Machine learning- Ex4, Osher Elhadad 318969748

1. (20 pts) Logistic regression

(a) Prove that the logistic regression classifier is equivalent to a softmax over a linear multiclass classifier for two classes $y = \text{“}a\text{”}, y = \text{“}b\text{”}$, where their separating hyperplanes obey $\mathbf{w}_a = -\mathbf{w}_b$.

תשובה-

נוכיח שקילות של $logistic\ regression\ classifier$
ל$softmax\ over\ linear\ multiclass\ classifier$ ע״י הוכחת הטענה הבאה-

לכל קלט מתאים x עבור 2 ה$classes$ "a" ו"b" אזי מתקיים כי קיים סט משקולות כלשהו
w' שמייצג $logistic\ regression\ classifier$ עבורו מתקיים
$\hat{y}_{logistic\ regression\ classifier} =$ "a" אמ״מ קיימים 2 סטים של משקולות כלשהם w_a, w_b
שמייצגים $softmax\ over\ linear\ multiclass\ classifier$ כך ש-
$\hat{y}_{softmax\ over\ linear\ multiclass\ classifier} =$ "a" כאשר מתקיים
$their\ separating\ hyperplanes\ obey\ w_a = -w_b$

(מכיוון שאנו מבצעים $classification$ עבור 2 $classes$ לכן אם נוכיח טענה זו אזי זה גם יוכיח את הטענה השקולה עבור הערך "b" מכיוון שינם 2 ערכים- נתון $binary\ classifier$.

כיוון ראשון:

יהי קלט x עבור 2 ה$classes$ "a" ו"b" ועבור $logistic\ regression\ classifier$ בהינתן תווית ליניארית $y \in \{0,1\}$ (כאשר 1 מייצג את האות "a" ו0 את האות "b") וסט משקולות w' נניח כי
$\hat{y}_{logistic\ regression\ classifier} =$ "a" ונוכיח כי עבור 2 סטים של משקולות כלשהם w_a, w_b
שמייצגים $softmax\ over\ linear\ multiclass\ classifier$ מתקיים כי
$\hat{y}_{softmax\ over\ linear\ multiclass\ classifier} =$ "a" כאשר מתקיים
$the\ classes\ separating\ hyperplanes\ obey\ w_a = -w_b$

נשים לב כי ב$logistic\ regression\ classifier$ בהינתן תווית ליניארית $y \in \{0,1\}$ (כאשר 1 מייצג את האות "a" ו0 את האות "b") וסט משקולות w' אזי נקבל כי

$$p_{logistic\ regression}(y = 1|x) = \sigma(w'^T x) = \frac{1}{1 + e^{-w'^T x}} = \frac{e^{w'^T x}}{1 + e^{w'^T x}}$$

$$p_{logistic\ regression}(y = 0|x) = 1 - \sigma(w'^T x) = 1 - \frac{1}{1 + e^{-w'^T x}} = \frac{e^{-w'^T x}}{1 + e^{-w'^T x}}$$

ולכן למעשה ניתן להסתכל על הסתברות זו כפונקציה של w, נגדיר

$$f(w) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

ונקבל כי $p_{logistic\ regression}(y = 0|x) = f(-w')$ וגם $p_{logistic\ regression}(y = 1|x) =$
$f(w')$ מכיוון שנתון $\hat{y}_{logistic\ regression\ classifier} =$ "a" אזי לפי הגדרת
ה$logistic\ regression\ classifier$ מכיוון שה$classifier$ חזה כי התווית היא "a" לכן

למעשה ההסתברות לקבל אותו הייתה גדולה מההסתברות לקבל את "b" , כלומר -
$p_{logistic\ regression}(y=1|x)=f(w') > p_{logistic\ regression}(y=0|x)=f(-w')$

$$f(w')>f(-w')\rightarrow \frac{e^{w'^Tx}}{1+e^{w'^Tx}}>\frac{e^{-w'^Tx}}{1+e^{-w'^Tx}}\rightarrow e^{w'^Tx}\left(1+e^{-w'^Tx}\right)$$
$$>e^{-w'^Tx}\left(1+e^{w'^Tx}\right)$$
$$\rightarrow e^{w'^Tx}-e^{-w'^Tx}+e^{w'^Tx}e^{-w'^Tx}-e^{w'^Tx}e^{-w'^Tx}>0$$
$$\rightarrow e^{w'^Tx}-e^{-w'^Tx}+e^{0}-e^{0}>0\rightarrow e^{w'^Tx}>e^{-w'^Tx}$$
$$\rightarrow \frac{e^{w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}>\frac{e^{-w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}$$

כאשר המעבר האחרון מוצדק מכיוון שכפלנו את 2 הצדדים בכופל חיובי $\frac{1}{e^{-w'^Tx}+e^{w'^Tx}}$.

