Theory Questions

1. **(20 points) SVM with multiple classes.** One limitation of the standard SVM is that it can only handle binary classification. Here is one extension to handle multiple classes. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and now let $y_1, \ldots, y_n \in [K]$, where $[K] = \{1, 2, \ldots, K\}$. We will find a separate classifier $\mathbf{w}_j$ for each one of the classes $j \in [K]$, and we will focus on the case of no bias ($b = 0$). Define the following loss function (known as the *multiclass hinge-loss*):

$$\ell(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i) = \max_{j \in [K]} (\mathbf{w}_j \cdot \mathbf{x}_i - \mathbf{w}_{y_i} \cdot \mathbf{x}_i + \mathbb{1}(j \neq y_i)),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Define the following multiclass SVM problem:

$$f(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i)$$

After learning all the $\mathbf{w}_j$, $j \in [K]$, classification of a new point $\mathbf{x}$ is done by $\arg\max_{j \in [K]} \mathbf{w}_j \cdot \mathbf{x}$. The rationale of the loss function is that we want the "score" of the true label, $\mathbf{w}_{y_i} \cdot \mathbf{x}_i$, to be larger by at least 1 than the "score" of each other label, $\mathbf{w}_j \cdot \mathbf{x}_i$. Therefore, we pay a loss if $\mathbf{w}_{y_i} \cdot \mathbf{x}_i - \mathbf{w}_j \cdot \mathbf{x}_i \leq 1$, for $j \neq y_i$.

Consider the case where the data is linearly separable. Namely, there exists $\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*$ such that $y_i = argmax_y \mathbf{w}_y^* \cdot \mathbf{x}_i$ for all i. Show that any minimizer of $f(\mathbf{w}_1, \ldots, \mathbf{w}_K)$ will have zero classification error.

סימון: $\bar{w} = (w_1, \ldots, w_K)$

~~טענה~~ טענת עזר: $f(\bar{w}) = 0 \Rightarrow$ $\bar{w}$ מסווג עם שגיאה אמפירית 0 ומתקיים $\forall i \in [n]: y_i = \operatorname{argmax}_{j \in [K]} \langle w_j, x_i \rangle$

הוכחת טענת עזר:

ראשית, נזכיר כי $0 \leq \ell(\bar{w}, x_i, y_i)$ לכל i; כיוון שאם $j = y_i$ אז $\ell(\bar{w}, x_i, y_i) = 0$

ולכן $\ell(\bar{w}, x_i, y_i)$ היא מקסימום בין 0 למספרים אחרים, כלומר הוא אי שלילי.

לכן אם $0 = f(\bar{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\bar{w}, x_i, y_i)$; ℓ אי שלילי נקבל שלכל i; $\ell(\bar{w}, x_i, y_i) = 0$,

כלומר השגיאה האמפירית 0.

כעת אם $j = y_i$ אז ברור שמתקיים $\forall i \in [n]: y_i = \operatorname{argmax}_j \langle w_j, x_i \rangle$

אחרת אם $j \in [K] \setminus \{y_i\}$ נקבל:

$$0 = f(\bar{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\bar{w}, x_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq y_i} \{\langle w_j, x_i \rangle - \langle w_{y_i}, x_i \rangle + 1, \underbrace{0}_{\text{כאשר } j=y_i}\}$$

$\iff w_j \cdot x_i - w_{y_i} \cdot x_i + 1 \leq 0 \iff w_j \cdot x_i < w_{y_i} \cdot x_i$ ולכן $y_i = \operatorname{argmax}_j w_j \cdot x_i$ כנדרש.

לכן עלינו למצוא מסווג w' כך ש $f(w') = 0$ וזה יבטיח של w' יש שגיאה אמפירית 0.

מההנחה שהדאטה "ניתן להפרדה לינארית", קיים $w^* = (w_1^*, \ldots, w_K^*)$ כך ש $y_i = \operatorname{argmax}_j w_j^* \cdot x_i$

לכל i.

כדי להראות ש $f(w^*, x_i, y_i) = 0$ לכל i צריך שיתקיים לכל $i \in [n], j \in [K] \setminus \{y_i\}$

כדי להראות ש $f(w^*, x_i, y_i) = 0$ לכל i צריך שיתקיים לכל $i \in [n]$, ולכל $j \in [K] \setminus \{y_i\}$

$w_j^* \cdot x_i - w_{y_i}^* \cdot x_i \le -1$ שכן אם האינדיקטור שונה ל-1 אז f תגדיר מקסימום של 0 עם מס' שלילי שבה 0. (אם $j = y_i$ אז ברור ש f מתאפס).

