## תרגול #4 ג' – 28/4/20

נלמד על אלגוריתם סיווג בינארי: Support Vector Machines (SVM)

### אלגוריתם SVM:

הבעיה: בין מודלים לינאריים (decision bound לינארי), נמצא את המודל הטוב ביותר. כלומר נבחר את המישור שעבורו ה-margin מקסימליים: המרחק ממנו לנק' ה-data הקרובות ביותר. נקודות אלו נקראות support vectors

**שיעור היום:**
(1) SVM: (i) Hard SVM – פרמול מתמטי
(ii) נשתמש ב-QP לפתרון SVM
(iii) Soft SVM
(2) Kernels: (i) מיפוי פיצ'רים למימד גבוה.
(ii) טריק ה-kernel

## SVM (1)

**כלי:**

**מישור מפריד:** מוגדר ע"י וקטור $w\in\mathbb{R}^d$ ו-$b\in\mathbb{R}$, הוא הקבוצה: $HP_{w,b} := \{u \mid \langle w,u\rangle + b = 0\}$. עבור הנתונים $\{x_i, y_i\}_{i=1}^m$

כאשר: $x_i\in\mathbb{R}^d$, $y_i\in\{-1,1\}$, נאמר שמישור-על מפריד. הוא מישור שמקיים לכל $i\in[m]$: $y_i(\langle x_i,w\rangle+b) > 0$.

(*) נסביר: $\langle x_i, w\rangle = 0$ אם $x_i$ על המישור $w^\perp$, הוא $>0$ אם $x_i, w$ נמצאים באותו צד של המישור, והוא $<0$ אם $x_i, w$ נמצאים בצדדים הנגדיים של המישור.

נרצה לדעת האם תייגנו נכון.

נבחין שאם $y_i=-1$, אז אם תייגנו נכון $y_i(\langle x_i,w\rangle) > 0$
ואם $y_i=1$, אז אם תייגנו נכון $y_i(\langle x_i,w\rangle) > 0$.

⇐ לכן זה אכן מישור מפריד אם (ורק אם) שומר על התכונה: $y_i(\langle x_i,w\rangle) > 0$, לכל דגימה ב-$S$.

**הערה:** עבור $c>0$ ו- $w\in\mathbb{R}^d$ מתקיים: $HP_{w,b} = HP_{cw,cb}$

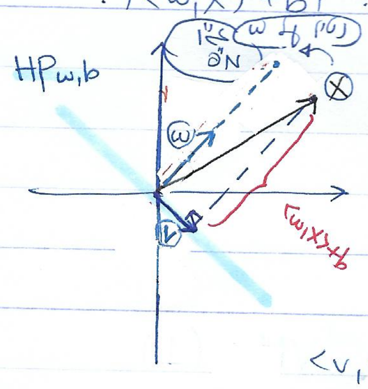
**הוכחה:** $HP_{w,b} = \{u \mid \langle w,u\rangle + b = 0\} = \{u \mid c(\langle w,u\rangle + b) = 0\}$
$= \{u \mid \langle cw,u\rangle + cb = 0\} = HP_{cw,cb}$

(*) ולכן אפשר להגביל עצמנו ל- $\|w\|=1$ (נוח חישובית)

**$d(x,B)$:** $B\subseteq\mathbb{R}^d$, $x\in\mathbb{R}^d$. אזי: $d(x,B) = \inf_{v\in B} \|x-v\|^2$

(*) בהינתן מישור, ה- $v$ שיקיים $\langle v,x\rangle=0$ עבור $v$ על המישור.
הוא הקרוב ביותר ל- $x\in\mathbb{R}^d$.