וקיבלנו כי לפי הגדרת $softmax\ over\ linear\ multiclass\ classifier$ נקבל כי מכיוון שישנם 2 $classes$ - "a" ו"b" (בהתאמה לפי הסדר ל1 ו0) בעצם מכיוון שהראנו כי מתקיימים- $p_{logistic\ regression}(y=1|x)=f(w')$ וגם $p_{logistic\ regression}(y=0|x)=f(-w')$ לכן למעשה עבור w' נקבל את ההסתברות של ה$logistic\ regression\ classifier$ שיחזה את הערך "a" (ששקול ל1 כפי שהגדרנו) ועבור $-w'$ נקבל את ההסתברות של ה$logistic\ regression\ classifier$ שיחזה את הערך "b" (ששקול ל0 כפי שהגדרנו) בדיוק כמו שמוגדר $softmax\ over\ linear\ multiclass\ classifier$ ונגדיר אותו כך- נסמן כי $w_a=w'$ ו $w_b=-w'$ (אכן מתקיים הנתון בשאלה- $w_a=-w_b$) שהם וקטורי המשקולות שלאחר $softmax$ על התוצאה של מכפלת כל אחד מהם בנפרד (w_a^T ו w_b^T) בקלט x נקבל למעשה כי $p_{Soft\ max}(y=1|x)=\frac{e^{w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}$ $p_{Soft\ max}(y=0|x)=\frac{e^{-w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}$ ומכיוון ש$\frac{e^{-w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}>\frac{e^{w'^Tx}}{e^{-w'^Tx}+e^{w'^Tx}}$ לכן $p_{Soft\ max}(y=1|x)>p_{Soft\ max}(y=0|x)$ ולכן ה$softmax\ over\ linear\ multiclass\ classifier$ יחזה את הערך "a" (ששקול ל1), כלומר $\hat{y}_{softmax\ over\ linear\ multiclass\ classifier}=$ "a" כאשר מתקיים $the\ classes\ separating\ hyperplanes\ obey\ w_a=-w_b$ כדרוש.

כיוון שני:

יהי קלט x עבור 2 ה$classes$ "a" ו"b" ועבור $softmax\ over\ linear\ multiclass\ classifier$ בהינתן תווית ליניארית $y\in\{0,1\}$ (כאשר 1 מייצג את האות "a" ו0 את האות "b") ו2 סטים של משקולות w_a,w_b כך ש$\mathrm{w_a}=-w_b$ נניח כי $\hat{y}_{softmax\ over\ linear\ multiclass\ classifier}=$ "a" ונוכיח כי עבור סט של משקולות w' כלשהו שמייצג $logistic\ regression\ classifier$ מתקיים כי $\hat{y}_{logistic\ regression\ classifier}=$ "a" .

יהי קלט x עבור 2 ה$classes$ "a" ו"b" נניח כי $\hat{y}_{softmax\ over\ linear\ multiclass\ classifier}=$ "a" כאשר מתקיים $the\ classes\ separating\ hyperplanes\ obey\ w_a=-w_b$ ונוכיח כי $\hat{y}_{logistic\ regression\ classifier}=$ "a"

נשים לב כי ב$softmax\ over\ linear\ multiclass\ classifier$ בהינתן תווית ליניארית $y\in\{0,1\}$ (כאשר 1 מייצג את האות "a" ו0 את האות "b") וסט משקולות w' אזי נקבל כי

$$p_{Soft\ max}(y=1|x)=\frac{e^{w_a^Tx}}{e^{w_a^Tx}+e^{w_b^Tx}}=\frac{e^{w_a^Tx}}{e^{w_a^Tx}+e^{-w_a^Tx}}$$

$$p_{Soft\,max}(y=0|x)=\frac{e^{w_b^Tx}}{e^{w_a^Tx}+e^{w_b^Tx}}=\frac{e^{-w_a^Tx}}{e^{w_a^Tx}+e^{-w_a^Tx}}$$

