרגיל הראשון נקבל ש:

$$\alpha_1 = \alpha_2 = \alpha$$

נחשב את הערכים של $K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$:

$$\begin{aligned} K(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) &= \exp(0) = 1 \\ K(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) &= \exp(0) = 1 \\ K(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \exp(-(2^2 + 2^2)) = e^{-8} \end{aligned}$$

בעיית האופטימיזציה הינה:

$$\begin{aligned} \alpha^* &= \arg \max_{\alpha} 2\alpha - \frac{\alpha^2}{2} \left(K(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) - 2K(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) + K(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) \right) \\ \text{s.t. } & \alpha_i \geq 0 \quad \forall i \\ &= \arg \max_{\alpha} 2\alpha - \alpha^2(1 - e^{-8}) \\ \text{s.t. } & \alpha_i \geq 0 \quad \forall i \end{aligned}$$

נגזור ונשווה ל-0:

$$\begin{aligned} \frac{d}{d\alpha} 2\alpha - \alpha^2(1 - e^{-8}) &= 0 \\ \Leftrightarrow 1 &= \alpha(1 - e^{-8}) \\ \Leftrightarrow \alpha &= \frac{1}{1 - e^{-8}} \end{aligned}$$

הוקטור \mathbf{w} נתון על ידי: $\mathbf{w} = \sum_i \alpha_i \mathbf{y}^{(i)} \Phi(\mathbf{x}^{(i)})$. נכתוב את הביטוי $\mathbf{w}^\top \Phi(\mathbf{x})$ בעבור נקודה כלשהי \mathbf{x} :

$$\begin{aligned} \mathbf{w}^\top \Phi(\mathbf{x}) &= \sum_i \alpha_i \mathbf{y}^{(i)} \Phi(\mathbf{x}^{(i)})^\top \Phi(\mathbf{x}) \\ &= \sum_i \alpha_i \mathbf{y}^{(i)} K(\mathbf{x}^{(i)}, \mathbf{x}) \\ &= \frac{1}{1 - e^{-8}} \left(K(\mathbf{x}^{(1)}, \mathbf{x}) - K(\mathbf{x}^{(2)}, \mathbf{x}) \right) \end{aligned}$$

נחשב את b על ידי שימוש בנקודה הראשונה:

$$\begin{aligned}
1 &= y^{(1)} (\mathbf{w}^\top \Phi(\mathbf{x}^{(1)}) + b) \\
\Leftrightarrow b &= 1 - \mathbf{w}^\top \Phi(\mathbf{x}^{(1)}) \\
\Leftrightarrow b &= 1 - \frac{1}{1 - e^{-8}} (K(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) - K(\mathbf{x}^{(2)}, \mathbf{x}^{(1)})) \\
\Leftrightarrow b &= 1 - \frac{1}{1 - e^{-8}} (1 - e^{-8}) = 0
\end{aligned}$$

המסוג יהיה:

$$\begin{aligned}
h(\mathbf{x}) &= \text{sign}(\mathbf{w}^\top \Phi(\mathbf{x}) + b) \\
&= \text{sign} \left(\frac{1}{1 - e^{-8}} (K(\mathbf{x}^{(1)}, \mathbf{x}) - K(\mathbf{x}^{(2)}, \mathbf{x})) \right) \\
&= \text{sign} (K(\mathbf{x}^{(1)}, \mathbf{x}) - K(\mathbf{x}^{(2)}, \mathbf{x})) \\
&= \text{sign} \left(\exp(-\|\mathbf{x}^{(1)} - \mathbf{x}\|^2) - \exp(-\|\mathbf{x}^{(2)} - \mathbf{x}\|^2) \right) \\
&= \text{sign} (\|\mathbf{x}^{(2)} - \mathbf{x}\|^2 - \|\mathbf{x}^{(1)} - \mathbf{x}\|^2) \\
&= \text{sign} ((x_1 + 1)^2 + (x_2 + 1)^2 - (x_1 - 1)^2 - (x_2 - 1)^2) \\
&= \text{sign} (4x_1 + 4x_2) \\
&= \text{sign} (x_1 + x_2)
\end{aligned}$$