**טענה:** יהיו $x\in\mathbb{R}^d$, $HP_{w,b}$, עבור $\|w\|=1$ הוכח: $|\langle w,x\rangle+b|$.




**הוכחה:** נסמן: $v = x - \underbrace{(\langle w,x\rangle+b)}_{d} w$. ~~מתקיים~~:

(1) $v\in HP_{w,b}$

(2) $v$ הוא הנק' הקרובה ביותר ב- $HP_{w,b}$ ל- $x$

(3) $d(v,x) = |\langle w,x\rangle + b|$

נוכיח:

(1) $\langle v,w\rangle + b = \langle x-dw,w\rangle + b = \langle x,w\rangle - d\underbrace{\langle w,w\rangle}_{1} + b$
$= \langle x,w\rangle - d + b = \langle x,w\rangle + b - (\langle x,w\rangle + b) = 0$

ולכן $v\in HP_{w,b}$

(2) יהי $u\in HP_{w,b}$. נראה ש- $v$ הקרוב ביותר. מתקיים:

$\|x-u\|^2 = \|(x-v)+(v-u)\|^2 = \langle (x-v)+(v-u), (x-v)+(v-u)\rangle =$
$= \langle x-v, x-v\rangle + 2\langle x-v, v-u\rangle + \langle v-u, v-u\rangle = \|x-v\|^2 +$
$+ 2\langle x-v, v-u\rangle + \underbrace{\|v-u\|^2}_{\geq 0} \geq \|x-v\|^2 + 2\langle x-v, v-u\rangle$

מתקיים בנוסף:

$\langle x-v, v-u\rangle = \langle dw, v-u\rangle = d(\underbrace{\langle w,v\rangle}_{-b} - \underbrace{\langle w,u\rangle}_{-b}) = 0$

$u,v\in HP_{w,b}$ ולכן:

וקיבלנו: לכל $u$: $\|x-u\|^2 \geq \|x-v\|^2$, כלומר $v$ אכן בעל המרחק המינימלי ל- $x$.

(3) $\|x-v\| = \|dw\| = |d|\underbrace{\|w\|}_{1} = |\langle w,x\rangle + b|$

$\blacksquare$

**margin**: יהא $S=\{x_i,y_i\}_{i=1}^m$. מרחק הנק' הקרובה ביותר מ-S

ל-$HP_{w,b}$ נקרא ה-margin, והוא שווה ל: $\min_{i\in[m]} |\langle w,x_i\rangle+b|$

## (i) Hard-SVM

**Hard SVM Rule** (1)

רוצים למצוא: $\boxed{\underset{\|w\|=1,\, b}{\operatorname{argmax}} \left(\min_{i\in[m]} |\langle x_i,w\rangle+b|\right)}$

כך שלכל i: $y_i(\langle x_i,w\rangle+b)>0$

(מעוניינים על מישור מפריד)

⇐ נראה כמו בעיית אופטימיזציה!

**בעיית אופטימיזציה**: פתרון

$$\boxed{\begin{aligned}&\max_{w\in\mathbb{R}^n} (f(w))\\ &s.t:\ g(w)\le d\end{aligned}}$$

עבור: $f:\mathbb{R}^n\to\mathbb{R}^n$, $g:\mathbb{R}^n\to\mathbb{R}^n$, $d\in\mathbb{R}^m$ קבועים.

f פונק' המטרה, g האילוצים.

(QP) **בעיית תכנון ריבועי**: דוגמה של אופטימיזציה:

$$\boxed{\begin{aligned}&\min_{w\in\mathbb{R}^n}\left(\frac{1}{2}w^TQw+a^Tw\right)\\ &s.t\ \ Aw\le d\end{aligned}}$$

עבור: $Q\in\mathbb{R}^{n\times n}$, $A\in\mathbb{R}^{m\times n}$, $a\in\mathbb{R}^n$, $d\in\mathbb{R}^m$ קבועים.

(*) כל מינימיזציה ניתן להעביר למקסימיזציה ע"י הכפלה ב-1-. (אם פונ' המטרה מקבלת רק ערכים חיוביים).

