מבוא ללמידה חישובית – תרגיל בית 4

שם : דורי רימון

ת"ז : 323996843

<u>שאלות תיאורטיות:</u>

שאלה 1:

א)

שאלה 1

א) נראה, שניתן לחשב באופן מפורש את $\Pi_K(x)$ במקרה זה.

נזכר כי $K = \{x \mid \|x\| \le R\}$.

$$\Rightarrow \Pi_K(x) = \begin{cases} x & \|x\| \le R \\ R \cdot \frac{x}{\|x\|} & \text{אחרת} \end{cases}$$

ההסבר חשוב כי מדובר בכדור סביב הראשית ברדיוס R, לכן הנקודה הקרובה ביותר ב-K לנקודה מבחוץ היא פשוט על הישר המחבר בין הנקודה והמרכז, והכיוון נשמר, ולכן את הוקטור נחלק בנורמה כדי לקבל נורמה 1 ($\frac{x}{\|x\|}$) ובכפל ב-R נקבל את הנקודה המבוקשת.

אם כן, אלגוריתם SGD יפעל כדלקמן:

* נתחיל $w_1 \in K$, למשל, $w_1 = 0$.

* כעת, לכל t, נבחר $i \in \{1, \dots, m\}$ באופן אקראי.

$$w_{t+1} = \begin{cases} w_t - \eta_t \nabla f_i(w_t) & \text{כאשר } \|w_t - \eta_t \nabla f_i(w_t)\| \le R \\ R \cdot \frac{w_t - \eta_t \nabla f_i(w_t)}{\|w_t - \eta_t \nabla f_i(w_t)\|} & \text{אחרת} \end{cases}$$

במקרה שלנו, f_i היא הפונקציה hinge loss -

$$f_i(w) = \max\{0, 1 - y_i w x_i\}$$

$$\Rightarrow \nabla f_i(w) = \begin{cases} 0 & y_i w x_i > 1 \\ -y_i w & \text{אחרת} \end{cases}$$

$$\Rightarrow w_t - \eta_t \nabla f_i(w_t) = \begin{cases} w_t & y_i w_t x_i > 1 \\ w_t + \eta_t \cdot y_i w_t & \text{אחרת} \end{cases}$$

סה״כ,

$$w_{t+1} = \begin{cases} w_t & y_i w_t x_i > 1 \\ \begin{cases} w_t(1 + \eta_t y_i) & \| w_t(1 + \eta_t y_i) \| \le R \\ R \cdot \dfrac{w_t(1 + \eta_t y_i)}{\| w_t(1 + \eta_t y_i) \|} & \text{אחרת} \end{cases} \end{cases}$$

ב)

ק)

* טריוויאלית, ואם $y \in K$, אז ברור ש- $x = y$, ואז אי-השוויון מתקיים ע"י שוויון.

* נניח $y \notin K$. נניח בשלילה, $\exists z \in K, \|y-z\| < \|x-z\|$.

$x = \Pi_K(y) \Rightarrow \|y-x\| < \|y-z\|$

$\Rightarrow \|y-x\| < \|y-z\| < \|x-z\|$

נתבונן במשולש z, y, x. הצלע (z, x) היא הארוכה ביותר כי אי-השוויונים מעלה מספרים. אז מול הזווית הגדולה ביותר. כלומר זווית ש- y גדולה.

* θ נקודה על (z, x) (שמחוץ ל- K (קמורה))
* הזווית ש- y כמעט קהה (המצב נייצב שהזווית הגדולה ביותר).

אז נקבל מרחק בין הזווית ל- $(\hat{z}, x)$ בנקודה $\alpha \in K$. נשים לב כי $\|x-y\| \geq \|\alpha - y\|$ אבל המרחק הקצר ביותר נקודה הוא לקצר אותו.
אז כן, מתקיים כי $x = \Pi_K(y)$, $\alpha = x$, אז קיבלנו לסתירה כי $\alpha = x$.
אז כן זה סתירה מדובר בשוקי $\overset{\text{זווית}}{\text{מעוי}}$, שתי הזוויות הישרה בקודקוד x, סתירה לכך שהזווית הגדולה ביותר נמצאת בקודקוד y.