לכן לפי הגדרת ה $softmax\ over\ linear\ multiclass\ classifier$ מכיוון שה$classifier$ חזה כי התווית היא "a" לכן למעשה ההסתברות לקבל אותו הייתה גדולה מההסתברות לקבל את "b" , כלומר- $p_{Soft\,max}(y=1|x)\ >p_{Soft\,max}(y=0|x)$ ולכן

$$\frac{e^{w_a^Tx}}{e^{w_a^Tx}+e^{-w_a^Tx}}>\frac{e^{-w_a^Tx}}{e^{w_a^Tx}+e^{-w_a^Tx}}\xrightarrow{e^{w_a^Tx}+e^{-w_a^Tx}\geq 0}e^{w_a^Tx}>e^{-w_a^Tx}$$
$$\rightarrow e^{w_a^Tx}-e^{-w_a^Tx}+e^0-e^0>0$$
$$\rightarrow e^{w_a^Tx}-e^{-w_a^Tx}+e^{w_a^Tx}e^{-w_a^Tx}-e^{w_a^Tx}e^{-w_a^Tx}>0$$
$$\rightarrow e^{w_a^Tx}\left(1+e^{-w_a^Tx}\right)>e^{-w_a^Tx}\left(1+e^{w_a^Tx}\right)\rightarrow\frac{e^{w_a^Tx}}{1+e^{w_a^Tx}}>\frac{e^{-w_a^Tx}}{1+e^{-w_a^Tx}}$$

כעת נביט ב$logistic\ regression\ classifier$ עם תווית ליניארית $y\in\{0,1\}$ (כאשר 1 מייצג את האות "a" ו0 את האות "b") וסט המשקולות w_a אזי נקבל כי

$$p_{logistic\ regression}(y=1|x)=\sigma(w_a^Tx)=\frac{1}{1+e^{-w_a^Tx}}=\frac{e^{w_a^Tx}}{1+e^{w_a^Tx}}$$

$$p_{logistic\ regression}(y=0|x)=1-\sigma(w_a^Tx)=1-\frac{1}{1+e^{-w_a^Tx}}=\frac{e^{-w_a^Tx}}{1+e^{-w_a^Tx}}$$

ומכיוון ש$\frac{e^{w_a^Tx}}{1+e^{w_a^Tx}}>\frac{e^{-w_a^Tx}}{1+e^{-w_a^Tx}}$ לכן

$$p_{logistic\ regression}(y=1|x)>p_{logistic\ regression}(y=0|x)$$

ולכן ה$logistic\ regression\ classifier$ יחזה את הערך "a" (ששקול ל1), כלומר $\hat{y}_{logistic\ regression\ classifier}=$ "a" כדרוש.

(b) Consider a softmax function for k classes that is scaled by a constant $T \in R$

$$f_i(\mathbf{x}) = \frac{\exp(\frac{1}{T}\mathbf{w_i}^T\mathbf{x})}{\sum_k \exp(\frac{1}{T}\mathbf{w_k}^T\mathbf{x})} \quad (1)$$

Show that when $T \to 0$, the softmax converges to the max function. What happens when $T \to \infty$?

תשובה-

נראה כי עבור $T \in R$ כך ש $T \to 0$ אזי $softmax\ scaled\ by\ T$ מתכנס ל$\max function$, כלומר $f_i(x)$ (כאשר $1 \le i \le k$) מתכנס ל1 עבור האינדקס $i = j_1 \vee \dots \vee i = j_n$ כך ש $1 \le j_h \le l$ לכל $j_h \in \{j_1, \dots j_n\}$ עבורו מתקיים כי $w_{j_h}^T x > w_r^T x\ \forall 1 \le r \le k \wedge r \notin \{j_1, \dots j_n\}$ כלומר לכל $j_h \in \{j_1, \dots j_n\}$ הערך $w_{j_h}^T x$ מקבל ערך מקסימלי ($w_{j_1}^T x = \cdots = w_{j_n}^T x$), אחרת $f_i(x)$ מתכנס ל0 (עבור כל שאר האינדקסים $\forall 1 \le r \le k: r \notin \{j_1, \dots j_n\}$) ולכן למעשה נקבל שערכי ה$argmax_i(w_i^T x)$ יקבלו ערך 1 ושאר האינדקסים עם ערך 0- כלומר פונקציית הmax שרצינו.

הפונקציה $softmax\ scaled\ by\ T$ הינה מהצורה-

$$f_i(x) = \frac{e^{\frac{1}{T}w_i^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x}}$$

לכל $i = j_h \in \{j_1, \dots j_n\}$ נקבל:

$$\lim_{T\to0} f_{j_h}(x) = \lim_{T\to0} \frac{e^{\frac{1}{T}w_{j_h}^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x}} = \lim_{T\to0} \frac{e^{\frac{1}{T}w_{j_h}^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x}} \cdot \frac{e^{-\frac{1}{T}w_{j_h}^T x}}{e^{-\frac{1}{T}w_{j_h}^T x}} = \lim_{T\to0} \frac{e^{\frac{1}{T}w_{j_h}^T x - \frac{1}{T}w_{j_h}^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x - \frac{1}{T}w_{j_h}^T x}}$$
$$= \lim_{T\to0} \frac{1}{e^{\frac{1}{T}w_{j_h}^T x - \frac{1}{T}w_{j_h}^T x} + \sum_{k' \ne j_h} e^{\frac{1}{T}w_{k'}^T x - \frac{1}{T}w_{j_h}^T x}}$$
$$= \lim_{T\to0} \frac{1}{1 + \sum_{k' \ne j} e^{\frac{1}{T}\left(w_{k'}^T x - w_{j_h}^T x\right)}} = 1$$

