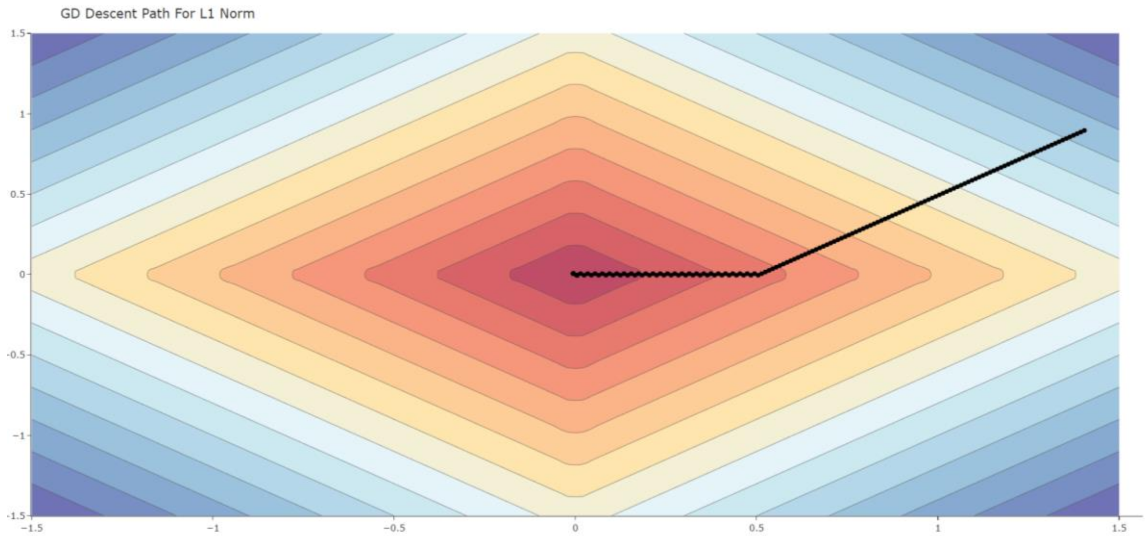
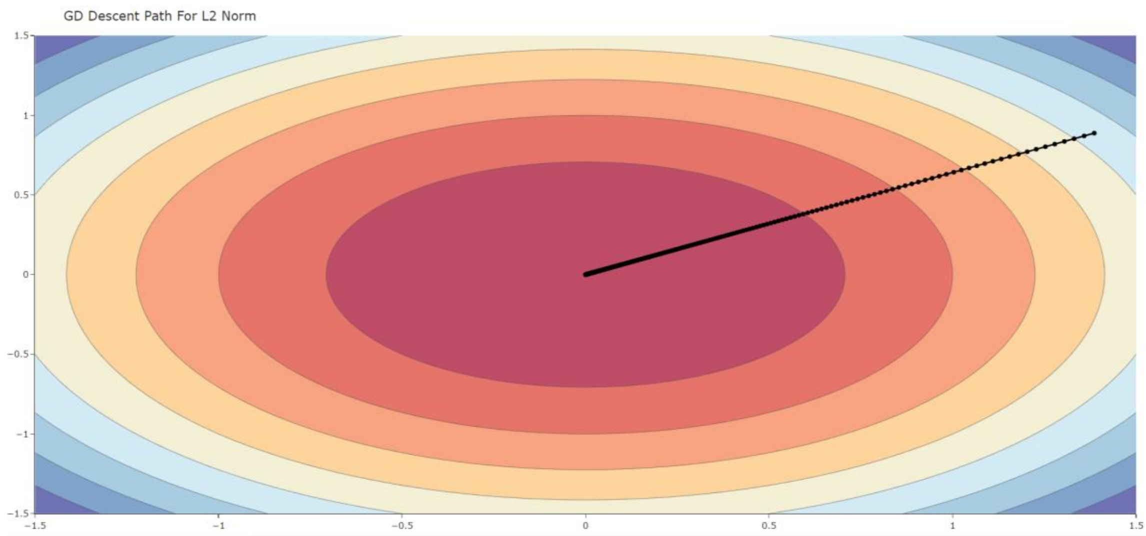
שאלה 1 :

1. Plot the descent path for each of the settings described above (you can use the `plot_descent_path`). Add below the plots for $\eta = 0.01$ and explain the differences seen between the `L1` and `L2` modules.