ג)

ג) אנו רוצים להוכיח כי בהינתן דיוק $\epsilon > 0$, פונק' f ק-ל שהגרדיאנט חסום ע"י G, ו- w^* אופטימלי ב- K כך ש- $\|w^*\| \le B$, אז בחירת צעד $\eta = \epsilon / G^2$ ו- $T = B^2 G^2 / \epsilon^2$ איטרציות, מתקיים

$$f(\bar{w}) - f(w^*) \le \epsilon \ , \quad \bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$$

$$f\left(\frac{1}{T} \sum_{t=1}^{T} w_t\right) \le \frac{1}{T} \sum_{t=1}^{T} f(w_t)$$

$$\Rightarrow f(\bar{w}) - f(w^*) \le \frac{1}{T} \sum_{t=1}^{T} \left(f(w_t) - f(w^*)\right)$$

$$\underset{f \text{ קמורה}}{\le} \frac{1}{T} \sum_{t=1}^{T} \nabla f(w_t)(w_t - w^*)$$

$$\frac{1}{T} \sum_{t=1}^{T} \nabla f(w_t)(w_t - w^*) = \frac{1}{T} \sum_{t=1}^{T} \frac{\|w_t - w^*\|^2 - \|w_t - w^* - \eta \nabla f(w_t)\|^2}{2\eta} - \frac{\eta}{2} \|\nabla f(w_t)\|^2$$

מהלמה מההרצאה -

$$w_{t+1} = \Pi_K (w_t - \eta \nabla f(w_t))$$

$$\underset{\forall t}{\Rightarrow} \|w_t - \eta \nabla f(w_t) - w^*\| \ge \|w_{t+1} - w^*\|$$

מש.ל קמורה,
$w^* \in K$

$$\Rightarrow \frac{1}{T} \sum_{t=1}^{T} \frac{\|w_t - w^*\|^2 - \|w_t - w^* - \eta \nabla f(w_t)\|^2}{2\eta} - \frac{\eta}{2} \|\nabla f(w_t)\|^2$$

$$\le \frac{1}{T} \sum_{t=1}^{T} \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta} - \frac{\eta}{2} \|\nabla f(w_t)\|^2$$

ומכאן ניתן להמשיך כמו בהוכחה מההרצאה - להתכנסות GD.

שאלה 2:

א)

שאלה 2

א) נתבונן בפונקציית ה- multi - class SVM עבור $K=2$:

$$f(w_1, w_2) = \frac{\beta}{2} \sum_{j \in \{1,2\}} \|w_j\|^2 + \frac{1}{n} \sum_{i=1}^{n} l(w_1, w_2, x_i, y_i) =$$

$$= \frac{\beta}{2} \left(\|w_1\|^2 + \|w_2\|^2\right) + \frac{1}{n} \sum_{i=1}^{n} \max_{j \in \{1,2\}} \{ w_j x_i - w_{y_i} x_i + \mathbb{1}(j \neq y_i) \}$$

נתבונן ב- l :

$$l(w_1, w_2, x_i, y_i) = \max \begin{cases} (w_1 - w_{y_i}) x_i + \mathbb{1}(1 \neq y_i) \\ (w_2 - w_{y_i}) x_i + \mathbb{1}(2 \neq y_i) \end{cases}$$

נתבונן בכל אחד מהביטויים בנפרד

$$(w_1 - w_{y_i}) x_i + \mathbb{1}(1 \neq y_i) = \begin{cases} 0 & y_i = 1 \\ (w_1 - w_2) x_i + 1 & y_i = 2 \end{cases}$$

$$= \begin{cases} 0 & y_i = 1 \\ (2y_i - 3)(w_1 - w_2) x_i + 1 & y_i = 2 \end{cases}$$

$$(w_2 - w_{y_i}) x_i + \mathbb{1}(2 \neq y_i) = \begin{cases} (w_2 - w_1) + 1 & y_i = 1 \\ 0 & y_i = 2 \end{cases} =$$

$$= \begin{cases} (2y_i - 3)(w_1 - w_2) x_i + 1 & y_i = 1 \\ 0 & y_i = 2 \end{cases}$$

סוגר, תחת הסימון $z = w_1 - w_2$,

$$l(w_1, w_2, x_i, y_i) = \max \{ 0, (2y_i - 3) \cdot (z \cdot x_i) + 1 \}$$

וניתן לראות כי פונק' ההפסד תלויה ב- z בלבד, ולא ב- w_2, w_1 בנפרד.