## (ii) פתרון SVM עם QP.

**טענה**: יהיו $v^*, c^*$ פתרונות אופטימליים של (2)

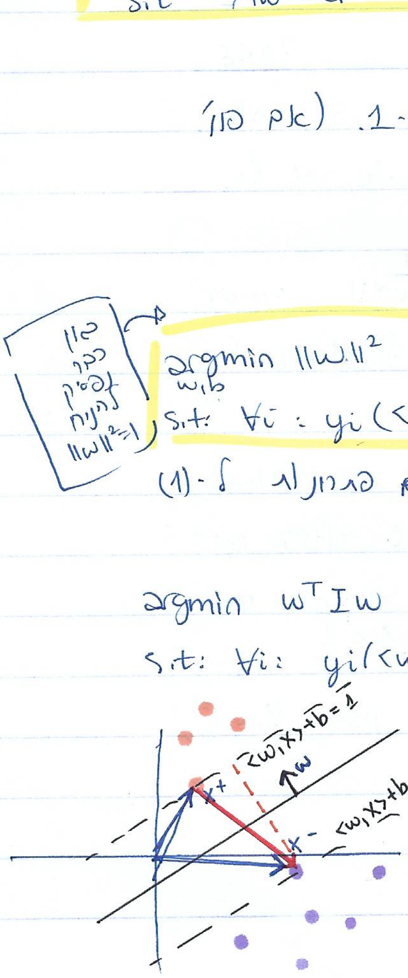
$$\boxed{\begin{aligned}&\underset{w,b}{\operatorname{argmin}}\ \|w\|^2\\ &s.t:\ \forall i:\ y_i(\langle x_i,w\rangle+b)\ge 1\end{aligned}}$$

כיון כבר נפסיק להניח $\|w\|^2=1$

נסמן: $\gamma=\frac{1}{\|v^*\|}$. אזי: $w^*=\gamma v,\ b^*=\gamma c^*$ הם פתרונות ל-(1).

**הוכחה**: נראה ש-(2) בעיית QP; ואכן:

$$\begin{aligned}&\operatorname{argmin}\ w^TIw\\ &s.t:\ \forall i:\ y_i(\langle w,x_i\rangle+b)\ge 1\end{aligned}$$




אינטואיציה:

כעת נראה ש (2) שקול ל-(1):

$$\underset{\text{margin}}{\ell} = (x^+-x^-)\frac{\vec{w}}{\|w\|} = \left(\langle x^+,w\rangle - \langle x^-,w\rangle\right)\frac{1}{\|w\|} =$$

$$= \left((1-b)+(-1-b)\right)\frac{1}{\|w\|} = \frac{2}{\|w\|}$$

⇐ אם נמצא w שנורמה קטנה ⇐ נמקסם את ה-margin!

כי $(\operatorname{argmin}\|w\|^2) \Leftrightarrow (\operatorname{argmax}\{\min[|\langle x_i,w\rangle+b|]\})$

נתחיל מ- **Hard SVM Rule**:

$$(1)\quad \underset{\|w\|=1,\, b}{\operatorname{argmax}} \left( \min_i |\langle w, x_i\rangle + b| \right) \quad s.t:\ \forall i-\ y_i(\langle w, x_i\rangle + b) > 0$$

נבחין: $|\langle w, x_i\rangle + b| = y_i(\underbrace{\langle w, x_i\rangle + b}_{d}) \Leftrightarrow y_i$ ו-$d$ שווי סימן,
ולכן (1) שקולה ל-(2):

$$(2)\quad \underset{\|w\|=1,\, b}{\operatorname{argmax}} \left( \min_i (\langle w, x_i\rangle + b) \right) \quad s.t:\ \forall i-\ y_i(\langle w, x_i\rangle + b) > 0$$