נשים לב כי היעקוביאן של L2 קטן בצורה משמעותית יותר ככל שהאלגוריתם מתקרב למינימום ולכן גודל הצעד על הפונקציה קטן בכל איטרציה וזה עקב השיפוע התלול של נורמתL2 . לעומת זאת, ב 1L השיפוע קבוע וזה גורם לכך שהצעד על הפונקציה ישאר קבוע בכל איטרציה (כי η קבוע). נשים לב כי עקב גודל הצעדים האלגוריתם עובר לחפש את המינימום גם כאשר L1 נמצאת בעלייה (שכן הוא מרחיק יותר מידי) מה שגורם לשינוי כיוון של הגרדיאנט, בעוד שב L2 כיוון הגרדיאנט נשאר קבוע ורק הגודל שלו משתנה.





שאלה 2 :

2. Describe two phenomena that can be seen in the descent path of the ℓ_1 objective when using GD and a fixed learning rate.

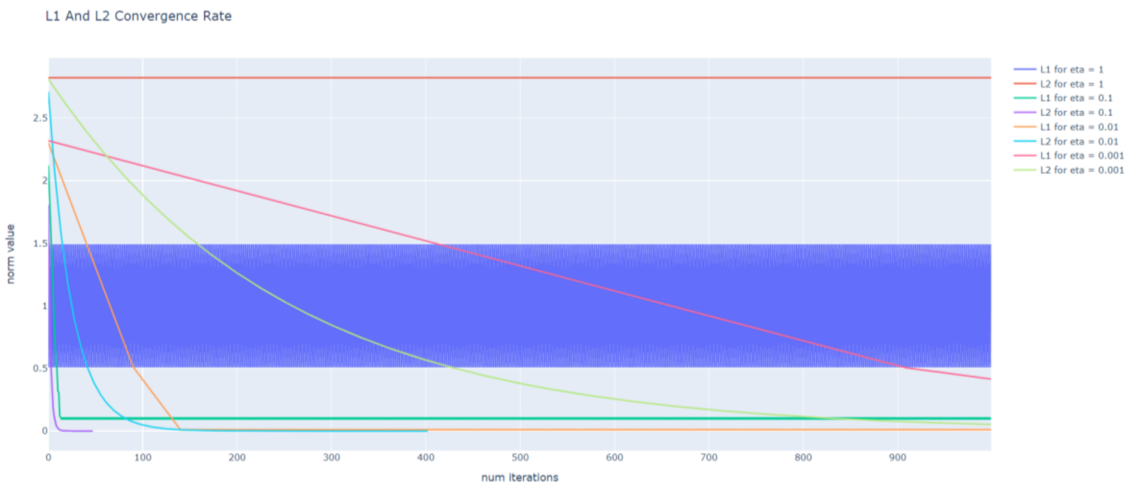
נבחין שישנה נקודה במסלול על נורמת L1 בה המסלול משנה כיוון כלפי מעלה, יורד למטה וכו׳ כך שהקו מתקרב למינימום אך לבסוף מתבדר ואינו מתכנס אליו, ולכן מינימום לא יתקבל לכל מספר של איטרציות .וזה כי קצב הלמידה קבוע ולכן אם האלגוריתם לא נפל בדיוק בנקודה בה הכיוון השתנה הוא ימשיך לחפש אותו בכיוון ההפוך בכל איטרציה מחדש, כך שהגרדיאנט ישנה את הכיוון כל איטרציה.