נסמן $c = w_1 + w_2$ ונכתוב את המטרה כפונק' $f(c, z)$:

$$f(c, z) = \frac{\beta}{8}\|c - z\|^2 + \frac{\beta}{8}\|c + z\|^2 + \frac{1}{n}\sum_{i=1}^{n} l(z, x_i, y_i)$$

כפי שציינו l

היא פונק' של

z, x_i, y_i

נשים לב כי

$$\|c - z\|^2 + \|c + z\|^2 = \|c\|^2 - \cancel{2cz} + \|z\|^2 + \|c\|^2 + \cancel{2cz} + \|z\|^2 =$$
$$= 2(\|c\|^2 + \|z\|^2)$$

מההנחה כי c לא משפיע על הסכום $\frac{1}{n}\sum_{i=1}^{n} l(z, x_i, y_i)$,

מכאן כי הפתרון האופטימלי יהיה רק אם $c = 0$.

כלומר, $w_1^* + w_2^* = c = 0 \Leftarrow w_1^*(\beta) = -w_2^*(\beta)$

כנדרש.

אם נחליף: $y_i \to 3 - y_i$, נקבל

$$\text{optimum}(f) = \frac{\beta}{4}\|z^*\|^2 + \frac{1}{n}\sum_{i=1}^{n} l_{hinge}(z^*, x_i, y_i)$$

וקיבלנו את הקשר $\beta' = \beta / 2$, $w^*(\beta') = z^* =$
$$= w_1^*(\beta) - w_2^*(\beta)$$

ב)

ק) ואז מניחים כי $\beta = 0$. אם כן,

$$f(w_1, \dots, w_k) = \frac{1}{n}\sum_{i=1}^{n} \ell(w_1, \dots, w_k, x_i, y_i) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} \max_{j \in \{1, \dots, k\}} \{(w_j - w_{y_i})x_i + \mathbb{1}[j \neq y_i]\}$$

מתקבל שה-f סכום אי-שליליות - סכום של ערכים אי-שליליים (מקס' של ערכים אחד מהם הוא 0).

נשים לב כי אם $w_1^*, \dots, w_k^*$ ממזערים את f, אז הם מפרידים את הנתונים בצורה מושלמת.

אם מתקיים כך היה, אז היה קיים $[k] \ni j \neq y_i$ עבורו

$$w_j \cdot x_i \geq w_{y_i} \cdot x_i \Rightarrow (w_j - w_{y_i}) \cdot x_i + 1 > 0$$

וזו סתירה לכך ש-f ממוזער.

מכיוון שלא ניתן להגיע תמיד לקיום קבוצה המפרידה את f. אולם, כיוון ש-f אי-שלילית, אז ערך ה-0 הוא המינימום האפשרי, ואם נשיג אותו, הוא יתקבל רק אם יש הפרדה מושלמת.

יהיו $w_1^*, \dots, w_k^*$ המפרידים הנתונים.

$$\Rightarrow y_i = \arg\max_{y} w_y^* \cdot x_i \Rightarrow \forall j \in [k],\ w_j^* \cdot x_i \leq w_{y_i}^* \cdot x_i$$

$$\Rightarrow \forall j \in [k],\ (w_j^* - w_{y_i}^*) \cdot x_i \leq 0$$

נסמן את ה-margin של המפרידים באופן הבא:

$$\lambda_i = \min_{j \in [k]} \{(w_{y_i}^* - w_j^*) \cdot x_i\}, \quad \lambda = \min_{i \in [n]} \{\lambda_i\}$$

וניקח כעת, $w_i = \frac{1}{\lambda} w_i^*$

$$\Rightarrow \forall j \in [k] \setminus \{y_i\},\ (w_j - w_{y_i}) \cdot x_i = \frac{1}{\lambda}(w_j^* - w_{y_i}^*) \cdot x_i =$$

$$= -\frac{1}{\lambda}(w_{y_i}^* - w_j^*) \cdot x_i \le -\frac{\lambda}{\lambda} \le -1$$

ונקבל שהוקטור $w_1, \dots, w_k$ מאפס את f, כפי שרצינו.