המעבר האחרון מוצדק לפי המעברים הבאים-

$$T \to 0 \Longrightarrow \frac{1}{T} \to \infty$$

$$\forall k' \notin \{j_1, \dots j_n\}: w_{j_h}^T x > w_{k'}^T x \Longrightarrow w_{k'}^T x - w_{j_h}^T x < 0$$

ונקבל-

$$\frac{1}{T}\left(w_{k'}^T x - w_{j_h}^T x\right) \to -\infty \Longrightarrow e^{\frac{1}{T}\left(w_{k'}^T x - w_{j_h}^T x\right)} \to 0 \Longrightarrow \frac{1}{1 + \sum_{k' \ne j} e^{\frac{1}{T}\left(w_{k'}^T x - w_{j_h}^T x\right)}} \to 1$$

לכל $i \notin \{j_1, \dots j_n\}$ נקבל:

$$\lim_{T\to 0} f_i(x) = \lim_{T\to 0}\frac{e^{\frac{1}{T}w_i^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x}} = \lim_{T\to 0}\frac{e^{\frac{1}{T}w_i^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x}}\cdot\frac{e^{-\frac{1}{T}w_i^T x}}{e^{-\frac{1}{T}w_i^T x}} = \lim_{T\to 0}\frac{e^{\frac{1}{T}w_i^T x-\frac{1}{T}w_i^T x}}{\sum_{k'} e^{\frac{1}{T}w_{k'}^T x-\frac{1}{T}w_i^T x}}$$
$$= \lim_{T\to 0}\frac{1}{\sum_{k_1} e^{\frac{1}{T}\left(w_{k_1}^T x-w_i^T x\right)}+\sum_{k_2} e^{\frac{1}{T}\left(w_{k_2}^T x-w_i^T x\right)}+\sum_{k_3} e^{\frac{1}{T}\left(w_{k_3}^T x-w_i^T x\right)}}$$

ונגדיר כל $k_1 \in \{1,\dots,k\}$ כך שמתקיים $w_{k_1}^T x - w_i^T x = 0$, בנוסף כל $k_2 \in \{1,\dots,k\}$ כך שמתקיים $w_{k_2}^T x - w_i^T x < 0$, וכל $k_3 \in \{1,\dots,k\}$ כך שמתקיים $w_{k_3}^T x - w_i^T x > 0$. ונקבל-

$$\lim_{T\to 0}\frac{1}{\sum_{k_1} e^{\frac{1}{T}\left(w_{k_1}^T x-w_i^T x\right)}+\sum_{k_2} e^{\frac{1}{T}\left(w_{k_2}^T x-w_i^T x\right)}+\sum_{k_3} e^{\frac{1}{T}\left(w_{k_3}^T x-w_i^T x\right)}}$$
$$= \lim_{T\to 0}\frac{1}{\sum_{k_1} e^{\frac{1}{T}\cdot 0}+\sum_{k_2} e^{\frac{1}{T}\left(w_{k_2}^T x-w_i^T x\right)}+\sum_{k_3} e^{\frac{1}{T}\left(w_{k_3}^T x-w_i^T x\right)}}\left(= "\frac{1}{1+0+\infty}"\right) = 0$$

המעבר האחרון מוצדק לפי המעברים הבאים-

$$T \to 0 \Longrightarrow \frac{1}{T} \to \infty$$
$$\forall k_2 \in \{1,\dots,k\}: w_{k_2}^T x - w_i^T x < 0$$

ונקבל-

$$\frac{1}{T}\left(w_{k_2}^T x - w_i^T x\right) \to -\infty \Longrightarrow e^{\frac{1}{T}\left(w_{k_2}^T x-w_i^T x\right)} \to 0$$
$$\forall k_3 \in \{1,\dots,k\}: w_{k_3}^T x - w_i^T x > 0$$

ונקבל-

$$\frac{1}{T}\left(w_{k_3}^T x - w_i^T x\right) \to \infty \Longrightarrow e^{\frac{1}{T}\left(w_{k_3}^T x-w_i^T x\right)} \to \infty$$

כלומר $f_i(x)$ (כאשר $1 \le i \le k$) מתכנס ל1 עבור האינדקס $i = j_1 \vee \dots \vee i = j_n$ כך ש $1 \le j_h \le k$ לכל $j_h \in \{j_1,\dots j_n\}$ עבורו מתקיים כי $w_{j_h}^T x > w_r^T x\ \forall 1 \le r \le k \wedge r \notin \{j_1,\dots j_n\}$ כלומר לכל $j_h \in \{j_1,\dots j_n\}$ הערך $w_{j_h}^T x$ מקבל ערך מקסימלי ($w_{j_1}^T x = \cdots = w_{j_n}^T x$), אחרת $f_i(x)$ מתכנס ל0 (עבור כל שאר האינדקסים $\forall 1 \le r \le k: r \notin \{j_1,\dots j_n\}$) כדרוש.

