# מבוא למערכות לומדות – 67577 – תרגיל 4

## אופק אבידן – 318879574

**1. חלק תיאורטי:**

**Convex Optimization 1.1**

.1

Based on Lecture 9 and Recitations 2,11

Here we will see a nice property that will help see some property of convexity

1. Let $f_1, \ldots, f_m : C \to \mathbb{R}$ be a set of convex functions and $\gamma_1, \ldots, \gamma_m \in \mathbb{R}_+$. Prove from definition that $g(\mathbf{u}) = \sum_{i=1}^{m} \gamma_i f_i(\mathbf{u})$ is a convex function.

מההגדרה, צריך להראות כי לכל $u, v \in C$, ולכל $\alpha \in [0,1]$,

$$g(\alpha v + (1-\alpha)u) \leq \alpha g(v) + (1-\alpha) g(u)$$

יהיו $\alpha \in [0,1], u, v \in C$.

כעת:

$$\begin{aligned} &g(\alpha v + (1-\alpha)u) \\ &\quad = \Sigma_i \gamma_i f_i(\alpha v + (1-\alpha)u) \leq_{\text{מהקמירות של } f} \Sigma_i \gamma_i \alpha f_i(v) + \gamma_i (1-\alpha) f_i(u) \\ &\quad = \alpha g(v) + (1-\alpha) g(u) \end{aligned}$$

כנדרש.

.2

2. Give a counterexample for the following claim: Given two functions $f, g : \mathbb{R} \to \mathbb{R}$, define a new function $h : \mathbb{R} \to \mathbb{R}$ by $h = f \circ g$. If $f$ and $g$ are convex then $h$ is convex as well.

נגדיר $f(x) = -x$ , $g(x) = x^2$.

$f(x)$ קמורה:

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= -\alpha x - (1-\alpha)y = \alpha(-x) + (1-\alpha)(-y) \\ &= \alpha f(x) + (1-\alpha) f(y). \end{aligned}$$

קמירות של $g(x)$:

$$\begin{aligned} g(\alpha x + (1-\alpha)y) &= (\alpha x + (1-\alpha)y)^2 = \alpha^2 x^2 + 2\alpha x(1-\alpha)y + (1-\alpha)^2 y^2 \\ &\leq \alpha x^2 + (1-\alpha)y^2 = \alpha g(x) + (1-\alpha) g(y) \end{aligned}$$

אבל $h(x) = f\big(g(x)\big) = -x^2$ איננה קמורה:

עבור $\alpha = \frac{1}{2}, u = -1, v = 1$

$$h(\alpha v + (1-\alpha)u) = h\left(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot(-1)\right) = h(0) = 0 > -1 = \frac{1}{2}\cdot(-1) + \frac{1}{2}(-1)$$
$$= \frac{1}{2}h(1) + \left(1 - \frac{1}{2}\right)h(-1) = \alpha h(v) + (1-\alpha)h(u)$$

## Sub-gradients for Soft-SVM Objective 1.2

.3

### 1.2 Sub-gradients for Soft-SVM Objective

Based on Lecture 9 and Recitations 2,11

The Soft-SVM objective, though convex, is not differentiable in all of its domain due to the use of the hinge-loss. Therefore, to implement a sub-gradient descent solver for this problem we must first describe sub-gradients of the objective.

3. Given $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$. Show that the hinge loss is convex in $\mathbf{w}, b$. That is, define

$$f(\mathbf{w}, b) := \ell^{hinge}_{\mathbf{x},y}(\mathbf{w}, b) = \max\left(0, 1 - y(\mathbf{x}^\top \mathbf{w} + b)\right)$$

and show that $f$ is convex in $\mathbf{w}, b$.

נגדיר

$$f_1(w,b) = 0$$

$$f_2(w,b) = 1 - y(x^T w + b)$$

נשים לב כי:

$$f(w,b) = \ell^{hinge}_{x,y} = \max\big(f_1(w,b), f_2(w,b)\big)$$

כעת,

1. $f_1$ היא קמורה כי פונקציית ה-0 קמורה
2. $f_2$ היא קמורה כי ראינו שפונקציות אפיניות הן קמורות

בנוסף מתקיים:

$$f(w,b) = \max\big(f_1(w,b), f_2(w,b)\big) = \sup\big(f_1(w,b), f_2(w,b)\big)$$

ולכן לפי תרגול 2 של איתן, $f$ קמורה גם היא.