כמו כן גודל הצעדים נשאר קבוע שכן השיפוע קבוע ו־ η קבועה לאורך כל התקדמות האלגוריתם

שאלה 3 :

3. For each of the modules, plot the convergence rate (i.e. the norm as a function of the GD iteration) for all specified learning rates. Explain your results

ניתן לראות את חשיבות הבחירה של η. עבור η גדולה מידי נקבל שהאלגוריתם לא מצליח להגיע למינימום ונתקע באותם הערכים במספר האיטרציות הנתון. עבור η לא גדולה מידי ולא קטנה האלגוריתם מצליח להגיע לערך הנמוך ביותר במספר האיטרציות הנתון. עבור η קטנה מידי האלגוריתם מתקדם בצעדים קטנים מידי ולא מצליח להגיע למספר קרוב למינימום במספר האיטרציות הנתון.



שאלה 4:

4. What is the lowest loss achieved when minimizing each of the modules? Explain the differences

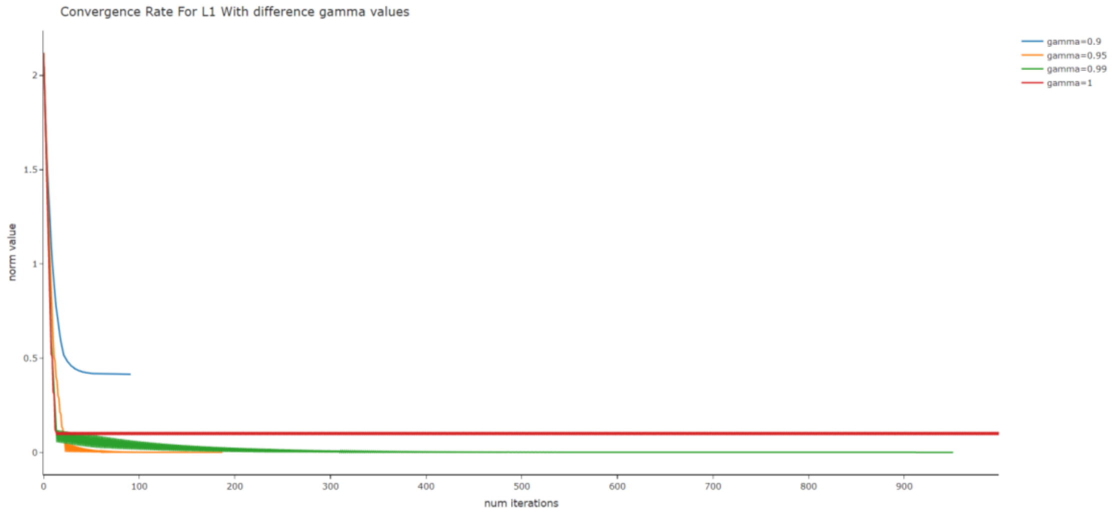
נשים לב כי אכן ה־ loss הקטן ביותר מתקבל עבור η = 0.01 בשתי הנורמות.

```
========================== Question 4 ==========================
    min loss for L1 Module for eta = 1 is 0.5081196195534134.

    min loss for L2 Module for eta = 1 is 2.821006233214517.

    min loss for L1 Module for eta = 0.1 is 0.09188038044658689.

    min loss for L2 Module for eta = 0.1 is 1.4029519498344255e-09.

    min loss for L1 Module for eta = 0.01 is 0.008119619553413011.

    min loss for L2 Module for eta = 0.01 is 2.3912283661218465e-07.

    min loss for L1 Module for eta = 0.001 is 0.4143075051928211.

    min loss for L2 Module for eta = 0.001 is 0.05146199526515047.

=================================================================
```

שאלה 5 :

5. Plot the convergence rate for all decay rates in a single plot. Explain your results.

ניתן לראות את חשיבות הבחירה של קצב הדעיכה. עבור γ גדולה מידי נקבל כי האלגוריתם לא מצליח להגיע למינימום ונתקע באותם הערכים במספר האיטרציות הנתון. עבור γ לא גדולה מידי האלגוריתם מצליח להגיע לערך הנמוך יותר במספר האיטרציות הנתון. עבור γ קטנה מידי האלגוריתם מתקדם בצעדים קטנים מידי ולא מצליח למספר קרוב למינימום במספר האיטרציות הנתון .



