תרגיל 4 מבוא ללמידה חישובית

.1

(a

נפתור את בעיית האופטימיזציה :

$$arg\min_{y}||x-y||$$

$$||y|| \leq R$$

ששקולה לבעיה הקמורה:

$$argmin_y ||x-y||^2$$

$$||y||^2 \leq R^2$$

$$\mathcal{L}(y,\lambda) = ||x-y|| + \lambda\left(||y||^2 - R^2\right)$$

$$\frac{\partial \mathcal{L}(y,\lambda)}{\partial y} = 2(y-x) + 2\lambda y = 0 \Rightarrow y = \frac{x}{1+\lambda}$$

בנוסף מComplementary Slackness מתקיים:

$$\lambda\left(||y||^2 - R^2\right) = 0$$

אם $\lambda = 0$: מתקיים $x = y$

אחרת $||y|| = R \Rightarrow ||x|| = R(1+\lambda) \Rightarrow \lambda = \frac{||x||-R}{R} \Rightarrow y = \frac{R}{||x||}x$

כלומר

$$\Pi_{\mathcal{K}}(x) = \begin{cases} x & ||x|| \leq R \\ \frac{R}{||x||}x & else \end{cases}$$

לכן

$$w_{t+1} = \begin{cases} w_t - \eta_t \nabla f_i(w_t) & ||w_t - \eta_t \nabla f_i(w_t)|| \leq R \\ \frac{R(w_t - \eta_t \nabla f_i(w_t))}{||w_t - \eta_t \nabla f_i(w_t)||} & else \end{cases}$$

נשים לב ש$y_i w x_i > 1$ אז $w_{t+1} = w_t$ אחרת $w_{t+1} = w_t + y_i x_i$

כלומר נקבל עבור hinge loss

$$w_{t+1} = \begin{cases} w_t & y_i w x_i > 1 \\ \begin{cases} w_t + y_i x_i & ||w_t + y_i x_i|| \le R \\ \frac{R}{||w_t + y_i x_i||}(w_t + y_i x_i) & else \end{cases} \end{cases}$$

b) תהי $z \in \mathcal{K}$ אם $y \in \mathcal{K}$ אז $x = y$ לכן $||y - z|| = ||x - z||$

אם $y \notin \mathcal{K}$ אז הישרים $x - z$, $y - x$ מאונכים מכיוון ש $x = \Pi_{\mathcal{K}}(y)$ לכן המשולש המוגדר על ידי הנקודות x, y, z הוא ישר זווית ומתקיים שהישר $y - z$ הוא היתר לכן מתקיים $||y - z|| > ||x - z||$ כנדרש

C) נעתיק את ההוכחה

5.4. *OPTIMIZATION ALGORITHMS*

Theorem 5.4.1. *Let $\epsilon > 0$. For $w_1 = w' = 0$, $\eta_t \equiv \frac{\epsilon}{G^2}$ and $T = \frac{B^2 G^2}{\epsilon^2}$ we have that*

$$f(\bar{w}) - f(w^*) \le \epsilon$$

Proof. From the definition of convexity:

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

Subtracting the optimal value from both sides, we have:

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

From convexity, $f(\cdot)$ lies above its tangent at the point w_t, and so:

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

For any two vectors u, v we have the property that $\|u\|_2^2 - \|u - \eta v\|_2^2 = 2\eta\langle u, v\rangle - \eta^2\|v\|_2^2$. Setting $u = w_t - w^*$, $v = \nabla f(w_t)$, while noting that $u - \eta v = w_t - \eta\nabla f(w_t) - w^* =$

מתקיים $w^*, w_{t+1} \in \mathcal{K}$ לכן לפי סעיף קודם :

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

לכן נקבל

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

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

$$= \frac{\|w_1 - w^*\|_2^2}{2\eta T} - \frac{\|w_{T+1} - w^*\|_2^2}{2\eta T} + \frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_2^2$$
$$\le \frac{B^2}{2\eta T} + \frac{\eta G^2}{2} = \epsilon$$