.4

4. Deduce some sub-gradient of the hinge loss function $g \in \partial \ell^{hinge}_{\mathbf{x},y}(\mathbf{w}, b)$.

נחשב את הגרדיאנט של $f$:

עבור $f(w,b) = 0$, $\nabla f(w,b) = \vec{0}$

עבור $f(w,b) \neq 0$:

$$\nabla f(w,b) = \nabla\big(1 - y(x^T w + b)\big) = \nabla(1 - yx^T w - yb)) = \nabla(-yx^T w - yb)$$
$$= (-yx_1, \dots, -yx_d, -y)$$

לכן:

$$g = \begin{cases} 0 & when\ f(w,b) = 0 \\ (-yx_1, \dots, -yx_d, -y) & else \end{cases}$$

.5

5. Let $f_1, \dots, f_m : \mathbb{R}^d \to \mathbb{R}$ be a set of convex functions and $\mathbf{g}_k \in \partial f_k(\mathbf{x})$ for all $k \in [m]$ be sub-gradients of these functions. Define $f : \mathbb{R}^d \to \mathbb{R}$ by $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$. Show that $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$.

יהיו $x, u \in dom(\Sigma_k f_k)$

$$f(u) = \Sigma_k f_k(u) \geq_{from\ defenition\ of\ sub-gradient\ foreach\ f_k} \Sigma_k(f_k(x) + \langle g_k, u - k\rangle)$$
$$= \Sigma_k f_k(x) + \Sigma_k \langle g_k, u - x\rangle = \Sigma_k f_k(x) + \langle \Sigma_k g_k, u - k\rangle$$
$$= f(x) + \langle \Sigma_k g_k, u - k\rangle$$

ולכן, מהגדרה:

$$\Sigma_k g_k \in \partial f(x) = \partial \Sigma_k f_k(x)$$

6. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{\pm 1\}$ be a sample and define $f : \mathbb{R}^d \to \mathbb{R}$ by:

$$f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \ell_{\mathbf{x}_i, y_i}^{hinge}(\mathbf{w}, b) + \frac{\lambda}{2} ||\mathbf{w}||^2$$

Find a sub-gradient of $f$ for any $\mathbf{w}$.

נזכר בסעיף 4 ונגדיר:

$$g_i = \begin{cases} 0 & \ell_{x_i,y_i}^{hinge}(w,b) = 0 \\ (-yx_1, \dots, -yx_d, -y_i) & else \end{cases}$$

מהתרגול נובע כי:

$$\partial\left(\frac{1}{m}\Sigma_i\, \ell_{x_i,y_i}^{hinge}(w,b) + \frac{\lambda}{2}||w||^2\right) = \frac{1}{m}\partial\left((\Sigma_i\, \ell_{x_i,y_i}^{hinge}(w,b)) + \lambda\partial\left|\left|\frac{1}{2}w\right|\right|^2\right)$$

מהסעיף הקודם נובע

$$\partial\left(\Sigma_i \ell_{x_i,y_i}^{hinge}(w,b)\right) = \Sigma g_i$$

נותר למצוא את $\partial\left|\left|\frac{1}{2}w\right|\right|^2$:

$$\partial\left(\frac{1}{2}||w||^2\right) = \nabla\left(\frac{1}{2}||w||^2\right) =_{\text{מהתרגול}} w^T =_{function\ of\ (w,b)} (w, 0)$$

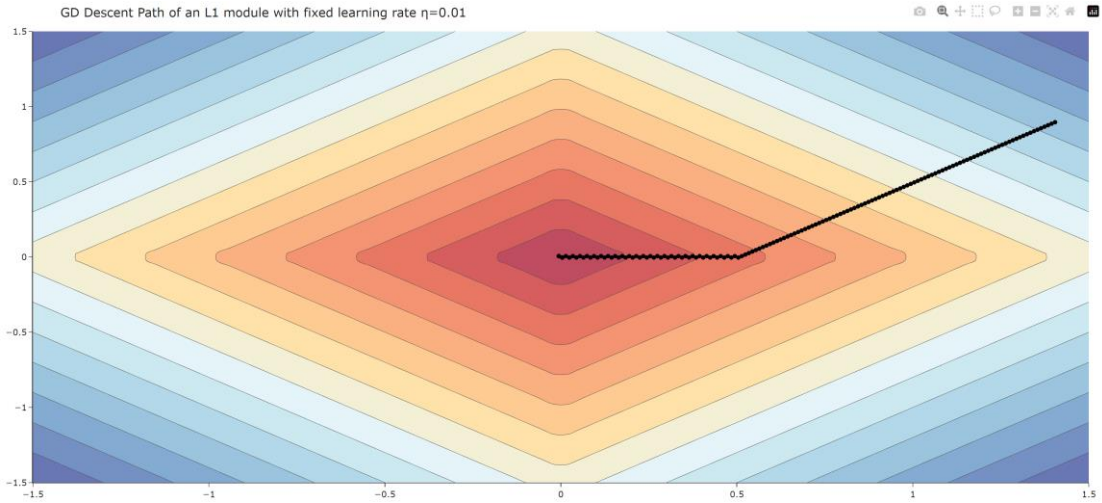
ונקבל:

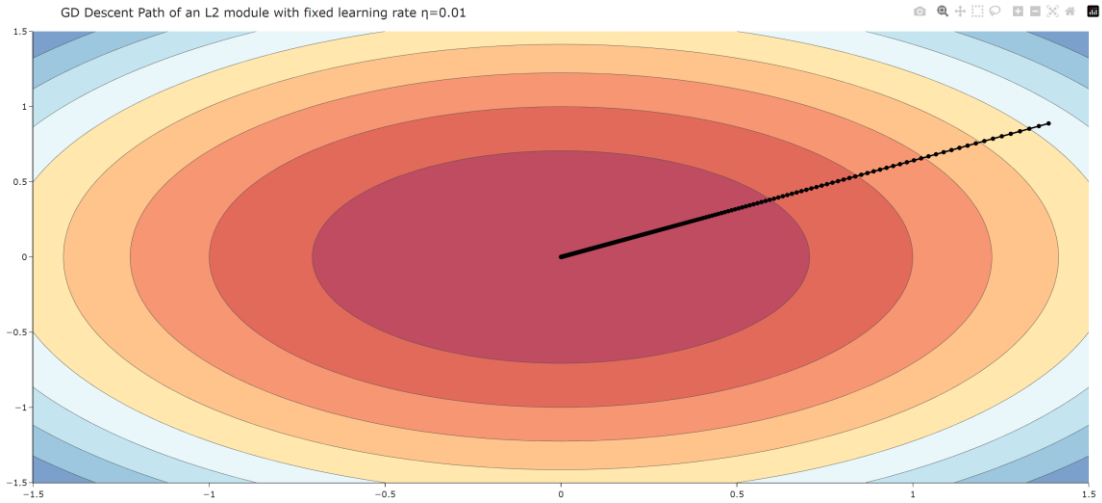
$$\frac{1}{m}\Sigma g_i + \lambda(w,0) \in \partial\left(\frac{1}{m}\Sigma_i \ell_{x_i,y_i}^{hinge}(w,b) + \frac{\lambda}{2}||w||^2\right)$$

## חלק פרקטי:

2.1 *Gradient Descent*

.1




בגרף של $L2$, נתיב הירידה יוצר קו ישר, הנובע מאופי הנורמה $L2$ המייצגת את המרחק האוקלידי במישור. נורמה זו גורמת לשיפוע ללכת בנתיב הירידה הישיר ביותר, וכתוצאה מכך מסלול חלק ומתמשך לעבר המינימום. הישירות של נתיב זה מצביעה על כך שתהליך האופטימיזציה יעיל יחסית, עם הפחתה עקבית בערך הפונקציה המובילה להתכנסות מהירה. הסדרת $L2$ מקדמת ירידה חלקה זו על ידי כיווץ אחיד של ערכי פרמטרים מבלי לאלץ מקדמי כלשהם לאפס, ובכך שומרת על ירידה מתמדת בתפקוד המטרה.

מצד שני, הגרף המתאים ל $L1$ מציג התנהגות שונה. נתיב הירידה עוקב בתחילה במסלול ישר אך לאחר מכן מבצע "שבירה" או זיגזג בולט כאשר הוא עובר למפלס המתאר הבא. הפסקה זו מתרחשת מכיוון שנורמת $L1$, בניגוד לנורמת $L2$, מקדמת דלילות על ידי העברת מקדמים מסוימים לאפס, וכתוצאה מכך שינויים פתאומיים בכיוון. נקודת המוצא של האלגוריתם משפיעה

באופן משמעותי על התנהגות זו; אם האלגוריתם התחיל ממיקום גבוה יותר על ציר ה-$y$, ייתכן שהוא היה ממשיך בקו ישר יותר לפני שהיה צריך להתאים. השינויים החדים בנתיב הם תוצאה של תהליך האופטימיזציה שמנסה לאפס מקדמים מסוימים במהירות, מה שמוביל למסלול ירידה פחות חלקה ולא יציבה בהשוואה לנורמה $L2$.