חלק מעשי: זיהוי מין הדובר

Code

בחלק זה, ננסה להשתמש ב-SVM כדי לזהות את מינו של הדובר באמצעות קולו. מוטיבציה למערכת כזאת יכולה להיות עוזר וירטואלי שרוצה לפנות לדובר לפי מינו. הרחבה לניסיון זה יכולה להיות זיהוי דובר על סמך קולו וכו'.

הרעיון וה-DATA נלקחו מ-Dataset והערכת ביצועים של קורי בקר, אשר נמצאים באתר הבא. בפרוייקט זה נאספו 3168 דגימות קול מתויוגות מהמקורות הבאים:

- [The Harvard-Haskins Database of Regularly-Timed Speech](#)
- [Telecommunications & Signal Processing Laboratory \(TSP\) Speech Database at McGill University](#)
- [VoxForge Speech Corpus](#)
- [Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University](#)

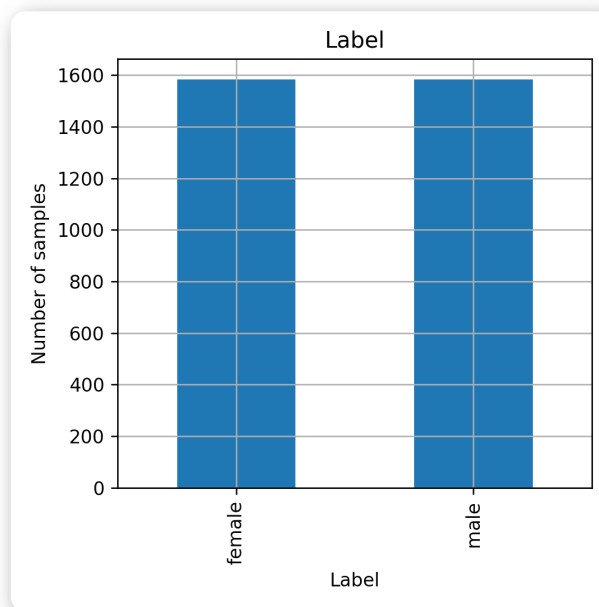
כל רצועת קול עברה עיבוד באמצעות כלי בשם [WarbleR](#) בכדי לייצר 20 Features לכל דגימה:

- **label:** The label of each track: male/female
- **meanfreq:** mean frequency (in kHz)
- **sd:** standard deviation of frequency
- **median:** median frequency (in kHz)
- **Q25:** first quantile (in kHz)
- **Q75:** third quantile (in kHz)
- **IQR:** interquantile range (in kHz)
- **skew:** skewness (see note in specprop description)
- **kurt:** kurtosis (see note in specprop description)
- **sp.ent:** spectral entropy
- **sfm:** spectral flatness
- **mode:** mode frequency
- **centroid:** frequency centroid (see specprop)
- **meanfun:** average of fundamental frequency measured across acoustic signal
- **minfun:** minimum fundamental frequency measured across acoustic signal
- **maxfun:** maximum fundamental frequency measured across acoustic signal
- **meandom:** average of dominant frequency measured across acoustic signal
- **mindom:** minimum of dominant frequency measured across acoustic signal
- **maxdom:** maximum of dominant frequency measured across acoustic signal
- **dfrange:** range of dominant frequency measured across acoustic signal
- **modindx:** modulation index. Calculated as the accumulated absolute difference between