סימונים: $\varphi_i := \min_{j \in [K]} w_{y_i}^* \cdot x_i - w_j^* \cdot x_i$, $\varphi = \min_{i \in [n]} \varphi_i$

כעת אם ניקח $\forall j \neq y_i : w_j = \frac{w_j^*}{\varphi}$ יתקיים:

$$\forall j \in [K] \setminus \{y_i\},\ i \in [n]: \quad w_j \cdot x_i - w_{y_i} \cdot x_i = \frac{w_j^* \cdot x_i}{\varphi} - \frac{w_{y_i}^* \cdot x_i}{\varphi} = \frac{1}{\varphi}\left(w_j^* \cdot x_i - w_{y_i} \cdot x_i\right)$$

$$= \underbrace{-\frac{1}{\varphi}}_{\le 0} \underbrace{\left(w_{y_i}^* \cdot x_i - w_j^* \cdot x_i\right)}_{\ge \varphi_i} \le -\frac{1}{\varphi} \cdot \varphi_i = -1 \cdot \underbrace{\frac{\varphi_i}{\varphi}}_{\ge 1} \le -1$$

לכן כל מחובר של f יתקיים לכל i $loss = 0$ וכן גם $f = 0$

ומאחר הערך יהיה טעות אמפירית 0. ■

3. **(15 points)** ℓ^2 **penalty.** Consider the following problem:

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

(a) Show that a constraint of the form $\xi_i \ge 0$ will not change the problem. Meaning, show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

לאורך כל השאלה: $\varepsilon := \xi_i$

נסמן ב- P את הבעיה המקורית ונסמן את הערך האופטימלי שלה ב $OPT(P)$.

נניח בשלילה כי עבור בעיה אחרת P' מתקיים $OPT(P) = OPT(P')$ אבל קיים $i \in [n]$ כך ש $\varepsilon_i < 0$.

נראה שאם נחליף את ε_i ב $\varepsilon_i' = 0$ נשמור על האילוצים ונקבל ערך פונ' מטרה קטן יותר מהמטרה האופטימלית.

נראה שאם נחליף את ε_i ב $\varepsilon'_i = 0$ נשמור על האילוצים ונקבל ערך פונ' מטרה קטן יותר
בסתירה לאופטימליות.

אכן האילוצים נשמרים: $y_i(w^T \cdot x_i + b) \geq 1 - \overset{<0}{\varepsilon_i} \geq 1 - \varepsilon'_i = 1 - 0 = 1$

ובנוסף, $(\varepsilon'_i)^2 = 0 < \varepsilon_i^2$ ולכן ערך פונ' המטרה קטן יותר, בסתירה. ∎

(b) What is the Lagrangian of this problem?

$$\mathcal{L}(w, b, \varepsilon, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\cdot\sum_{i=1}^{n}\varepsilon_i^2 + \sum_{i=1}^{n}\alpha_i\left[1 - \varepsilon_i - y_i(w^T x_i + b)\right]$$

(c) Minimize the Lagrangian with respect to $\boldsymbol{w}, b, \boldsymbol{\xi}$ by setting the derivative with respect to these variables to 0.

תחילה נפשט:

$$\mathcal{L}(w, b, \varepsilon, \alpha) = \overset{*}{\frac{1}{2}\|w\|^2} + \overset{*}{\frac{C}{2}\sum_{i=1}^{n}\varepsilon_i^2} + \overset{*}{\sum_{i=1}^{n}\alpha_i} - \overset{*}{\sum_{i=1}^{n}\alpha_i\varepsilon_i} - \overset{*}{\sum_{i=1}^{n}\alpha_i y_i w^T x_i} - \overset{*}{\sum_{i=1}^{n}\alpha_i y_i b}$$

נגזור לפי כל משתנה ונשווה לס:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad (i)$$

$$\nabla_b \mathcal{L} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \qquad (ii)$$

נשים לב שאם $\sum_{i=1}^{n}\alpha_i y_i \neq 0$ אז הלגרנז'יאן אינו חסום מלמטה ולכן לא קיים מינימום.