. 2

תופעה ראשונה:

ניתן לשים לב כי ההתקדמות לא נעשית במסלול הכי מהיר לנקודת המינימום (האופטימיזציה נוטה לנוע לאורך הקצוות).

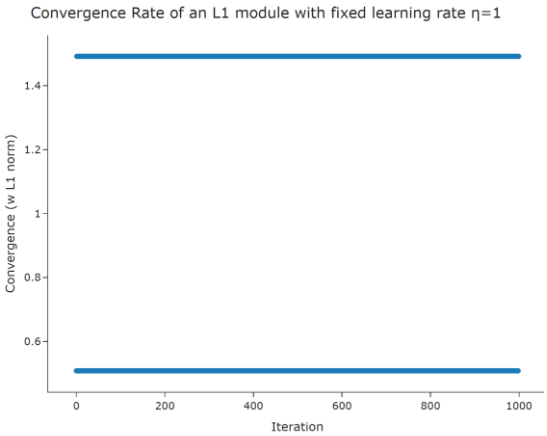
תופעה שניה:

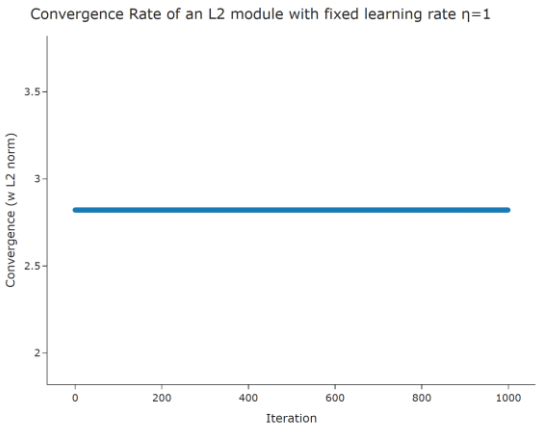
כמו כן, ניתן לשים לב כי ההתקדמות עבור כ"א מערכי ה-$\eta$ הולכות בקו ישר עד שהן מגיעות לאיזור של $y$=0 ואז נשברות וממשיכות במקביל לציר $x$ באופן יחסי. (אהיה כן, נעזרתי כאן קצת בצ'אט ג'יפיטי)

.3

כיאה לחומר ולתרגיל זה, נלך באופן יורד:

נתחיל ב-$\eta = 1$:

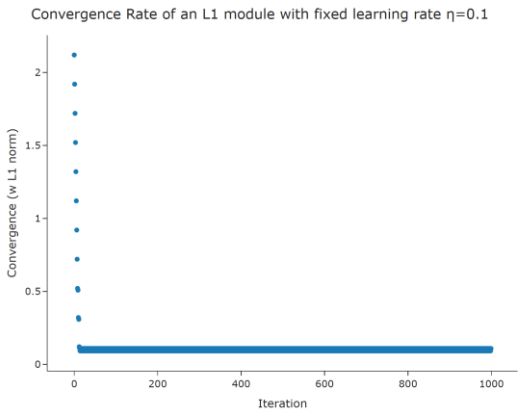


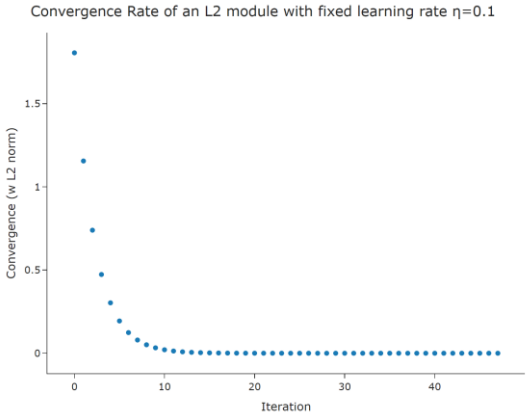




בגרף המתאים ל-$L2$ ניתן לשים לב כי ערכי השגיאה לא משתנים ונשארים קבועים על הערך2.75 ואין דעיכה במהלכו.