נבחין שהביטוי בפונק' המטרה דומה לאילוצים.
מה קורה אם אחד מהם לא מתקיים? הוא לא פתרון פיזיבילי
המינימום יבחר את הפתרון הלא פיזיבילי הזה, שנותן ערך שלילי
ואז argmax על כל הפתרונות האפשריים על כל המישורים האפשריים
לא יבחר את הפתרון הלא-פיזיבילי הזה.
לכן אפשר לשקול פתרונות לא פיזיבליים (הם לא ילקחו בגלל argmax)
לכן יש פתרונות מיותרים, ולכן ההצגה הזו שקולה גם היא:

$$(3)\quad \underset{\|w\|=1,\, b}{\operatorname{argmax}} \{ \min_i y_i(\langle w, x_i\rangle + b) \}$$

נבחין שעבור פתרון אופטימלי $c^*$, $v^*$ תמיד מתקיים:

$$\min_i \{ y_i(\langle v^*, w_i\rangle + c^*) \} = 1 \quad (\star)$$

__למה?__

אחרת היינו יכולים לחלק את $v^*$ בקבוע $>0$, ולקבל פתרון פיזיבילי שממקסם אף יותר את פונ' המטרה (QP), בסתירה לאופטימליות $v^*$.

נסמן: $\gamma = \frac{1}{\|v^*\|}$. מתקיים: $w^* = \gamma v^*$, $b^* = \gamma c^*$ פתרון פיזיבילי
למשוואה (3) עם פונק' המטרה $\min_i \{ y_i(\langle \gamma v^*, x_i\rangle + \gamma c^*) \} = \gamma$ (מ-$\star$).

נראה שזה פתרון אופטימלי ל-(3):
נניח בשלילה שקיים פתרון טוב יותר: $\min_i \{ y_i(\langle w_2, x_i\rangle + b_2) \} = \delta > \gamma$.
אז מתקיים: $\frac{b_2}{\delta}$, $\frac{w_2}{\delta}$ פתרון פיזיבילי למשוואה (QP)
עם $\left\|\frac{w_2}{\delta}\right\| = \frac{1}{\delta} < \frac{1}{\gamma}$, כלומר פתרון טוב יותר מ-$v^*, c^*$ בסתירה לאופטימליות של פתרון QP זה.

■

הערה: נוח לעבור למרחבי למידה הומוגניים ע"י הוספת פיצ'ר (bias) לכל דוגמה $x_i$
כלומר ע"י הוספת מימד ל-X.
אך הגדרנו את הבעיה כבעיה שיש אילוץ על נורמה, כלומר הוספת
ה-bias משפיע את האילוצים, מוסיף אילוץ שלא היה.
יש טענה שאומרת שזה לא באמת משנה (לא משפיע על הפתרון
והוא עדיין קרוב לאופטימלי, והרבה יותר קל).

## Soft SVM (iii)

יש חפיפה בין הקבוצות — איך אפשר לבחור w-מישור מפריד אם כך?
נגדיר קב' משתנים $\geq 0$: $\{\xi_i\}$ שמציינים את כמות החריגה של
הדוגמה ה-i מה-margin.

ממקסם את ה-margin    לאפשר חריגות

נגדיר את הבעיה:

$$\min_{w,\{\xi_i\}} \left( \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \right) \quad (1)$$
$$\text{s.t} \quad \forall i: \; y_i \langle w, x_i \rangle \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

(לכל נאפשר רק חריגה מסוימת מהשוליים)

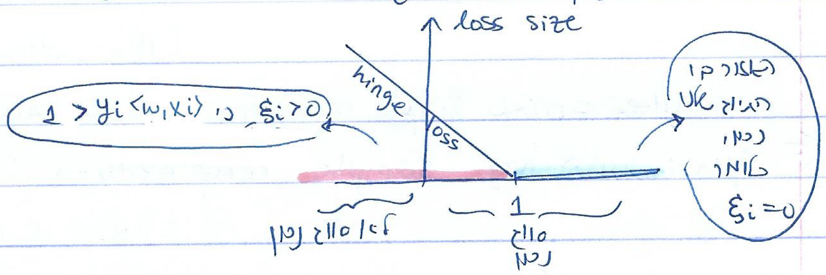
- אם λ גבוהה: $\|w\|^2$ החלק הזה במטרת מינימום. לא מעניין כמות חריגות, רק מעניין ש-$\|w\|$ קטנה (ולכן margin גדול).
- אם λ קטן: לא מעניין $\|w\|^2$, ה-margin, עדיף להקטין את כמות החריגות, על חשבון הקטנת ה-margin.