שאלה 3:

שאלה 3

א) נכון כי בבעיה עם התנאי האופטימלי של אילוץ האי-שוויון, עבור ק-ם i צריך $\varepsilon_i < 0$.

נכון כי פונק' המטרה מקטינה את סכום ε_i ולכן סותר את אופטימליות הפתרון הנוכחי.

* ההבדלה ל $\varepsilon_i \in \mathbb{R}$ בפונק' המטרה היא רק ריבועי $\frac{C}{2}\sum_{i=1}^{m} \varepsilon_i^2$. כיוון ש- ε_i היה שלילי, הפיכתו ל-0 מקטינה עוד את הריבוע שלו.

* נכון כי בנוסף, האילוצים נשמרים. האילוץ היחידי הרלוונטי הינו $y_i(w^T x_i + b) \geq 1 - \varepsilon_i$, ומכיוון ש- $\varepsilon_i < 0$, אז $1 - \varepsilon_i > 1$ $\Rightarrow y_i(w^T x_i + b) > 1$ ולכן גם אם נהפוך את $\varepsilon_i \to 0$ האילוץ עדיין נשמר.

ב)
$$L(w, b, \varepsilon, \lambda) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\|\varepsilon\|^2 + \sum_{i=1}^{m} \lambda_i (1 - \varepsilon_i - y_i(w^T x_i + b)) =$$

$$= \frac{1}{2}\|w\|^2 - \lambda\varepsilon + \frac{C}{2}\|\varepsilon\|^2 + \sum_{i=1}^{m} \lambda_i (1 - y_i(w \cdot x_i + b))$$

ג) נגזור בהתאם ונשווה ל-0:

$$\nabla_w L = w - \sum_{i=1}^{m} \lambda_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \lambda_i y_i x_i$$

$$\nabla_b L = -\sum_{i=1}^{m} \lambda_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \lambda_i y_i = 0$$

$$\nabla_\varepsilon L = C \cdot \varepsilon - \lambda = 0 \Rightarrow \varepsilon = \frac{1}{C} \cdot \lambda$$

ד) היה ניתן לפתור את המשוואות, וכך מציבים אותן ניתן לקבל את הלגרנז'יאן

הטיפוסי בסעיף ל ה.

ב)

שאלה 4:

שאלה 4

K הוא אכן קרנל. נראה זאת בעזרת $\phi(x):[K]\to\{0,1\}^K$ הבאה:

$$\phi(x)=(\underbrace{1,\dots,1}_{x\text{ פעמים}},\underbrace{0,\dots,0}_{K-x\text{ פעמים}})^t$$

$$\Rightarrow \phi(x)\cdot\phi(x')=\min\{x,x'\}=K(x,x')$$

כרצוי.

שאלה 5:

שאלה 5

$$(1+x_i x_j)^q \underset{\text{נוסחת הבינום}}{=} \sum_{k=0}^{q} \binom{q}{k} (x_i x_j)^k = \sum_{k=0}^{q} \left(\binom{q}{k}^{1/2} x_i^k\right)\left(\binom{q}{k}^{\frac{1}{2}} x_j^k\right)$$

על מנת שמיפוי זה יהיה תוצאת המכפלה הפנימית — נדרש שיתקיים

$$\phi(x) = (x^0, \binom{q}{1}^{\frac{1}{2}} x^1, \dots, \binom{q}{q}^{\frac{1}{2}} x^q)^T$$

כתיבה מטריצית -

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^q \\ 1 & x_2 & & \\ \vdots & & & \\ 1 & x_n & \dots & x_n^q \end{pmatrix}$$

מדובר במטריצת ונדרמונדה, עם כן מהצגה הרגילה, בגודל n.

$$\phi = \begin{pmatrix} \binom{q}{0}^{\frac{1}{2}} \cdot 1 & \binom{q}{1}^{\frac{1}{2}} \cdot x_1 & \dots & \binom{q}{q}^{\frac{1}{2}} \cdot x_1^q \\ \vdots & \vdots & & \vdots \\ \binom{q}{0}^{\frac{1}{2}} \cdot 1 & \binom{q}{1}^{\frac{1}{2}} \cdot x_n & & \binom{q}{q}^{\frac{1}{2}} \cdot x_n^q \end{pmatrix}$$

וכיוון שמטריצת ונדרמונדה, של מטריצה זה הפיכה בקבוע שלהו, מכיוון שדרגתה מלאה n.