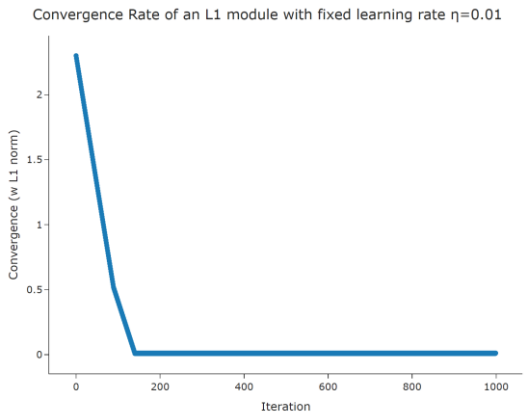
לעומת זאת, בגרף של $L1$, ערכי השגיאה קופצים בין 0.2~ ל1.5~ באופן קבוע. הדבר נובע ככל הנראה מהמקדם הגבוה יחסית שאיתו האלגוריתם מתקשה למצוא את נקודת המינימום ונעדיף מקדם נמוך יותר. בנוסף לכל זאת נשים לב שלא הפסקנו לפני 1000 איטרציות כלומר כל האיטרציות מומשו בגלל הקושי למצוא את המינימום.

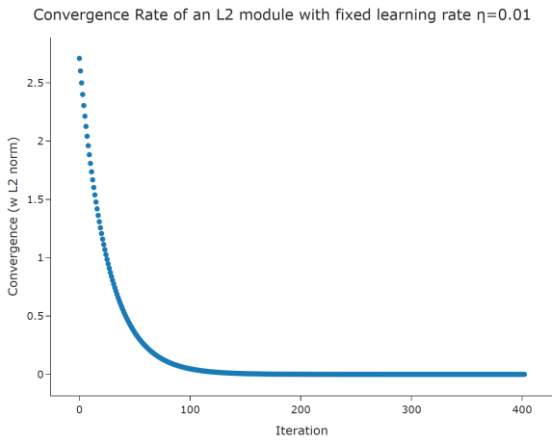
אז כאמור, נעבור למקדם נמוך יותר, $\eta = 0.1$



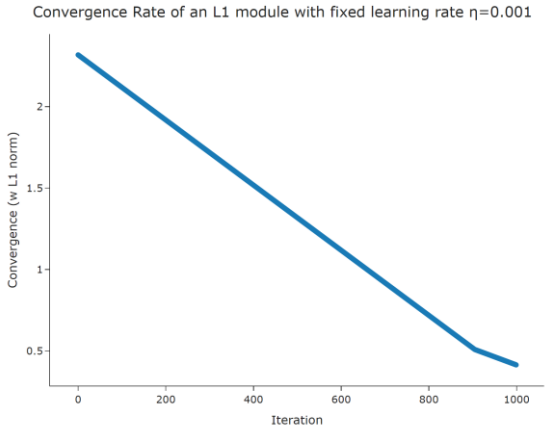



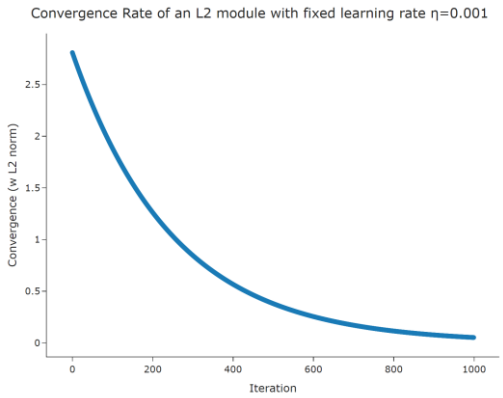

חיפוש המינימום עובד טוב יותר! האלגוריתם ב$L2$ הפסיק כבר אחרי 50 איטרציות! ב$L1$ עוד לא הגענו לכינוס למינימום, ספוילר: ה-$eta$ עדיין גדולה מדי




קיבלנו התכנסות בשתי הנורמות! אם כי ראוי לציין שאיבדנו קצת מהמהירות לכך בנורמה2 (במקום 50 איטרציות, אזור ה400).




זהו המקדם המינימלי – ואיתו באים גם צעדים קטנים יותר. מה שגורם ל: דעיכה "חלקה" יותר, וגם: התכנסות בשניהם. נשים לב שעכשיו לא רק נורמה1 תרוץ 1000 איטרציות אלא גם נורמה2. דהיינו כל האיטרציות.