2

נציב $k = 2$

$$f(w_1, w_2) = \frac{\beta}{2}\left(||w_1||^2 + ||w_2||^2\right) + \ell(w_1, w_2, x_i, y_i)$$
$$= \frac{\beta}{2}\left(||w_1||^2 + ||w_2||^2\right) + \frac{1}{2}\max_{j\in[2]}(w_j x_i - w_{y_i} x_i + \mathbb{1}(y_1 \neq j))$$
$$= \frac{\beta}{2}\left(||w_1||^2 + ||w_2||^2\right) + \frac{1}{2}\max(0,1 - (2y_i - 3)(w_2 - w_1)x_i)$$

נבצע החלפת משתנים נסמן $u = w_1 + w_2, v = w_2 - w_1$ נקבל

$$\frac{\beta}{2}\left(\left|\left|\frac{1}{2}(u - v)\right|\right|^2 + \left|\left|\frac{1}{2}(u + v)\right|\right|^2\right) + \frac{1}{2}\max(0,1 - y_i' u x_i), \quad y_1' = 1, y_2' = -1$$

$$\frac{\beta}{8}\left(||u - v||^2 + ||u + v||^2\right) + \frac{1}{2}\ell_{hinge}(u, x_i, y'_i)$$
$$= \frac{\beta}{8}\left(||u||^2 - 2 < u, v > + ||v||^2 + ||u||^2 + 2 < u, v > + ||v||^2\right)$$
$$+ \frac{1}{2}\ell_{hinge}(u, x_i, y'_i) = \frac{\beta}{4}\left(||v||^2 + ||u||^2\right) + \ell_{hinge}(u, x_i, y'_i)$$

נשים לב ש$\ell_{hinge}(u, x_i, y'_i)$ לא תלויה בv לכן המינימום של $f(u, v)$ התקבל כאשר $v^* = 0$ והפתרון האופטימלי הוא $\frac{\beta}{4}||u^*|| + \ell_{hinge}(u^*, x_i, y'_i)$ לכן

$$w^*(\beta') = u^*(\beta) = w_2^*(\beta) - w_1^*(\beta) \quad and\ \beta' = \frac{1}{2}\beta$$

.3

$$\min_{w,b,\xi} \frac{1}{2}w^T w + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2$$
$$\text{s.t. } y_i(w^T x_i + b) \ge 1 - \xi_i \quad \forall i = 1, \ldots, m$$

(a

נראה שהאילוץ $\forall i\ \xi_i \geq 0$ אינו משנה את הפתרון של בעיית האופטימיזציה כלומר נניח בשלילה שקיים פתרון אופטימלי כך ש$\xi_{i_0} < 0$ אז

$$y_{i_0}\left(w^T x_{i_0} + b\right) \geq 1 - \xi_{i_0} \Rightarrow\ y_{i_0}\left(w^T x_{i_0} + b\right) > 1$$

לכן מצאנו פתרון פיזיבלי שמקטין את פונקציית המטרה בסתירה לכך שהפתרון שמקיים $\xi_{i_0} < 0$

(b

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 + \sum_{i=1}^{m}\lambda_i(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b}) - \xi_i) =$$

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 + \sum_{i=1}^{m}\lambda_i\left(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b})\right) - < \boldsymbol{\xi}, \boldsymbol{\lambda} >$$

(c

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

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

$$\nabla_\xi\mathcal{L} = C\boldsymbol{\xi} - \boldsymbol{\lambda} = 0 \Rightarrow \boldsymbol{\xi} = \frac{\boldsymbol{\lambda}}{C}$$

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 + \sum_{i=1}^{m}\lambda_i\left(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b})\right) - < \boldsymbol{\xi}, \boldsymbol{\lambda} > \Rightarrow$$