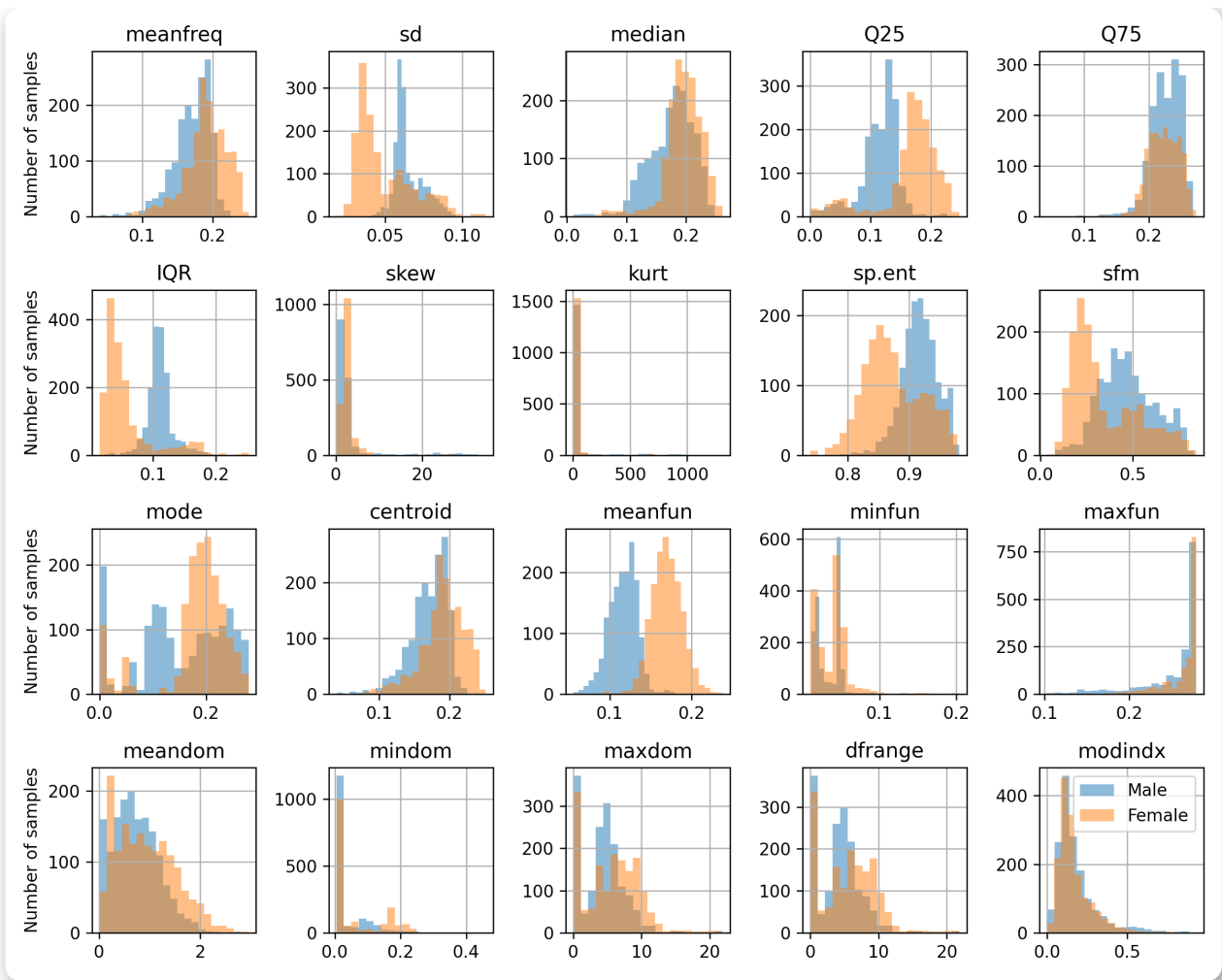
נציג מספר עמודות ושורות מן המדגם:

	label	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt
0	male	0.059781	0.0642413	0.0320269	0.0150715	0.0901934	0.075122	12.8635	274.403
1	male	0.0660087	0.06731	0.0402287	0.0194139	0.0926662	0.0732523	22.4233	634.614
2	male	0.0773155	0.0838294	0.0367185	0.00870106	0.131908	0.123207	30.7572	1024.93
3	male	0.151228	0.0721106	0.158011	0.0965817	0.207955	0.111374	1.23283	4.1773
4	male	0.13512	0.0791461	0.124656	0.0787202	0.206045	0.127325	1.10117	4.33371
5	male	0.132786	0.0795569	0.11909	0.067958	0.209592	0.141634	1.93256	8.3089
6	male	0.150762	0.0744632	0.160106	0.0928989	0.205718	0.112819	1.53064	5.9875
7	male	0.160514	0.0767669	0.144337	0.110532	0.231962	0.12143	1.39716	4.76661
8	male	0.142239	0.0780185	0.138587	0.0882063	0.208587	0.120381	1.09975	4.07028
9	male	0.134329	0.08035	0.121451	0.07558	0.201957	0.126377	1.19037	4.78731

הפילוג של התוויות (גברים-נשים) במדגם הינו:



נציג את ההיסטוגרמה של כל אחד מהמאפיינים בעבור כל אחד מתוויות:



Soft SVM

ננסה לפתור את בעיית הסיווג של מין הדובר בעזרת soft SVM. נרצה לשם כך לפתור את בעיית האופטימיזציה הבאה (הבעיה הדואלית):

$$\begin{aligned} \{\alpha_i\}^* &= \arg \max_{\{\alpha_i\}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y^{(i)} y^{(j)} \alpha_i \alpha_j \mathbf{x}^{(i)\top} \mathbf{x}^{(j)} \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i \\ & \sum_i \alpha_i y^{(i)} = 0 \end{aligned}$$

את \mathbf{w} נמצא בעזרת:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

את b ניתן לחשב על ידי בחירה של נקודה שעבורה $0 < \alpha_i$ ולהשתמש במשוואה: $y_i (\mathbf{w}^T \mathbf{x}_i + b) = 1 - \xi_i$

כדי לפתור את הבעיה נצטרך לבחור פרמטר משקל C אשר קובע את גודל הקנס לנקודות שחורגות מה margin. נתחיל ב $C = 1$ ואחר כך ננסה למצוא את הערך המיטבי.

חלוקת המדגם

כרגיל נחלק את המדגם ל 20%-20%-60% שהם train-validation-test.