כתיבה שקולה:

$$\min_{w,b} \; \lambda\|w\|^2 + \frac{1}{m}\sum \ell^{hinge}(y_i\langle w, x_i\rangle) \quad (2)$$

כאשר: $\ell^{hinge}(a) = \max\{0, 1-a\}$

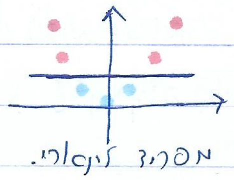
(מחליפה $\xi_i$ והאילוצים)    פונ' רציפה ואי-גזירה



# Kernels

אם נתבונן על קב' נקודות $x\in\mathbb{R}$ :  בלתי ניתנות להפרדה לינארית, ניתן להתבונן על אותה הקב' ע"י פונק' שממפה את הנקודות שלנו למימד גבוה יותר. לדוג':

$x \to (x, x^2)$



וכעת ניתן למצוא בקלות מפריד לינארי.
משם נוכל להטיל על הקב' המקורית, ולקבל שם מפריד לינארי.

## (i) High dimensional Feature Mapping

להעביר את הדאטה למימד גבוה ושם להפריד

נבחר מיפוי $\psi : \mathcal{X} \to \mathcal{F}$ שממפה את מרחב הקלט למרחב פיצ'רים.
$\mathcal{F}$ מוגדר להיות מרחב מ"פ שלם: מרחב הילברט.
לרוב: $\mathcal{X}\subseteq\mathbb{R}^d$ , $\mathcal{F}\subseteq\mathbb{R}^k$ ו- $k>d$ .

כעת, ישנו את המידע: $\{(\psi(x_i), y_i)\}_{i=1}^{m}$ , ונרצה למצוא עבורו את בעית soft-SVM, תחת פונק' המיפוי $\psi$, כלומר:

$$\boxed{\min_{w,b}\left\{\lambda \|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}(y_i\langle w, \psi(x_i)\rangle)\right\}}$$

## הקשיים בשימוש ב-Kernels

(1) בחירת $\psi$
(2) עלות חישובית גבוהה יותר ככל שהמימד עולה.
(3) מס' הדגימות שאנחנו צריכים כדי לאמן את המודל עולה ככל שהמימד עולה (sample complexity)

איך נתמודד עם בעיות אלו?

(3) יש חסמים שאומרים שכאשר ה-margin גדול במרחב שאנו ממפים אליו, מס' ה-sample complexity עולה בצורה מתונה.

פרוט בהרצאות הבאות.

(2) כדי לפתור בעיה זו נשתמש ב:
(ii) טריק הקרנל:

להפריד במימד המקורי

Kernels - מתאר מכפלות פנימיות במרחב הפיצ'רים.

פונק' הקרנל: $K(x,x') = \langle \psi(x), \psi(x') \rangle$

כלומר המ"פ של המיצג הממופה במרחב אליו מיפינו.

כאשר $\psi(x)$ קשה לחישוב, אך: $K(x_i,x_j) = \langle \psi(x_i), \psi(x_j) \rangle$
קל לחישוב עבור 2 דגימות $x_i, x_j$, נחשב למעשה את $K(x_i,x_j)$.
(*) מסתבר שזו תכונה שמתקיימת לרוב.

דוג': $X=\mathbb{R}$, $\psi(x)_n = n!^{-\frac{1}{2}} e^{-\frac{x^2}{2}} x^n$, לכל רכיב n.
נבחין ש- $\psi$ ממפה מיצג מ- $\mathbb{R}$ למרחב $\infty$-מימדי, ולכן לא
ניתן לחשב ישירות את $\psi$, אולם עבור $x,x' \in \mathbb{R}$, נבחין ש:
$\langle \psi(x), \psi(x') \rangle = e^{-\frac{(x-x')^2}{2}}$, פונק' פשוטה חישובית בהרבה.