שאלה 6 :

6. How does the algorithm perform using the exponential decay compared to the fixed learning rate? What is the lowest ℓ_1 norm achieved using the exponential decay. Explain why there are differences.

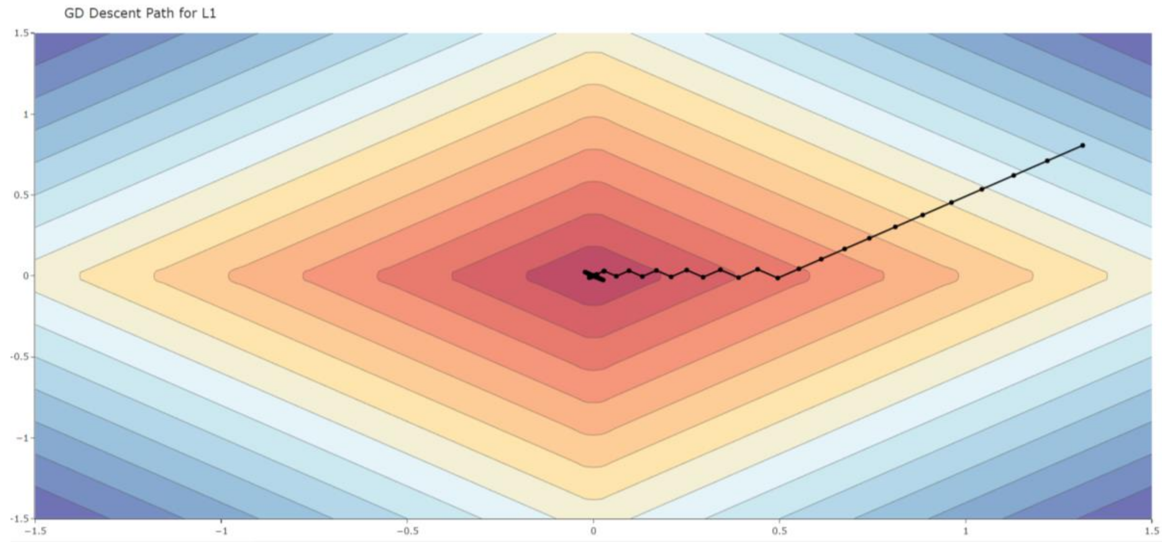
נשים לב כי הפעם ה־ loss המינימלי אשר התקבל עבור L1 כאשר קצב הלמידה אקספוננציאלי קטן משמעותית מאשר ה־ loss המינימלי שהתקבל עבור קצב הלמידה הקבוע ההבדל נובע מכך שקצב הלמידה האקספוננציאלי מתאים את קצב הלמידה עם ההתקדמות ולכן פחות תלוי בנקודת ההתחלה של האלגוריתם.