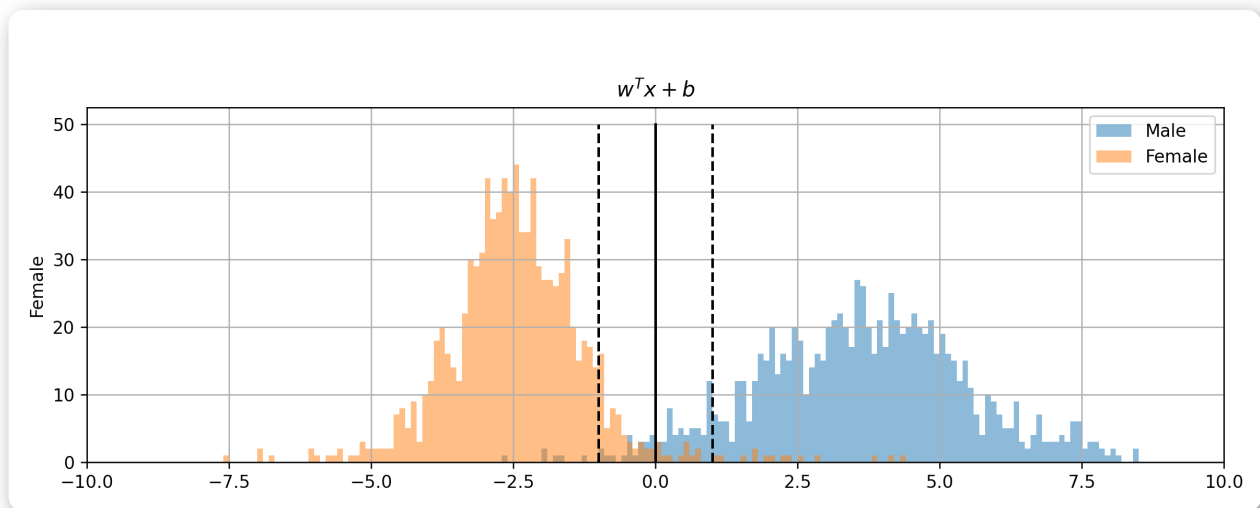
נרמול

חשוב לנרמל את העמודות של המדגם לפני הרצת האלגוריתם משתי סיבות עיקריות:

1. המדגם מכיל מאפיינים ביחידות וסקלות שונות.
2. האלגוריתם מנסה למזער objective אשר מבוסס מרחק, מה שהופך אותו לרגיש לגודל של כל מאפיין. לדוגמא, אם נכפיל מאפיין מסויים בקבוע גדול מ-1 אנו ניתן לו חשיבות יתרה ב objective.

תוצאות

נשתמש באופטימיזציה נומרית על מנת למצוא את הפרמטרים של משטח ההפרדה. לשם המחשה נשרטט את הפילוג של ה signed distance של הנקודות בכל אחת מהמחלקות

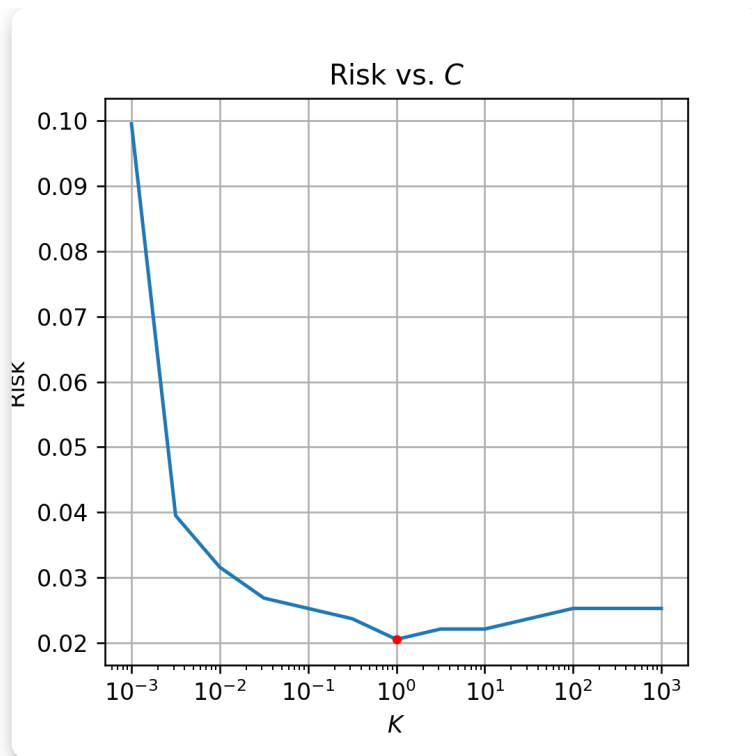


ניתן לראות כי המשטח באמת מצליח לחלק את המדגם עד כדי כמה נקודות שחורגות.

ה miscalssification rate שמתקבל על ה test set הינו 0.028.

מציאת ה C האופטימאלי

נבחן סדרה של ערכי C בתחום 10^{-3} עד 10^3 . כרגיל, בעבור כל ערך נבנה מסווג ונבחן אותו על ה validation set. נקבל את התוצאה הבאה:



קיבלנו כי הערך שאיתו התחלנו של $C = 1$ הוא האופטימאלי מבין הערכים שבחרנו.