כמו כן עבור $T \to \infty$ נקבל:

$$\lim_{T\to 0} f_i(x) = \lim_{T\to 0}\frac{e^{\frac{1}{T}w_i^T x}}{\sum_k e^{\frac{1}{T}w_k^T x}} = \frac{e^0}{\sum_k e^0} = \frac{1}{k}$$

ונקבל במקרה זה כי $f_i(x)$ מתכנסת ל-$\frac{1}{k}$ לכל i (כלומר למספר קבוע k- מספר ה$classes$). כלומר זו תהיה פונקציה קבועה לכל i וx שנקבל- מאבדים פה את כל ההבדלים בין כל w_i שאנו חיפשנו מלכתחילה ב$softmax$ כדי לקבל הסתברויות או מקסימום כמו שהצגנו פה למעלה.

(c) Write the gradient update rule for a logisitic regression model, when the usual loss of the negative log likelihood is now regularized with the square of the L_2 norm over the weight vector $\frac{1}{2}||\mathbf{w}||^2$.

תשובה-

נראה את חוק העדכון של מודל ה$logistic\ regression$ כאשר פונקציית ה$loss$ הינה $negative \log likelihood$ בתוספת רגולריזציה של $square\ of\ L2\ norm\ over\ weight\ vector\ \frac{1}{2}||w||^2$.

כלומר נקבל כי פונקציית ה$loss$ היא $Err_{new}(w) = -\log(likelihood) + \frac{1}{2}||w||^2$

$$Err_{new}(w) = -\log(likelihood) + \frac{1}{2}||w||^2$$

נזכיר מההרצאה כי-

$$likelihood = \prod_{y_i=1} p(y=1|x_i) \prod_{y_i=0} p(y=0|x_i) \overset{\forall z \in \mathbb{R} | z^0 = 1}{=}$$

$$= \prod_{i\in\{1,\dots,n\}} p(y=1|x_i)^{y_i} \prod_{i\in\{1,\dots,n\}} p(y=0|x_i)^{1-y_i}$$

$$\overset{p(y=1|x_i)=\sigma(w^Tx_i) \wedge p(y=0|x_i) = 1-\sigma(w^Tx_i)}{=}$$

$$= \prod_{i\in\{1,\dots,n\}} \sigma(w^Tx_i)^{y_i} \prod_{i\in\{1,\dots,n\}} \left(1-\sigma(w^Tx_i)\right)^{1-y_i}$$

ולכן נקבל-

$$-\log(likelihood) =$$

$$-\sum_{i\in\{1,\dots,n\}} y_i \log\left(\sigma(w^Tx_i)\right) - \sum_{i\in\{1,\dots,n\}} (1-y_i)\log\left(1-\sigma(w^Tx_i)\right)$$

כעת כדי לחשב את הגרדיאנט נשתמש בכללים אלו-

$$\frac{d\sigma(z)}{dz} = \sigma(z)\left(1-\sigma(z)\right)$$

ולפי כלל השרשרת-

$$\frac{d\log(likelihood)}{dw} = \frac{d\log\left(\sigma(w^Tx_i)\right)}{d\sigma(w^Tx_i)} \frac{d\sigma(w^Tx_i)}{dw^Tx_i} \frac{dw^Tx_i}{dw}$$

בנוסף כפי שראינו בהרצאה

$$\frac{d\frac{1}{2}||w||^2}{dw} = w$$

ולכן כעת נמצא את הגרדיאנט של Err_{new} נסמן את $\sigma(w^Tx_i)$ בקיצור σ_1-

$$\nabla Err_{new}(w) = w - \sum_{i\in\{1,\dots,n\}} y_i \cdot \frac{1}{\sigma_1} \cdot \sigma_1 \cdot (1-\sigma_1) \cdot x_i + (1-y_i) \cdot \frac{-1}{1-\sigma_1} \cdot \sigma_1 \cdot (1-\sigma_1) \cdot x_i$$

$$= w - \sum_{i\in\{1,\dots,n\}} y_i \cdot (1-\sigma_1) \cdot x_i - \left(1-y_i\right) \cdot \sigma_1 \cdot x_i$$