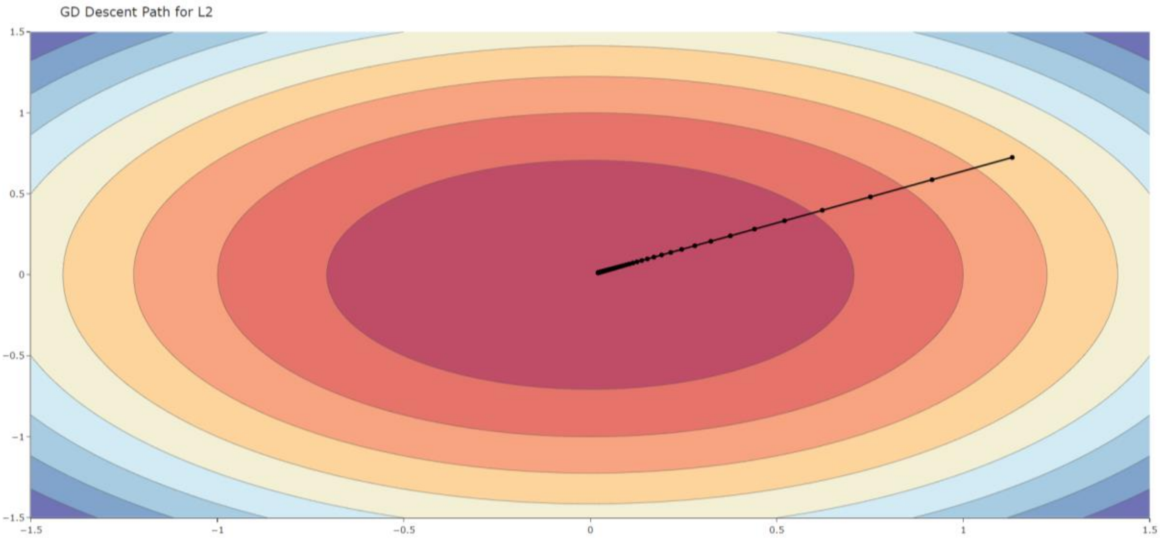
```
=========================== Question 6 ===========================
    The Lowest L1 Loss Rate for gamma=0.95
    and eta=0.1, is 3.493826212995591e-07.
===================================================================
```

שאלה 7 :

7. Plot the descent path for the $\gamma = 0.95$. Describe how the descent path changed from when using a fixed learning rate.

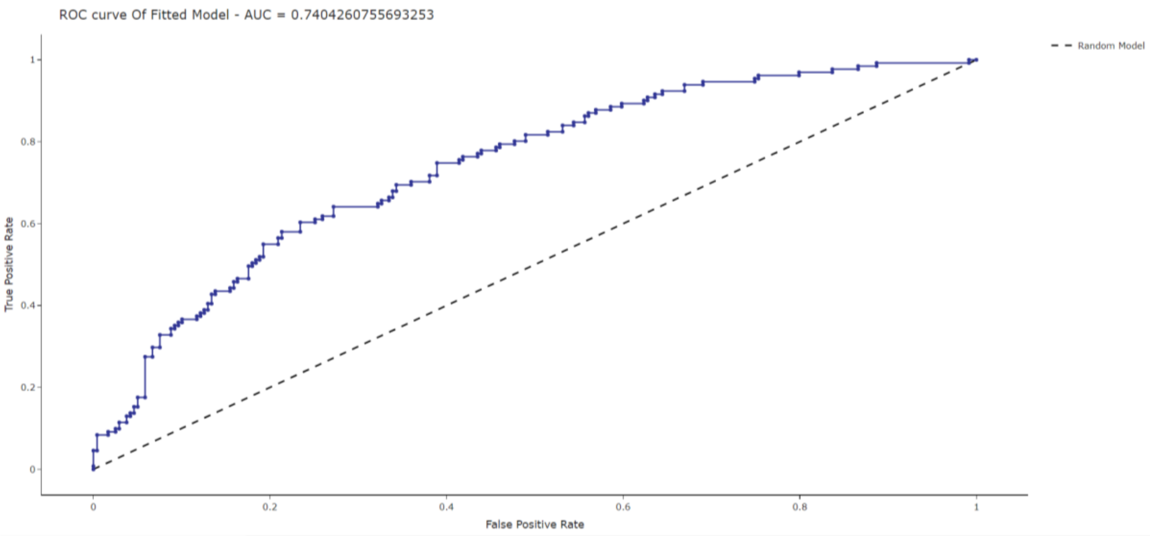
נשים לב כי שני המקרים ב L2 דומים (שניהם מתכנסים למינימום אבל המקרה האקספוננציאלי מתכנס
מהר יותר), לעומת המקרה הנוכחי ב־ L1 השונה מהמקרה של קצב הלמידה הקבוע .במקרה של קצב
הלמידה הקבוע קיבלנו כי הקצב הקבוע הוביל להתבדרות סביב המינימום, לעומת המקרה הנוכחי בו ניתן
לראות שיש התכנסות אל המינימום שכן האלגוריתם מחשב את הצעד הכדאי ביותר בכל איטרציה ובכך
"דילוג" על 14 הנקודה בה הגרדיאנט משנה את כיוונו (0) y = לא משנה את ההתכנסות אל המינימום.