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{m}\lambda_j\lambda_i y_i y_j \boldsymbol{x_i}\boldsymbol{x_j} + \frac{C}{2}\left|\left|\frac{\boldsymbol{\lambda}}{C}\right|\right|^{\boldsymbol{2}} - \frac{\left||\boldsymbol{\lambda}|\right|^{\boldsymbol{2}}}{C} + \sum_{i=1}^{m}\lambda_i$$
$$- \sum_{j=1}^{m}\sum_{i=1}^{m}\lambda_j\lambda_i y_i y_j \boldsymbol{x_i}\boldsymbol{x_j} - \boldsymbol{b}\sum_{i=1}^{m}\lambda_i y_i$$
$$= \sum_{i=1}^{m}\lambda_i - \frac{\left||\boldsymbol{\lambda}|\right|^2}{2C} - \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{m}\lambda_j\lambda_i y_i y_j \boldsymbol{x_i}\boldsymbol{x_j}$$

(d

נשים לב שאם $\sum_{i=1}^{m}\lambda_i y_i \neq 0$ אז $\mathcal{L}(\lambda) = -\infty$ לכן הבעיה הדואלית היא :

$$\max_{\lambda}\sum_{i=1}^{m}\lambda_i - \frac{||\boldsymbol{\lambda}||^2}{2C} - \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{m}\lambda_j\lambda_i y_i y_j \boldsymbol{x_i}\boldsymbol{x_j}$$

$$s.t: \qquad \sum_{i=1}^{m}\lambda_i y_i \neq 0$$

4. נמצא מיפוי ϕ נגדיר:

$$\phi: [k] \to \{0,1\}^k \ , \phi(x) = (1,\dots,1,0,\dots,0) \ \left(x \text{ פעמים } 1\right)$$

$$\phi(x)\phi(y) = \min\{x,y\}$$

5.

יהי $x, x' \in \mathbb{R}^d$ נתון כי *K* גרעין לכן מתקיים

$$k(x,x') = \phi(x)\phi(x')$$

נחזיק מערך *A* בגודל *n* שהכניסה ה $i-1$ מכילה את המקדם של $\phi(x_i)$

נסמן $\mathbf{0}$ $w_0 =$ כלומר *A=[0,…,0]* לכן $w_1 = Cy_i\phi\left(x_{i_1}\right)$ כלומר נעדכן את כניסה ה $i_1 - 1$ במערך מאפס ל-Cy_i ואז $w_2 = (1-\eta)w_1 + Cy_{i_2}\phi\left(x_{i_2}\right)$ לכן צריך להעדכן את הכניסות $i_1 - 1, i_2 - 1$ לאחר *t* צעדים $w_t = \sum_{i=1}^{n}A_{i-1}\phi(x_i)$ נשים לב שכל צעד דורש עדכון של לכל היותר *n* תאים ובסוף נוכל לחשב $w^* = \sum_{i=1}^{n}A_{i-1}\phi(x_i)$

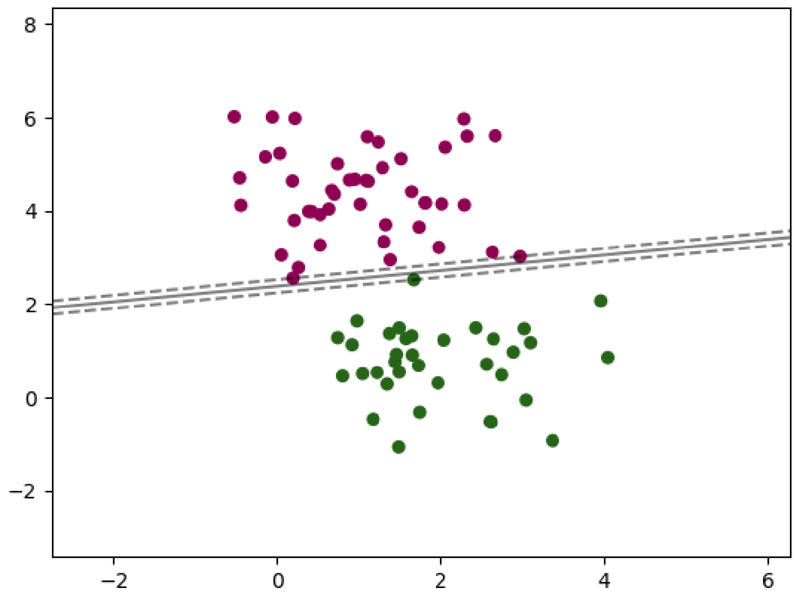
ולבצע סיווג על ידי

$$sign(w^*x) = sign\left(\sum_{i=1}^{n}A_{i-1}\phi(x_i)\phi(x)\right) = sign\left(\sum_{i=1}^{n}A_{i-1}k(x_i,x)\right)$$