ממילא מתקיים, נקבל שיש w המפריד את הנקודות ϕ ספרתי.

בהמשך כמו כאשר בהנחה מפרידה. אז אחי' Hard SVM פתור את שגיאת אימון אפס.

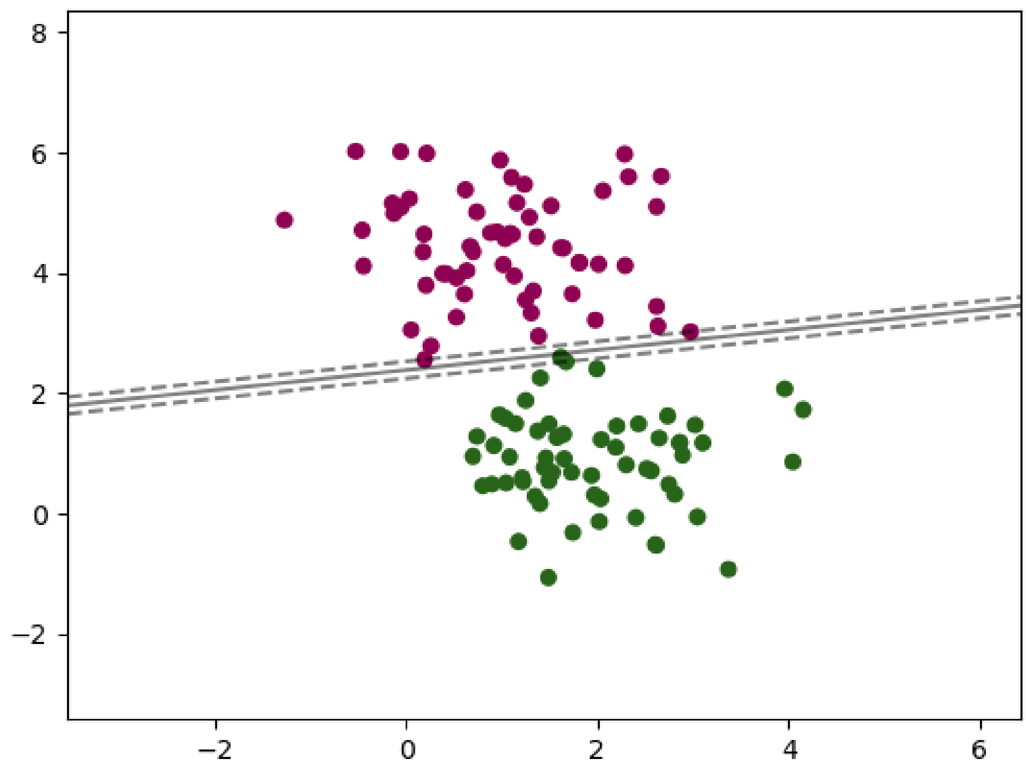
קיבלנו את הדרוש.