שאלה 8 :

8. Using your implementation, fit a logistic regression model over the data. Use the `predict_proba` to plot an ROC curve. You can use sklearn's `metrics.roc_curve` function and the code provided in Lab 04.



שאלה 9 :

9. Which value of α achieves the optimal ROC value according to the criterion below. Using this value of α^* what is the model's test error?

$$\alpha^* = \text{argmax}_\alpha\{\text{TPR}_\alpha - \text{FPR}_\alpha\}$$

```
=========================== Question 9 ===========================
    Best alpha for the logistic regression is 0.42675449507831764
    with loss of 0.5108695652173914 on the test set.
===================================================================
```

שאלה 10 :

10. Fit an ℓ_1-regularized logistic regression by passing `penalty="l1"` when instantiating a logistic regression estimator
    - Set $\alpha = 0.5$
    - Use your previously implemented cross-validation procedure to choose λ
    - After selecting λ repeat fitting with the chosen λ and $\alpha = 0.5$ over the entire train portion.

For values of What value of λ was selected and what is the model's test error?

```
=========================== Question 10 ===========================
    the best Lambda is 0.002 with test score of 0.6847826086956522.
====================================================================
```

שאלה 11 :

11. Repeat question 10 for ℓ_2 regularized logistic regression. What value of λ was selected and what is the model's test error?

```
=========================== Question 11 ===========================
    the best Lambda is 0.1 with test score of 0.31521739130434784.
===================================================================
```

חלק תיאורטי

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

נרצה להוכיח כי $g(u)$ היא פונקציה קמורה כאשר לכל

$i \in [m]$ f_i פונ' קמורה ו- γ_i מספר ממשי חיובי לכל $i \in [m]$

נתון שמכיון ש $\gamma_i > 0$ לכל $i \in [m]$ אז גם $\gamma_i f_i(u)$ קמורה

שהרי מתקיים $f(\alpha v + (1-\alpha)u) \le \alpha f(v) + (1-\alpha) f(u)$

ואם נכפול בקבוע האי"ש ישמר ולכן

$$g(\alpha v + (1-\alpha)u) \le \sum_{i=1}^{m} \gamma_i f_i(\alpha v + (1-\alpha)u) \le$$

$$\le \sum_{i=1}^{m} \alpha \gamma_i f_i(v) + (1-\alpha) f(u) = \alpha \sum_{i=1}^{m} \gamma_i f_i(v) + (1-\alpha) \sum_{i=1}^{m} \gamma_i f_i(u) =$$

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

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 than h is convex as well.

דוגמא נגד.ית : $g(x) = x^2$, $f(x) = -x$ שתי הפונקציות קמורות

אבל $h(x) = -x^2$ שהיא ההרכבה אינה קמורה.

3. Let $f : C \to \mathbb{R}$ be a function defined over a convex set C. Prove that f is convex iff its *epigraph* is a convex set, where $\text{epi}(f) = \{(u,t) : f(u) \le t\}$.

נניח ש-f קמורה ונראה ש $epi(f)$ סט קמור

נניח בשלילה שקיימות 2 נקודות שונות $(a,b),(c,d)\in epi(f)$

וגם $\lambda\in(0,1)$ ו- $\lambda(a,b)+(1-\lambda)(c,d)\in epi(f)$

נקבל ש: $f(a)\le b$ וגם $f(c)\le d$

וגם $f(\lambda a+(1-\lambda)f(c))\le \lambda f(a)+(1-\lambda)f(c)\le$

$\le \lambda b+(1-\lambda)d$ בסתירה.