<u>חלק מעשי</u>

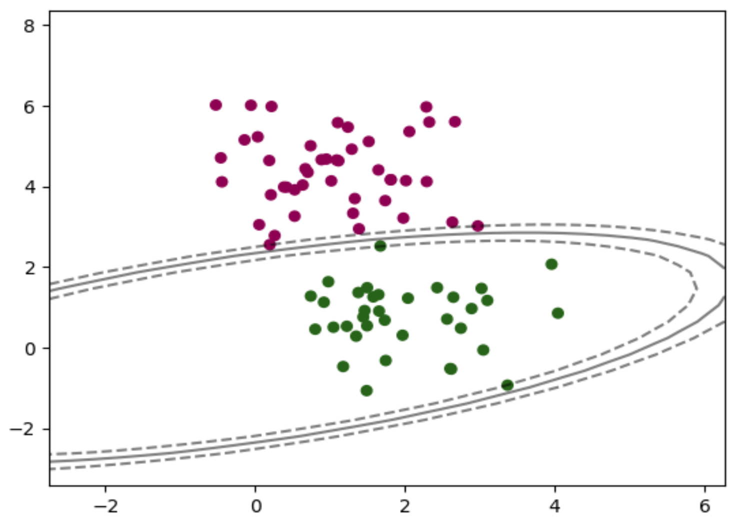
.1

a) המסווג עם הגרעין הלינארי:



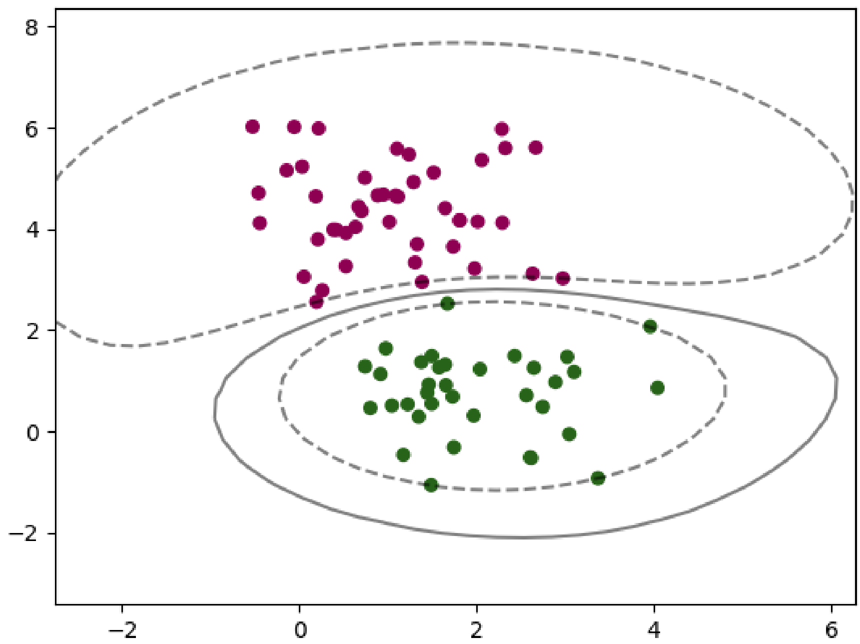
כפי שניתן לראות הוא מפריד נכון את הדאטה על ידי פונקציה לינארית ומספר הווקטורים התומכים הוא 3

המסווג הריבועי:



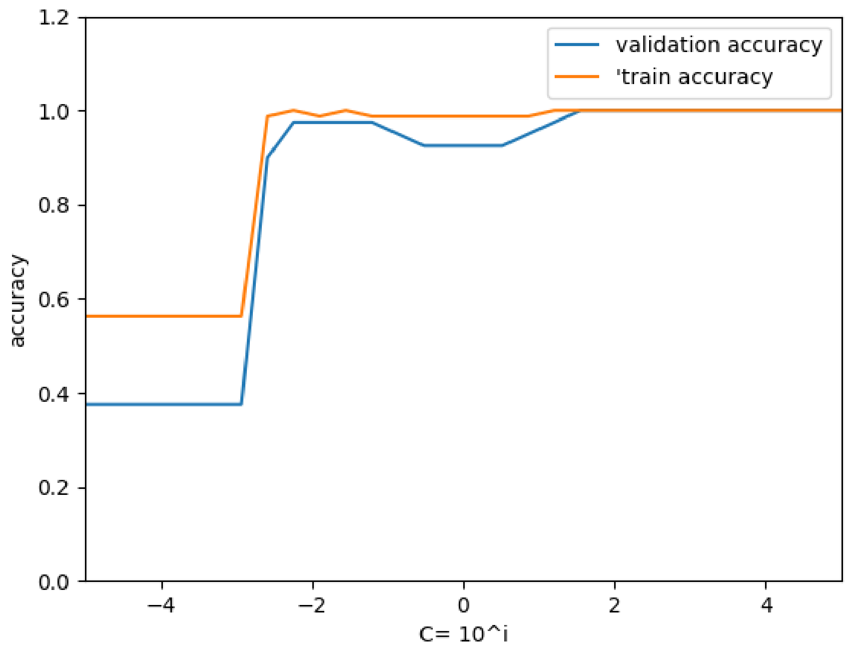
רואים שהמסווג הוא פונקציה ריבועית(פרבולות) ושיש 4 וקטורי תמיכה בנוסף הוא נותן צורה לקבוצה הירוקה

מסווג עם גרעין *RBF*:



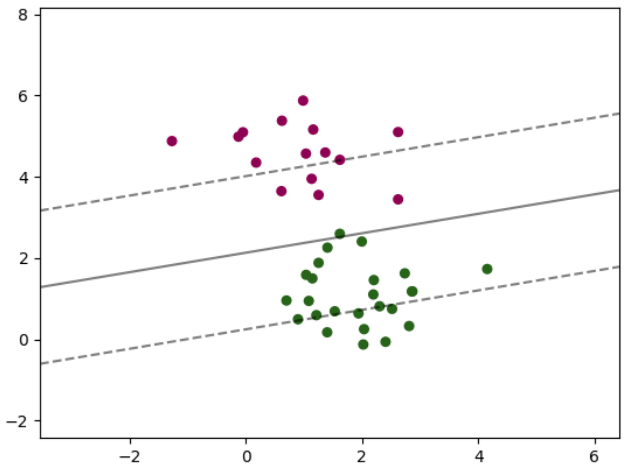
ניתן לראות שהמסווג נותן צורה לשני הקבוצות ושיש שבעה וקטורי תמיכה

b(

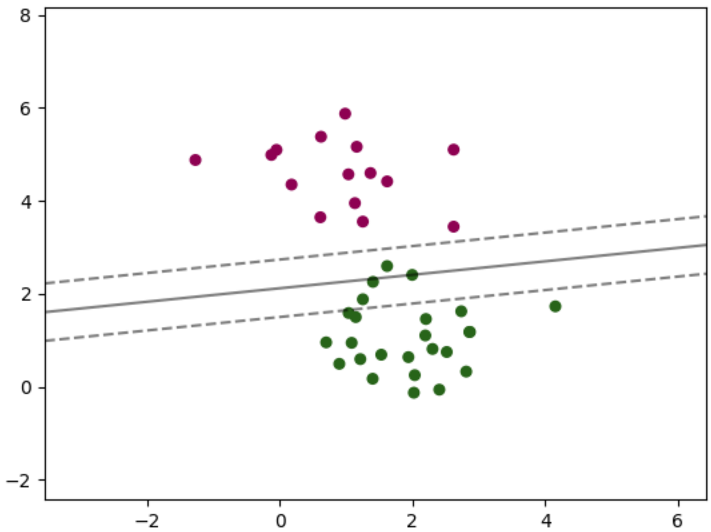


ניתן שרטוטים של $C = 10^{-2}, 1, 100$ ניתן לראות שככל שC נמוך יותר המסווג נותן עדיפות למרחק ההחלטה על גבי סיווג נכון של הנקודות כמו כן ניתן לראות שמ $C = 100$ התוצאות הדיוק הוא זהה לכן זהו הC הטוב ביותר