<u>מטלה מעשית:</u>

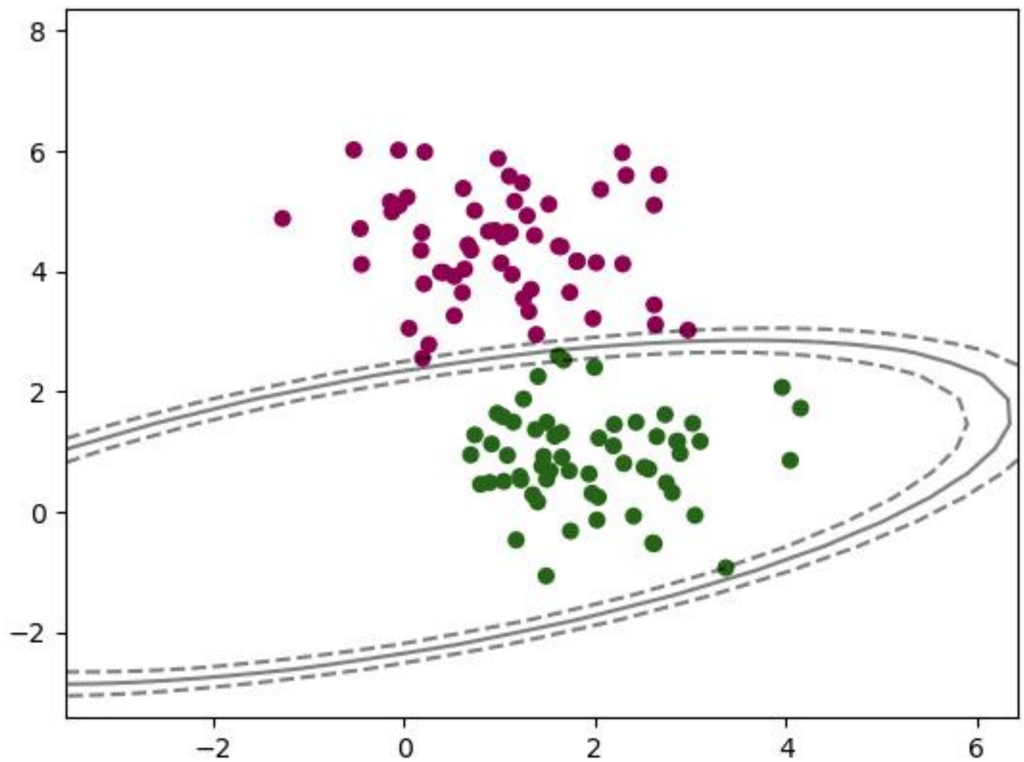
א.

ראשית, מבחינת כמות הוקטורים התומכים (support vectors), במקרה הלינארי ישנם 3, במקרה הריבועי ארבעה, ובמקרה של RBF ישנם שישה.

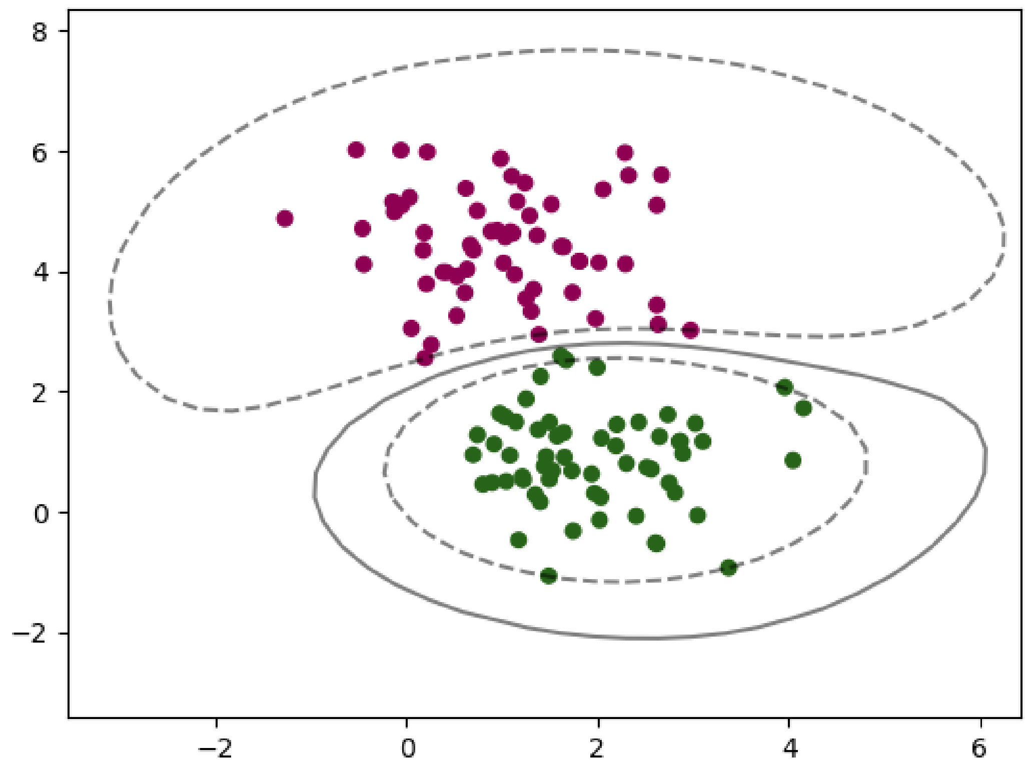
<u>המפריד הלינארי:</u>



המפריד הריבועי :