$$\nabla_\varepsilon \mathcal{L} = C\varepsilon - \alpha = 0 \iff \varepsilon = \frac{\alpha}{C} \qquad (iii)$$

נשתמש במשוואות אלו ונציב בלגרנז'יאן, לפי הסימונים של הצבעים:

$$* = \frac{1}{2}\|w\|^2 - \overbrace{w^T\sum_{i=1}^{n}\alpha_i y_i x_i}^{\text{לפי (i)}\; w} = \frac{1}{2}\|w\|^2 - \|w\|^2 = -\frac{1}{2}\|w\|^2 \overset{(i)}{=} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$$

$$* = -\underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{\text{לפי (ii)}\; 0} b = 0$$

לפי (iii)

(ii) לפי

(iii) לפי

$$* = \frac{C}{2}\sum_{i=1}^{n} \varepsilon_i^2 - \sum_{i=1}^{n}\alpha_i \varepsilon_i \overset{\downarrow}{=} \frac{C}{2}\sum_{i=1}^{n}\frac{\alpha_i^2}{C^2} - \sum_{i=1}^{n}\alpha_i \cdot \frac{\alpha_i}{C} = \frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2 - \frac{1}{C}\sum_{i=1}^{n}\alpha_i^2 =$$

$$= \frac{1}{C}\sum_{i=1}^{n}\alpha_i^2 \cdot \left(\frac{1}{2}-1\right) = -\frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2 = -\frac{1}{2C}\|\alpha\|^2$$

סה"כ:

$$\min_{w,b,\varepsilon} \mathcal{L}(w,b,\varepsilon,\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2C}\|\alpha\|^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$$

(d) What is the dual problem?

נסמן $g(\alpha) = \min_{w,b,\varepsilon} \mathcal{L}(w,b,\varepsilon,\alpha)$

הבעיה הדואלית היא $\max_{\alpha} g(\alpha)$ כאשר האילוצים הם האילוץ $\sum_{i=1}^{n}\alpha_i y_i = 0$ שקיבלנו כבר

ואילוץ אי שלילות על כופלי הלגרנז' α_i.

סה"כ:

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2C}\|\alpha\|^2 - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \qquad (D)$$

$$\alpha_i \geq 0 \qquad i \in [n]$$

5. **(15 points) Separability using polynomial kernel.** Let $x_1, ..., x_n \in \mathbb{R}$ be distinct real numbers, and let $q \geq n$ be an integer. Show that when using a polynomial kernel, $K(x, x') = (1 + xx')^q$, hard SVM achieves zero training error. Use the following fact: Given distinct values $\alpha_1, ..., \alpha_n$, the Vandermonde matrix defined by,

$$\begin{pmatrix} 1 & \alpha_1^1 & \dots & \alpha_1^q \\ 1 & \alpha_2^1 & \dots & \alpha_2^q \\ \vdots & & & \\ 1 & \alpha_n^1 & \dots & \alpha_n^q \end{pmatrix},$$

is of rank n. (Hint: use the lemma from slide 7 in recitation 7).

$$\langle\phi(x_i),\phi(x_j)\rangle = K_q(x_i,x_j) \overset{x_i\in\mathbb{R}}{=} (1+x_ix_j)^q \overset{\text{בינום}}{=} \sum_{m=0}^{q} \binom{q}{m}\cdot 1^{q-m}\cdot(x_ix_j)^m = \sum_{m=0}^{q}\binom{q}{m}x_i^m x_j^m =$$

$$\sum_{m=0}^{q} \binom{q}{m}^{1/2}\cdot\binom{q}{m}^{1/2} x_i^m\cdot x_j^m = \sum_{m=0}^{q} \left[\binom{q}{m}^{1/2}\cdot x_i^m\right]\cdot\left[\binom{q}{m}^{1/2}\cdot x_j^m\right]$$

כלומר קיבלנו:

$$[\phi(x_i)]^t = \left[\binom{q}{0}^{1/2}\cdot x_i^0,\ \binom{q}{1}^{1/2}\cdot x_i^1,\ \dots,\ \binom{q}{q}^{1/2}\cdot x_i^q\right]$$

$$[\phi(x_j)]^t = \left[\binom{q}{0}^{1/2}\cdot x_j^0,\ \binom{q}{1}^{1/2} x_j^1,\ \dots,\ \binom{q}{q}^{1/2}\cdot x_j^q\right]$$

כעת נשים לב שמכיוון ש $x_1,\dots,x_n$ שונים זה מזה, מתקיים שהמטריצה

$$\begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & \cdots & x_1^q \\ x_2^0 & x_2^1 & x_2^2 & \cdots & x_2^q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n^0 & x_n^1 & x_n^2 & & x_n^q \end{bmatrix}$$

הינה מטריצת Vandermonde ולכן דרגתה n.