.4

```
The lowest loss achieved by an L1 module with a fixed learning rate n=1 is 0.5081196195534
The lowest loss achieved by an L2 module with a fixed learning rate n=1 is 2.8210062332145
The lowest loss achieved by an L1 module with a fixed learning rate n=0.1 is 0.0918803804466
The lowest loss achieved by an L2 module with a fixed learning rate n=0.1 is 1.403e-09
The lowest loss achieved by an L1 module with a fixed learning rate n=0.01 is 0.0081196195534
The lowest loss achieved by an L2 module with a fixed learning rate n=0.01 is 2.391228e-07
The lowest loss achieved by an L1 module with a fixed learning rate n=0.001 is 0.4143075051928
The lowest loss achieved by an L2 module with a fixed learning rate n=0.001 is 0.0514619952652
```

אז:

```
The lowest loss achieved by an L1 module with a fixed learning rate n=0.01 is 0.0081196195534
```

```
The lowest loss achieved by an L2 module with a fixed learning rate n=0.1 is 1.403e-09
```

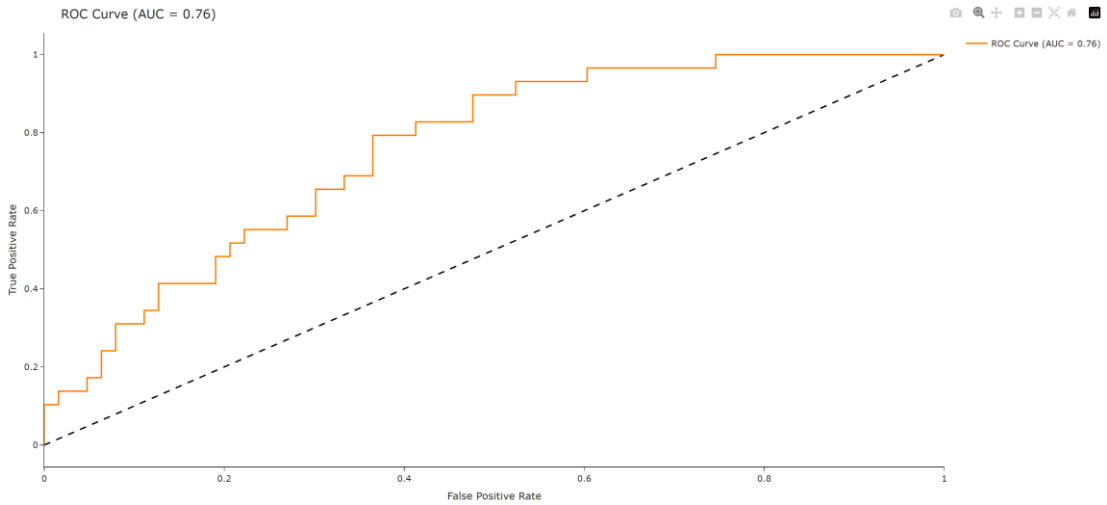
ניתן לראות שערך השגיאה המינימלי יותר נמוך משמעותית כאשר משתמשים ב$L2$ עם מקדם 0.1.

.