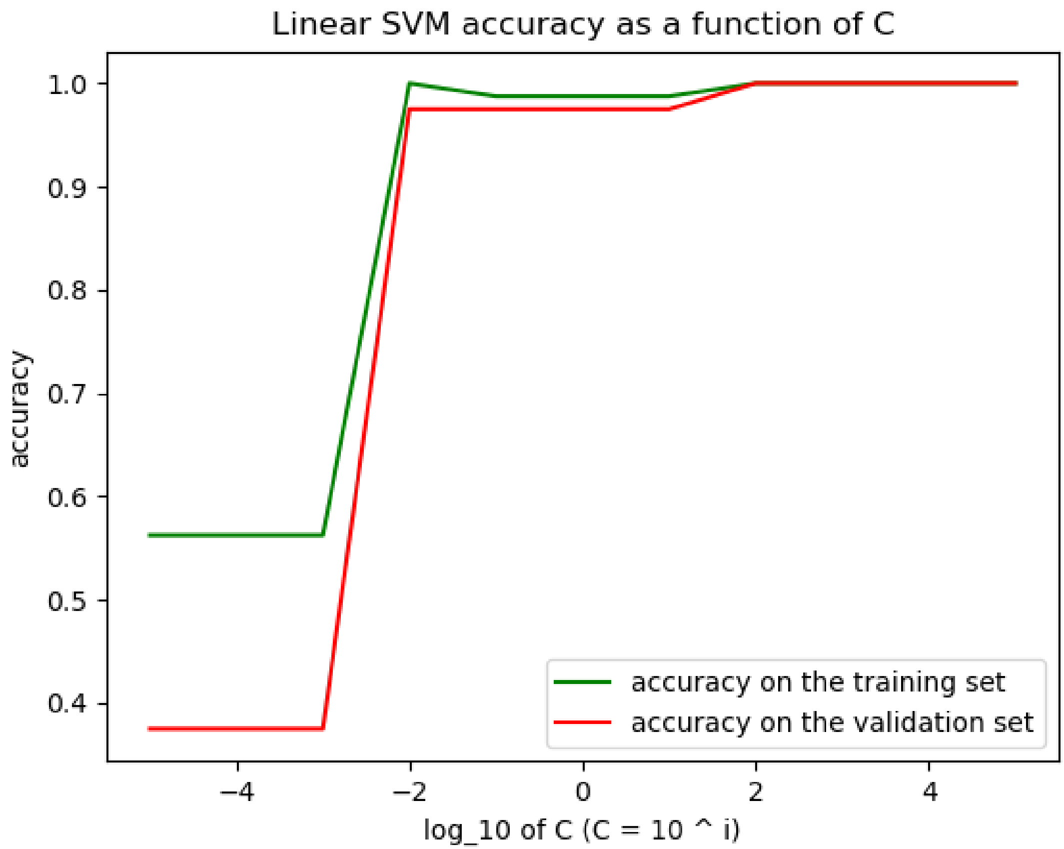


מפריד RBF:



ניתן לראות כי כל המפרידים מפרידים היטב כפי שהיינו מציעים, אך כי במעבר מכל אחד מהם המפריד הולך ונהיה יותר ויותר אקספרסיבי. כלומר המפריד הראשון בסה"כ מראה כי ניתן להפריד את הנקודות על ידי ישר, עד אשר שבמפריד האחרון כבר ניתן לראות כמעט "מעטפת" סביב הנקודות – צורתן מובנת לחלוטין.

ב.



כפי שניתן לראות בגרף, ה – C האופטימלי הוא 100 (C הקטן ביותר עם הדיוק הגבוהה ביותר).