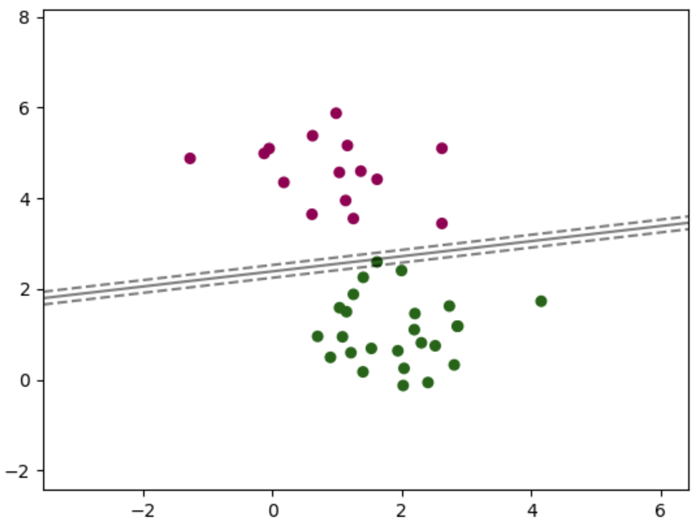
$C = 10^{-2}$



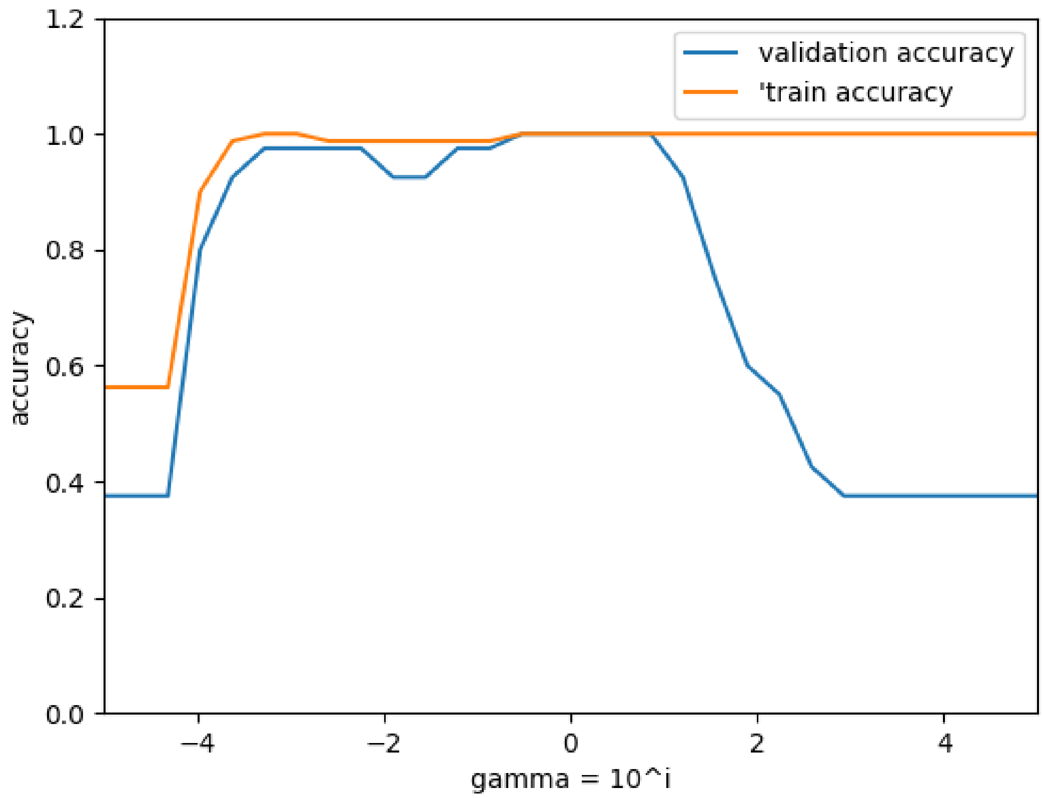
$C = 1$



$C = 100$

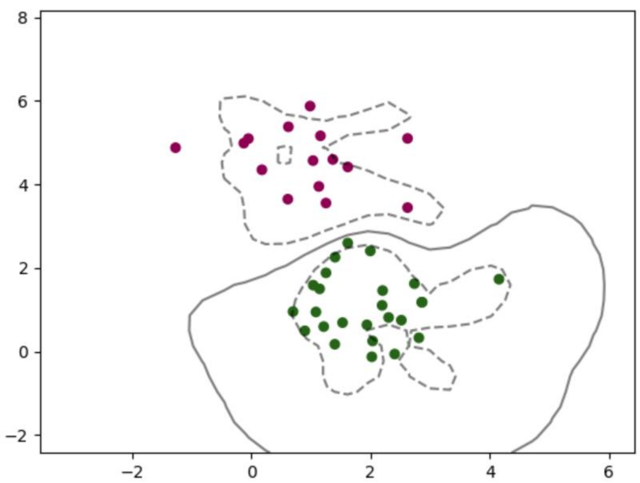

8
6
4
2
0
−2
−2
0
2
4
6

$($c


validation accuracy
'train accuracy
accuracy
gamma = 10^i
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4

ניתן לראות שהשגיאה הולכת ומשפרת עד לאיזור $\gamma = 1$ ולאחר מכן אנו מבצעים אוברפיט על הדאטה