$$= w - \sum_{i\in\{1,\dots,n\}} x_i (y_i \cdot (1-\sigma_1) - (1-y_i) \cdot \sigma_1)$$

$$= w - \sum_{i\in\{1,\dots,n\}} x_i (y_i - y_i \cdot \sigma_1 - \sigma_1 + y_i \cdot \sigma_1) = w - \sum_{i\in\{1,\dots,n\}} x_i \left(y_i - \sigma\left(w^T x_i\right)\right)$$

$$\nabla Err_{new}(w) = \sum_{i\in\{1,\dots,n\}} \frac{1}{n} w - x_i\left(y_i - \hat{y}_i\right)$$

כלומר הגרדיאנט עבור דגימה i הוא $\frac{1}{n} w - x_i\left(y_i - \hat{y}_i\right)$.

כעת נגדיר את כלל העדכון של w (בדומה להרצאה לפי הדגימה הi)-

$$w^{t+1} = w^t - \eta\left(\frac{1}{n} w^t - x_i\left(y_i - \hat{y}_i\right)\right) = w^t - \frac{\eta}{n} w^t + \eta x_i\left(y_i - \hat{y}_i\right)$$

וכלל העדכון של w לפי n דגימות יהיה-

$$w^{t+1} = w^t - \eta\left(w^t - \sum_{i\in\{1,\dots,n\}} x_i\left(y_i - \hat{y}_i\right)\right) = w^t - \eta w^t + \eta \sum_{i\in\{1,\dots,n\}} x_i\left(y_i - \hat{y}_i\right)$$

2. (10 pts) Support Vector Machines

At class we discussed the SVM formulation

$$\min_{\mathbf{w},\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \qquad (2)$$

where C is a trade-off hyper parameter that weighs the cost of misclassification o(thrugh the slack variable). You are given a binary classification problem for which the cost of misclassifying positive samples is different than the cost of misclassifying negative examples. Formulate the problem as an SVM with two types of slack variables. How many variables are now optimized over in your optimization problem, and what are their dimensions?

<u>תשובה-</u>

נמדל את בעיית האופטימיזציה ש- SVM פותר כאשר אנחנו מתבוננים בשני סוגי slack variables (ξ_i). הראשון הינו עבור דגימות חיוביות שמסווגות לא נכון (כשליליות) או שמפספסות את ה$margin$ (כפי שהוצג בהרצאה). והשני הינו עבור דגימות שליליות שמסווגות לא נכון (כחיוביות) או שמפספסות את ה$margin$ (כפי שהוצג בהרצאה).

הבעיה המקורית מההרצאה הוגדרה כך- $\min_{w\in R^d,\xi_i} \frac{1}{2}\left||w|\right|^2 + C\cdot \Sigma_{i=1}^{n}\xi_i$

כאשר $w^T x_i y_i \geq 1-\xi_i$ וגם $\forall i: \xi_i \geq 0$.

למעשה הווקטורים w שיבחרו יקיימו שהמרחק ל$support\ vectors$ הוא 1 ולכן למעשה החלק השמאלי של הרגולציה בעצם דואג לכך שהנורמה של הווקטור w תהיה קטנה, עם ערכים קטנים לw ושעדיין המרחק מה$support\ vectors$ הוא 1. כמו כן המשתנה ξ_i עבור כל דגימה, והוגדר בהרצאה שמתקיים $\xi_i = \max(0,1-w^T x_i y_i)$ כלומר אם הדגימה אכן סווגה נכונה וגם נמצאת מעל הטווח שהגדרנו (מעבר ל$support\ vector$ שבצד של הסיווג הנכון לדגימה זו) אזי ξ_i יהיה למעשה 0- כלומר לא יהיה עונש, אחרת יקבע ע״י המרחק מה$support\ vector$ (שבצד של הסיווג הנכון לדגימה זו) כעונש לנקודה זו שלא נמצאת בטווח שרצינו. כלומר נקבל $hinge\ loss$ בחלק הימני של הביטוי. כעת נשים לב כי C הוא למעשה היפר פרמטר שכאשר הוא גדול יותר הוא נותן יותר חשיבות לעונש שהגדרנו לנקודות שלא בטווח הנכון (מעבר ל$support\ vector$ בכיוון הסיווג הנכון של הדגימה) ופחות לנורמה של w , כלומר יאפשר w יותר גדולים. לעומת זאת כאשר C קטן יותר נקבל כי אנחנו נותנים בעיקר חשיבות לנורמה של w ופחות לטעויות של ξ_i.