כפי שראינו בתרגול, אם נכפול כל עמודה במטריצה בסקלר שאינו 0, הדרגה תישאר n.

לכן נכפול כל עמודה $j=0,\dots,q$ ב $\binom{q}{j}^{1/2}$ ונקבל את המטריצה:

$$\begin{bmatrix} \binom{q}{0}^{1/2}\cdot x_1^0 & \binom{q}{1}^{1/2} x_1^1 & \cdots & \binom{q}{q}^{1/2} x_1^n \\ \binom{q}{0}^{1/2}\cdot x_2^0 & \ddots & & \vdots \\ \vdots & & & \\ \binom{q}{q}^{1/2} x_n^0 & \cdots & & \binom{q}{q}^{1/2}\cdot x_n^n \end{bmatrix}$$

נשים לב שהשורה ה-i של המטריצה היא בדיוק $[\phi(x_i)]^t$.

נסמן מטריצה זו ב ϕ_s. עפ"י הלמה מהתרגול אם ϕ_s מדרגה n אזי המדגם לאחר הטרנספורמציה ϕ "ניתן להפרדה לינארית" ולכן השגיאה האמפירית של HardSVM עם $K_q(x_i,x_j)$ תהיה 0. ∎

4. **(15 points) Soft SVM on separable data.** Consider the soft-SVM problem with linearly separable data (assume no bias for simplicity):

$$\min_{\mathbf{w},\xi} 0.5\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$s.t\ \forall i: \quad y_i \mathbf{w}\cdot\mathbf{x}_i \ge 1-\xi_i$$
$$\xi_i \ge 0$$

Let $\mathbf{w}^\star$ be the solution of **hard SVM**. Show that if $C \ge \|\mathbf{w}^\star\|^2$ then the solution of soft SVM separates the data. (Hint: Show that in the optimal solution $(\mathbf{w}', \boldsymbol{\xi}')$ of the soft SVM problem, the sum of ξ_i''s is bounded by some constant smaller than 1).

סימון לאורך כל השאלה $\varepsilon_i := \xi$

יהי w^* פתרון אופטימלי לבעיית Hard SVM ; (w',ε') פתרון אופטימלי לבעיית soft SVM.

ראשית נראה כי מתקיים $\|w^*\| \ge \frac{1}{2}\|w'\|^2 + C\cdot\sum_{i=1}^{n}\varepsilon'_i$.

קודם כל w^* קיים כיוון שנתון שהדאטה ספרבילי.

כעת, אם נסתכל על בעיית ה Hard SVM כך שנתון פונ' המטרה שלה יהיה $\frac{1}{2}\|w^*\|^2 + C\cdot\sum_{i=1}^{n}\varepsilon^*_i$ כש $\forall i: \varepsilon^*_i = 0$ נקבל שזוהי בעיה שקולה לבעיית ה Hard SVM המקורית (נניח שקיימים אילוצי אי שליליות על ε_i אבל גם נתון כל $\varepsilon_i = 0$).

כעת מכיוון ש (w',ε') הוא פתרון אופטימלי לבעיית ה soft ; (w^*,ε^*) הוא פתרון פיזיבילי כלשהו לבעיית ה soft נקבל,

$$\frac{1}{2}\|w'\|^2 + C\cdot\sum_{i=1}^{n}\varepsilon'_i \le \frac{1}{2}\|w^*\|^2 + C\cdot\overbrace{\sum_{i=1}^{n}\varepsilon^*_i}^{0} = \frac{1}{2}\|w^*\|^2$$

כפי שרצינו.