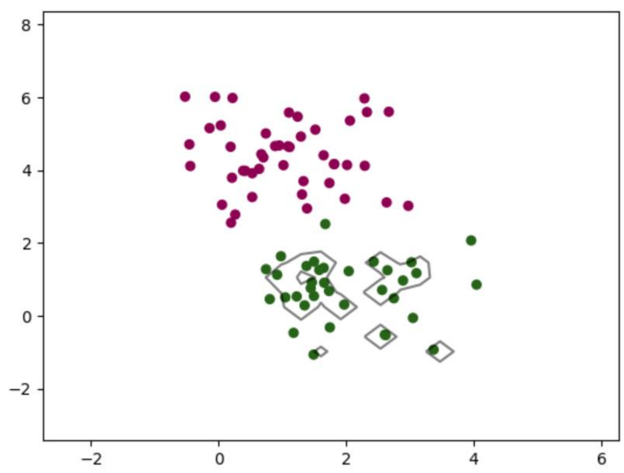
$\gamma = 1$:



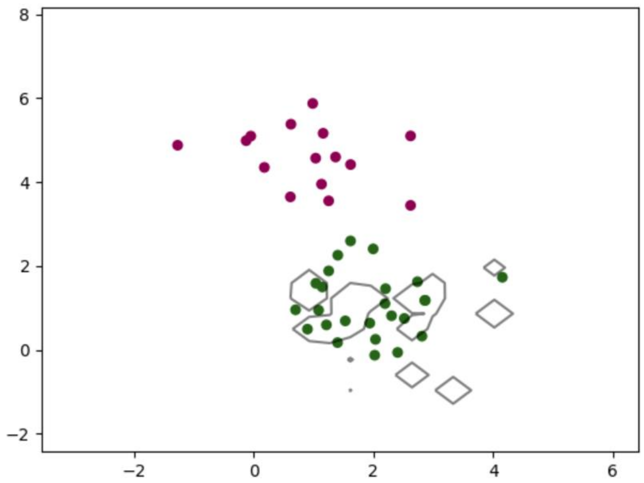
ניתן לראות שהמסווג מיצר צור שכבר ממש הדוקה על הדאטה

$\gamma = 100$:

על הדאטה האימון



על *validation data*



ניתן לראות שהמסווג ממש הדוק על דאטה האימון ככה שהוא מפספס את *validation data* כלומר שי אובר פיט של המסווג לדאטה האימון.