כעת נשים לב כי אנו משנים את הבעיה במעט- העונש על הטעות יהיה שונה לדגימות שליליות וחיוביות, כלומר נכפיל בערך C שונה לדגימות החיוביות והשליליות. נחלק את הדגימות כך-

לכל דגימה x_i^+ חיובית נסמן את הslack variable המתאים עם ξ_i^+ ואת ההיפר פרמטר C עם C^+, ולכל דגימה x_i^- שלילית נסמן את ה-slack variable המתאים עם ξ_i^- ואת ההיפר פרמטר C עם C^- (כך ש$C^+ \neq C^-$).

כעת נגדיר את הבעיה המתאימה לתשובה כך-

$$\min_{w\in R^d,\xi_i^+,\xi_i^-} \frac{1}{2}\left||w|\right|^2 + C^+\cdot\sum_{x_i^-}\xi_i^+ + C^-\cdot\sum_{x_i^+}\xi_i^-$$

כאשר $w^T x_i y_i \geq 1-\xi_i^-$, בנוסף $\forall i: \xi_i^- \geq 0$, כמו כן $w^T x_i y_i \geq 1-\xi_i^+$ וגם $\forall i: \xi_i^+ \geq 0$

בבעיה המקורית בהרצאה נאמר כי עקב הוספת ξ הפרמטרים שאנו לומדים הפכו להיות w, ξ אך מכיוון ש$\xi_i = \max(0,1 - w^T x_i y_i)$ לכן למעשה אנו לומדים את w ויש לנו אילוצים על ξ (ויש לנו n משתנים כאלו- כלומר מימד n) ומתוך זה גם נקבל את הξ_i (שהוא תלוי בw) שאנו מחפשים (את C אנו לא לומדים- היפר פרמטר). זאת בדיוק כמו בבעיה החדשה שעליה אנו מסתכלים שגם בה אנו מאפטמים את הw שמימדו הוא n ויש לנו אילוצים על ξ (ויש לנו n משתנים כאלו- כלומר מימד n) ומתוך כך גם נקבל את הξ_i (שהוא תלוי בw) שאנו מחפשים (גם פה אנו לא לומדים את ההיפר פרמטרים C^+, C^-).

3. (10 pts) Polynomial kernels

Given a pair of samples in input space $\mathbf{x}, \mathbf{z} \in R^2$, show that the polynomial kernel of degree 3, $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T\mathbf{z})^3$ can be expressed as a dot product in feature space $\phi(\mathbf{x})^T\phi(\mathbf{z})$ with the right choice of a feature map $\phi : R^2 \to R^d$ $(d > 2)$. Write the explicit form of ϕ.

How many operations are needed to compute $K(\mathbf{x}, \mathbf{z})$, and how many are needed for computing $\phi(\mathbf{x})\phi(\mathbf{z})$? How does the number of operation changes for polynomial kernel of higher degrees?

תשובה-

נסמן- $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$

למעשה נמצא את הפונקציה ϕ ע"י פתיחה של הביטוי $(1 + x^T z)^3$ עד הסוף ולאחר מכן בעצם כל מחובר שיהיה בביטוי הסופי יהיה בעצם מכפלה של 2 ביטויים בווקטורים $\phi(x), \phi(z)$ באותה השורה (כאשר אלו וקטורי עמודה) וכך למעשה נמצא כל ביטוי בכל שורה בווקטור שהפונקציה ϕ מחזירה.

$$1 + 3x_1z_1 + 3x_2z_2 + 3x_1^2z_1^2 + 6x_1z_1x_2z_2 + 3x_2^2z_2^2 + x_1^3z_1^3 + 3x_1^2z_1^2x_2z_2 + 3x_1z_1x_2^2z_2^2 + x_2^3z_2^3 =$$

$$1 + 3(x_1z_1 + x_2z_2) + 3(x_1z_1 + x_2z_2)^2 + (x_1z_1 + x_2z_2)^3 =$$

$$1 + 3x^Tz + 3(x^Tz)^2 + (x^Tz)^3 =$$

$$(1 + x^Tz)^3$$

$$\phi(x)^T \cdot \phi(z) =$$

$$\begin{pmatrix} x_2^3 & \sqrt{3}\cdot x_1x_2^2 & \sqrt{3}\cdot x_1^2x_2 & x_1^3 & \sqrt{3}\cdot x_2^2 & \sqrt{6}\cdot x_1x_2 & \sqrt{3}\cdot x_1^2 & \sqrt{3}\cdot x_2 & \sqrt{3}\cdot x_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} z_2^3 \\ \sqrt{3}\cdot z_1z_2^2 \\ \sqrt{3}\cdot z_1^2z_2 \\ z_1^3 \\ \sqrt{3}\cdot z_2^2 \\ \sqrt{6}\cdot z_1z_2 \\ \sqrt{3}\cdot z_1^2 \\ \sqrt{3}\cdot z_2 \\ \sqrt{3}\cdot z_1 \\ 1 \end{pmatrix} =$$

$$= 1 + 3x_1z_1 + 3x_2z_2 + 3x_1^2z_1^2 + 6x_1z_1x_2z_2 + 3x_2^2z_2^2 + x_1^3z_1^3 + 3x_1^2z_1^2x_2z_2 + 3x_1z_1x_2^2z_2^2 + x_2^3z_2^3$$