כעת נמשיך עם אלגברה:

$$\frac{1}{2}\|w'\|^2 + C\sum_{i=1}^{n}\varepsilon'_i \le \frac{1}{2}\|w^*\|^2 \iff \sum_{i=1}^{n}\varepsilon'_i < \frac{\frac{1}{2}\|w^*\|^2 - \frac{1}{2}\|w'\|^2}{C} \overset{C \ge \|w^*\|^2}{\le} \frac{\frac{1}{2}\|w^*\|^2 - \frac{1}{2}\|w'\|^2}{\|w^*\|^2} =$$

$$= \frac{1}{2} - \frac{\frac{1}{2}\|w'\|^2}{\|w^*\|^2}$$

מתקיים: $\frac{\|w'\|^2}{\|w^*\|^2} \in [0,1]$ כי כאמור $\|w'\|^2 \le \|w^*\|^2$ ולכן כשנכפול ב $\frac{1}{2}$ ונחסר את התוצאה מ $\frac{1}{2}$ נקבל ביטוי ששייך לקטע $(1,0]$. כלומר מתקיים $\sum_{i=1}^{n}\varepsilon'_i \le 1$.

כפי שראינו בהרצאה עבור $\varepsilon_i \in [0,1)$ יש לנו סיווג נכון ואנו טועים רק אם $\varepsilon_i \ge 1$.

מכיוון שהסכום $\sum_{i=1}^{n}\varepsilon'_i \le 1$ נקבל שהפתרון (w',ε') לבעיית ה soft מפריד את הדאטה במדויק. ∎

2. **(15 points) Solving hard SVM.** Consider two distinct points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ with labels $y_1 = 1$ and $y_2 = -1$. Compute the hyperplane that Hard SVM will return on this data, i.e., give explicit expressions for $\boldsymbol{w}$ and b as functions of $\mathbf{x}_1$, $\mathbf{x}_2$.
(Hint: Solve the dual problem by transforming it to an optimization problem in a single variable. Use your solution to the dual to obtain the primal solution).

תחילה נרשום את הבעיה הפרימאלית:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad (P)$$

$$\text{s.t. } y_1(w\cdot x_1+b)\geq 1$$
$$y_2(w\cdot x_1+b)\geq 1$$

כעת את הבעיה הדואלית:

$$\max_{\alpha_1,\alpha_2} \alpha_1+\alpha_2-\frac{1}{2}\left(\alpha_1^2y_1^2\|x_1\|^2+\alpha_1\alpha_2y_1y_2x_1\cdot x_2+\alpha_2\alpha_1y_2y_1x_2\cdot x_1+\alpha_2^2y_2^2\|x_2\|^2\right)$$

$$\text{s.t. } \alpha_1y_1+\alpha_2y_2=0 \qquad (D)$$
$$\alpha_1,\alpha_2\geq 0$$

נבחין כי מכיוון ש $y_1=1$, $y_2=-1$ האילוץ הראשון בדואלית הוא $\alpha_1-\alpha_2=0$ ולכן $\alpha_1=\alpha_2$.

נסמן: $\alpha_1=\alpha_2=\alpha$ ונציב את פונ' המטרה של הדואלית:

$$2\alpha-\frac{1}{2}\left(\alpha^2\|x_1\|^2-\alpha^2x_1\cdot x_2-\alpha^2x_2\cdot x_1+\alpha^2\|x_2\|^2\right)\underset{\uparrow}{=}2\alpha-\frac{1}{2}\left(\alpha^2\|x_1\|^2-2\alpha^2x_1\cdot x_2+\alpha^2\|x_2\|^2\right)$$

מ"פ היא סימטרית.

$$=2\alpha-\frac{1}{2}\alpha^2\|x_1\|^2+\alpha^2x_1\cdot x_2-\frac{1}{2}\alpha^2\|x_2\|^2$$

כלומר קיבלנו כי הבעיה הדואלית שקולה לבעיה:

$$\max_{\alpha} \underbrace{2\alpha-\frac{1}{2}\alpha^2\|x_1\|^2+\alpha^2x_1\cdot x_2-\frac{1}{2}\alpha^2\|x_2\|^2}_{=:g(\alpha)}$$

$$\text{s.t. } \alpha\geq 0$$

נמקסם ע"י גזירה לפי α והשוואה ל-0:

$$\frac{\partial g(\alpha)}{\partial\alpha}=2-\alpha\|x_1\|^2+2\alpha x_1\cdot x_2-\alpha\|x_2\|^2=0 \iff 2=\alpha\left(\|x_1\|^2-2x_1\cdot x_2+\|x_2\|^2\right)=\alpha\cdot\|x_1-x_2\|^2$$