בכיוון השני: נניח ש $epi(f)$ גם קמור אזי לכל

$(x,y),(z,t)\in epi(f)$ ולכל $a\in[0,1]$ מתקיים:

$$\alpha(x,y)+(1-\alpha)(z,t)=(\alpha x+(1-\alpha)z,\alpha y+(1-\alpha)t)\in epi(f)$$

כלומר: $f(\alpha x+(1-\alpha)z)\le\alpha y+(1-\alpha)z$

נבחר $y=f(x)$, $t=f(z)$ ונקבל

$$f(\alpha x+(1-\alpha)z)\le\alpha f(x)+(1-\alpha)f(z)$$

ולכן f קמורה.

4. Let $f_i:V\to\mathbb{R}$, $i\in I$. Let $f:V\to\mathbb{R}$ given by

$$f(u)=\sup_{i\in I}f_i(u).$$

If f_i are convex for every $i\in I$, then f is also convex.

נראה ש $epi(f)=epi(\sup f_i(u))$ גם קמור ומשאלה 3

זה מוכיח ש- $\sup f_i(\omega)$ פונקציה קמורה

מתקיים $(a,b) \in epi(\sup f_i(\omega))$ אמ"מ $f(a) \leq b$

$\Leftarrow f_i(\omega) \leq b \quad \forall i \in I \Leftarrow (a,b) \in epi(f_i)$

אם כן נקבל ש $epi(f) = \bigcap_{i \in I} epi(f_i)$ קמור כי חיתוך

של סטים קמורים הוא קמור

5. 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 $\mathbf{w}, b$.

ראינו ש- Pointwise-max קמורה כלומר בהנחה

שהפונקציות $1-y(w^t x+b)$ ו- 0 קמורות

$\max(0, 1-y(w^t x+b))$ קמורה

ונבחין שהפונקציה 0 קמורה כי היא ליניארית

ו- $1-y(w^t x+b)$ קמורה כי היא הרכבה על פונקציה אפינית

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

נבחין כי $\ell^{hinge}_{x,y}$ דיפרנציאבילית לכל w,b שלא מקיימים

$y(x^t w + b) = 1$

נבחין כי לכל (w,b) המקיימים $y(x^t w+b)\geq 1$ מתקיים

ש- $0\in\partial l^{hinge}_{x,y}(w,b)$ ולכל $y(x^t w+b)<1$ מתקיים:

$$l^{hinge}_{x,y}(w,b)=1-y(x^t w+b)=1-yx^t w-yb \Rightarrow$$

$$\Rightarrow \partial^{hinge}_{x,y}(w,b)=\left(\frac{\partial}{\partial w},\frac{\partial}{\partial b}\right)=(-yx,-y)$$

נבחר כי:

$$g=\begin{cases}0 & y(x^t w+b)\geq 1\\ (-yx,-y) & y(x^t w+b)<1\end{cases}$$

7. Let $f_1,\ldots,f_m:\mathbb{R}^d\to\mathbb{R}$ 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 $\sum_k \mathbf{g}_k\in\partial\sum_k f_k(\mathbf{x})$.

מהגדרת סאבגרדיאנט מתקיים לכל $i\in[m]$ ולכל $u\in\mathbb{R}^d$:

$$f_i(u)\geq f_i(x)+\langle g_i,u-x\rangle$$

$$\Rightarrow \sum_{i=1}^m f_i(u)\geq\sum_{i=1}^m f_i(x)+\langle g_i,u-x\rangle$$

$$\sum_k g_k\in\partial\sum f_k(x)$$ $\Leftarrow$ מההגדרה

8. 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:

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

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

בדומה לשאלה 6 נגדיר

$$g=\begin{cases}0 & y_i(x_i^t w+b)\geq 1\\ (-y_i x_i,-y_i) & y_i(x_i^t w+b)<1\end{cases}$$

ונקבל ש- $g_i \in \partial l^{hinge}_{x_i,y_i}(w,b)$

ומשאלה 7 נקבל:

$$\frac{1}{m}\sum_{i=1}^{m} g_i + \lambda(w,0) \in \partial \frac{1}{m}\sum_{i=1}^{m} l^{hinge}_{x_i,y_i}(w,b) + \frac{\lambda}{2}\|w\|^2$$