לכן, נגדיר את הפונקציה $\phi: \mathbb{R}^2 \to \mathbb{R}^{10}$ לכל v, מתקיים

$$\phi(v) = \phi\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_2^3 \\ \sqrt{3}\cdot v_1 v_2^2 \\ \sqrt{3}\cdot v_1^2 v_2 \\ v_1^3 \\ \sqrt{3}\cdot v_2^2 \\ \sqrt{6}\cdot v_1 v_2 \\ \sqrt{3}\cdot v_1^2 \\ \sqrt{3}\cdot v_2 \\ \sqrt{3}\cdot v_1 \\ 1 \end{pmatrix}$$

כעת נגדיר פעולה אחת ככפל סקלרי או חיבור סקלרי, ולכן חזקת i צורכת i פעולות.

נחשב כמה פעולות חישוב אנו צריכים לבצע על מנת לחשב את $\phi(x)^T \cdot \phi(z)$ ישירות-

תחילה נשים לב שעל מנת להגיע מכפל הווקטורים לביטוי הבא-

$$1 + 3x_1z_1 + 3x_2z_2 + 3x_1^2z_1^2 + 6x_1z_1x_2z_2 + 3x_2^2z_2^2 + x_1^3z_1^3 + 3x_1^2z_1^2x_2z_2 + 3x_1z_1x_2^2z_2^2 + x_2^3z_2^3$$

למעשה עלינו לבצע הכפלה של כל המספרים $\sqrt{3}, \sqrt{6}, 1$ בין האיברים המתאימים בכפל הוקטורי כלומר יש לנו 8 הכפלות סקלרים (8 פעולות כפי שהגדרנו) כאלו שיביאו לנו את האיברים הבאים בתוך הביטוי הגדול שלנו (שביניהם יהיה חיבור בתוספת האיברים שלא היה להם מספרים ממשיים שכפלנו בהם)-

$$1, 3x_1z_1, 3x_2z_2, 3x_1^2z_1^2, 6x_1z_1x_2z_2, 3x_2^2z_2^2, 3x_1^2z_1^2x_2z_2, 3x_1z_1x_2^2z_2^2$$

$3x_1z_1$ – צורך 2 פעולות כפל, $3x_2z_2$- צורך 2 פעולות כפל, $3x_1^2z_1^2$- צורך 4 פעולות כפל, $6x_1z_1x_2z_2$- צורך 4 פעולות כפל, $3x_2^2z_2^2$- צורך 4 פעולות כפל, $x_1^3z_1^3$- צורך 5 פעולות כפל, $3x_1^2z_1^2x_2z_2$- צורך 6 פעולות כפל, $3x_1z_1x_2^2z_2^2$- צורך 6 פעולות כפל, $x_2^3z_2^3$- צורך 5 פעולות כפל

סה״כ הגענו ל-10 מחוברים, כלומר 9 פעולות חיבור. נסכום הכל:

$$2 + 2 + 4 + 4 + 4 + 5 + 6 + 6 + 5 + 9 + 8 = 55$$

נחשב כמה פעולות חישוב אנו צריכים לבצע על מנת לחשב את הקרנל-

חישוב x^Tz יעלה לנו 2 פעולות כפל ופעולת חיבור לכפל הווקטורי בין 2 ווקטורים ממימד 2. כמו כן הוספת 1 (עוד פעולה אחת), והעלאה בשלישית מצריכה 2 פעולות כפל. סה״כ ביצענו 6 פעולות.

נחשב כמה פעולות חישוב אנו צריכים לבצע אם d היה גדול יותר-

אם המימד d היה גדול יותר, הקרנל היה מהצורה $(1 + x^Tz)^d$, ולכן בהנחה ש-$x, z \in R^2$, בהתאמה היינו נדרשים לבצע $d + 4$ פעולות. כלומר כמות הפעולות שהקרנל מצריך היא לינארית למימד של ϕ, בניגוד לחישוב ישיר שעולה לנו הרבה יותר פעולות כפי שראינו עבור $d = 3$ (מכיוון שנקבל יותר מחוברים בסוף ההעלאה בחזקה כפי שעשינו בתחילת התרגיל).