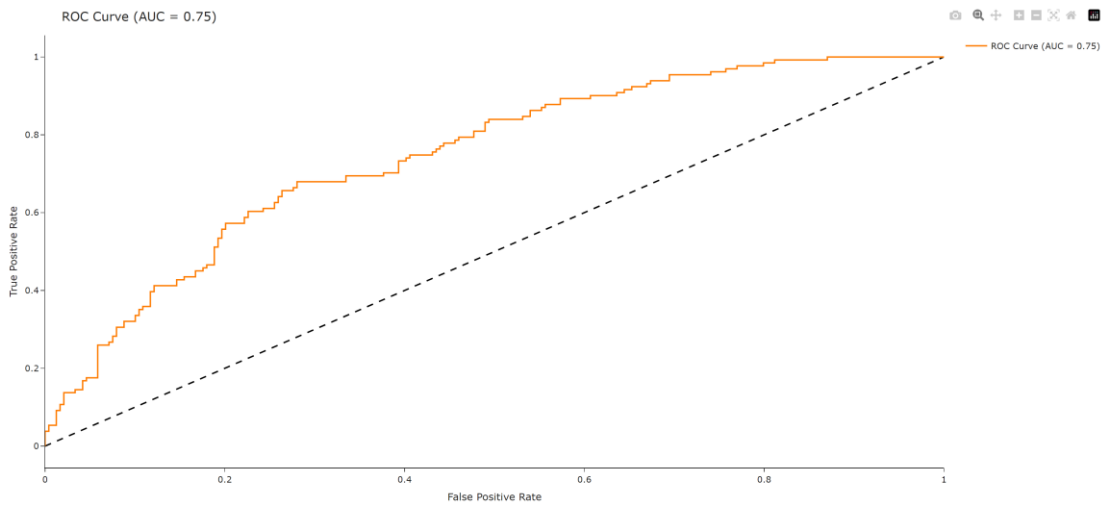
ההבדל מ$L1$ נובע ככל הנראה מכך שהשימוש ב$L2$ נכון יותר לבעיה בגלל שהוא מתאר את המרחק האוקלידי בין שתי נקודות ומתאר הכי טוב את הכיוון אליו האלגוריתם צריך ללכת כדי למזער את השגיאה כמה שיותר, בעוד שב-$L1$ הגרדיאנט הוא קבוע (0, 1 או 1-). מכאן שאין חשיבות לערכי המשקלים ב-$L1$ בעוד שב-$L2$ מטפל ב-*learning rate* גדולים יותר טוב יותר, מה שמוביל להתכנסות מהירה ומדויקת יותר (לכן ה-*loss* שלה נמוך יותר).

5. הבנתי מהפורום שצריך לצרף את שני הגרפים, ולהלן:

*ROC curve* עבור *regression logistic* מעל ה *set test*:




*ROC curve* עבור *regression logistic* מעל ה *set train*:



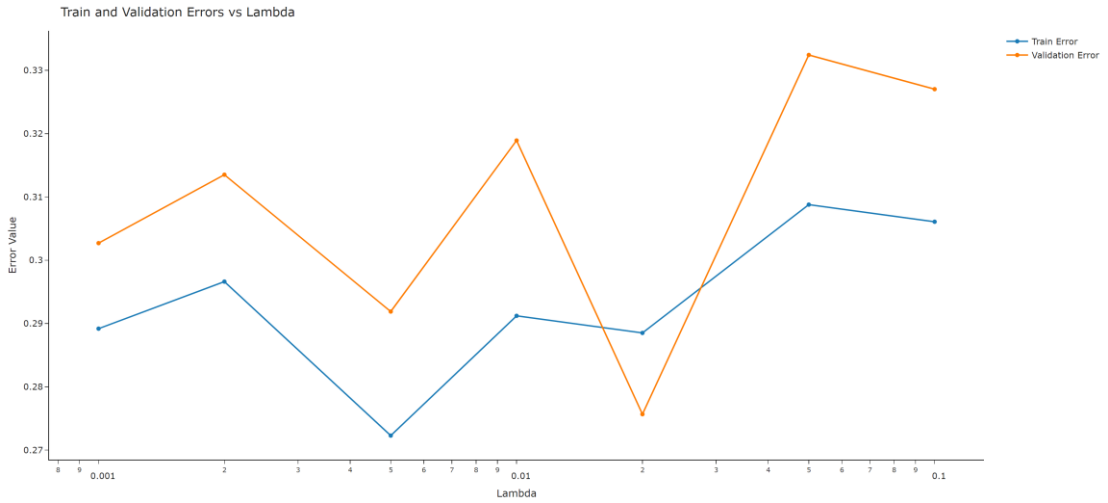

6. אפשר לראות בהדפסות כשמריצים את התרגיל:

```
Optimal alpha: 0.8586879152343948
Test error with optimal alpha: 0.31521739130434784
```

7. אפשר לראות בהדפסות כשמריצים את התרגיל:

```
Best lambda: 0.02
Test error with best lambda: 0.2826086956521739
```

והפלוט יהיה:




## ***שימוש ב-Chat GPT***

לצערי, בתרגיל זה השתמשתי בצ׳אט ג׳יפיטי יותר מהתרגילים הקודמים. אני לא תמיד מצליח להבין את הדברים בתלת מימד ולכן נעזרתי בו בתרגיל הזה גם בדו״ח, וכמובן שאני גם נעזר בו בחלק התכנותי כפי שעשיתי עד עכשיו בקורס.