משפט הייצוג: אם כלל הלמידה הוא מהצורה הבאה:

$$(\star) \quad w^* \in \arg\min_w f(\langle w, \psi(x_1)\rangle, \dots, \langle w, \psi(x_m)\rangle) + \lambda \|w\|^2$$

עבור: $\psi : X \to F$, כאשר $F$ מרחב מ"פ שלם (מרחב הילברט),
$f: \mathbb{R}^n \to \mathbb{R}$ פונק' כלשהי ו- $\lambda \in \mathbb{R}_+$.
אזי: קיים פתרון $w^*$ ווקטור $\alpha \in \mathbb{R}^m$ כך שמתקיים -

$$w^* = \sum_{i=1}^m \alpha_i \psi(x_i)$$

נרצה לראות איך המשוואה $(\star)$ מתקשרת ל- Hard-SVM rule:

$$\arg\min_{(w,b)} \|w\|^2 \quad s.t: \ \forall i \ \ y_i(\langle x_i, w\rangle + b) \geq 1$$

מכיוון שב-2 המקרים מופיע $\|w\|^2$, אזי $\lambda = 1$. מהי f?

$$f(a_1,..,a_m) = \begin{cases} 0 & : if \ \exists b \ s.t \ \forall i \ \ y_i(a_i+b) \geq 1 \\ \infty & : \text{אחרת} \end{cases}$$

$\boxed{\text{כשהתנאים מתקיימים נוסיף ל-}\lambda\|w\|^2 \text{ 0, ונקבל את H-SVM. אחרת, נוסיף } \infty \text{ ולא נמצא ש מישור מפריד}}$

נרצה לראות איך $(\star)$ מתקשרת ל- Soft-SVM rule:

נגדיר את $f(a_1,..,a_m) = \frac{1}{m}\sum_{i=1}^m \ell^{hinge}(y_i a_i)$, ונקבל:

$$\min_{w,b} \lambda \|w\|^2 + \frac{1}{m}\sum_{i=1}^m \ell^{hinge}(y_i \langle w, x_i\rangle)$$

או

$$\min_{w,b} \lambda \|w\|^2 + \frac{1}{m}\sum_{i=1}^m \ell^{hinge}(y_i \langle w, \psi(x_i)\rangle)$$

כלומר אפשר לתאר בצורה דומה למשפט הייצוג את כלי Hard-SVM ו- Soft-SVM.

<u>מה נעשה עם זה?</u>

נגדיר $K(x_i,x)=\langle \Psi(x_i),\Psi(x)\rangle$, ונשתמש במשפט הייצוג לקבלת $\alpha_1,\dots,\alpha_m$ שמספקים לנו $w^*$ פתרון אופטימלי.
$(w^*=\sum_{i=1}^{m}\alpha_i\Psi(x_i))$.

אז מה הסיווג של דוגמה x? אמור להיות $\Psi(x)$, אך זה קשה חישובית.
לכן נחשב: $\langle w^*,\Psi(x)\rangle=\sum_{i=1}^{m}\alpha_i K(x_i,x)$
ונבדוק אם התוצאה מעל או מתחת למישור, ועל פיה נתייג את x!
(ללא חישוב מפורש של $\Psi(x)$...)

(1) בחירת $\Psi$:
2 תשובות:
1. יש הרבה קב׳ כלליות של פונק׳ מיפוי שלפעמים יכולות לעזור, ולתת פתרון טוב (גם אם לא האופטימלי)
2. לפעמים ניתן להשתמש בידע על הבעיה כדי לבחור אותו בצורה חכמה: לשמור על קשר בין פיצ׳רים עם מימד הקרבה, למשל.