$$\Rightarrow \alpha=\frac{2}{\|x_1-x_2\|^2}$$

אכן $\alpha\geq 0$

$$\alpha^*=\frac{2}{\|x_1-x_2\|^2}$$

כעת נוכל לפתור את הבעיה הפרימאלית, כפי שראינו בהרצאה בהינתן שפתרנו את הדואלית הפתרון לפרימאלית יהיה:

$$w^*=\sum_{i=1}^{2}\alpha_i^*y_ix_i \quad , \quad b^*=y_i-w^*x_i$$

$$w^*=\alpha_1^*y_1\cdot x_1+\alpha_2^*y_2x_2=\alpha^*x_1-\alpha^*x_2=\frac{2x_1-2x_2}{\|x_1-x_2\|^2}=2\cdot\frac{x_1-x_2}{\|x_1-x_2\|^2}$$

$$w = \alpha_1 y_1 \cdot x_1 + \alpha_2 y_2 x_2 = \alpha' x_1 - \alpha x_2 = \frac{\underbrace{\alpha x_1 - \alpha x_2}}{\|x_1 - x_2\|^2} = 2 \cdot \frac{x_1 - x_2}{\|x_1 - x_2\|^2}$$

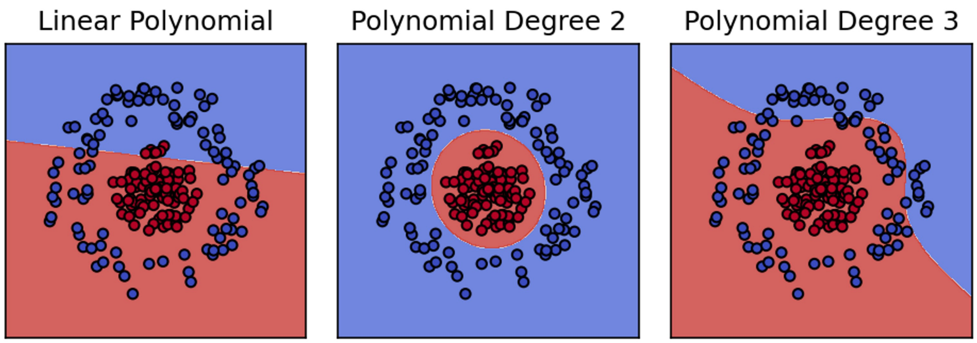
באשר ל b^* זה לא משנה אם נחשב לפי הנקודה הראשונה או השנייה כי אם $\alpha_i^* \neq 0$ אז (x_i, y_i) הוא "וקטור תומך" ולכל הרצ יש אותו margin ונקבל אותו b^*.

נחשב לפי (x_1, y_1):

$$b^* = y_1 - w^* \cdot x_1 = 1 - 2 \cdot \frac{x_1 - x_2}{\|x_1 - x_2\|^2} \cdot x_1$$

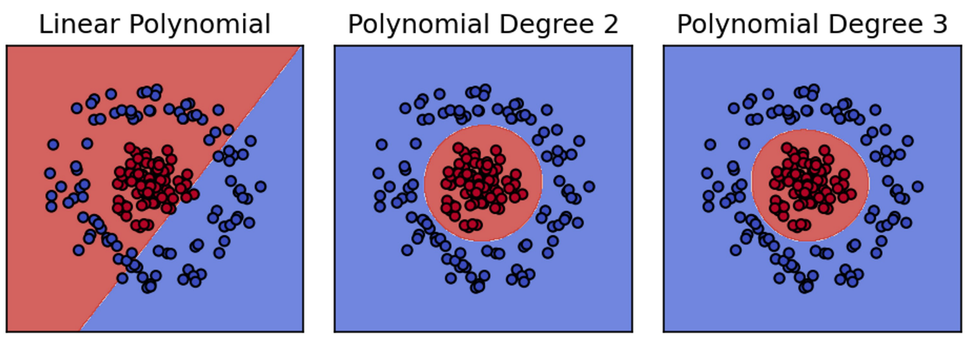
שאלה תכנות

(a) קרנל פולינומי הומוגני:



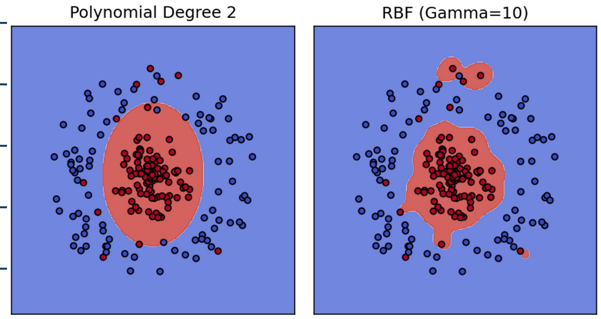
ניתן לראות בבירור שהקרנל הפולינומי מדרגה 2 מפריד את הדאטה הכי מדויק.
זו כמובן תוצאה הגיונית שכן הדאטה מסווג ע"י מעגל ולכן קרנל פולינומי מדרגה 2 ייצר את המישור המפריד הכי מתאים.

(b) קרנל פולינומי לא-הומוגני:



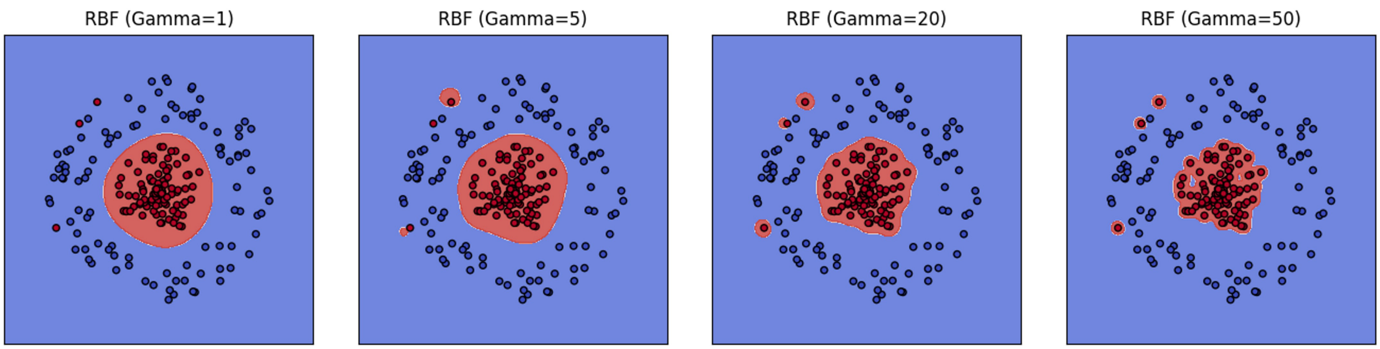
נראה כי הפולינומים מדרגה 1 ו-2 מתפקדים באופן דומה למקרה ההומוגני. אבל כעת הפולינום מדרגה 3 מפריד את הדאטה הרבה יותר טוב משום שבמקרה הלא הומוגני הגורם הקבוע שמתפקד כפרמטר חופשי מתקן את המישור המפריד ביחס למקרה ההומוגני.
עדיין, נראה שהפולי' מדרגה 2 מפריד הכי טוב את הדאטה באופן סימטרי.

(c) Poly vs RBF



ניתן לראות שבכאשר יש קשיים בדאטה ה RBF מפריד טוב יותר מהקרנל הפולינומי משום שהוא מסוגל לעבוד בממד גבוה ממימד הקרנל הפול' מדרגה 2. בנוסף, ניתן לראות שהמפריד בקרנל הפולינומי יותר אליפטי, כלומר מנסה "לסגור" את המעגל כדי לסווג עוד נקודה.

RBF with different Gamma values



ניתן לראות שככל שערכי Gamma גדולים יותר כך ההפרדה נעשית עדינה ומדויקת יותר על הדאטה, כלומר המודל נוטה יותר ל overfit על המדגם הנתון. בנוסף עוד שעבור ערכי γ קטנים (ובאופן שקול ערכי σ גדולים) יספקו מפריד יותר "כללי" כשעבור $\gamma=1$ למעשה אין הבדל גדול ביחס לקרנל הפולינומי.

ניתן לראות בדוגמאות אלו את עקרון ה: "bias variance tradeoff":
ככל שהמודל מתאים עצמו טוב יותר למדגם הנתון (overfit) כך השונות תהיה גבוהה בין מדגם למדגם כך שה bias יהיה קטן. לעומת זאת ככל שהמודל יהיה יותר "חלק" (underfit) כך השונות תקטן וה bias יגדל.