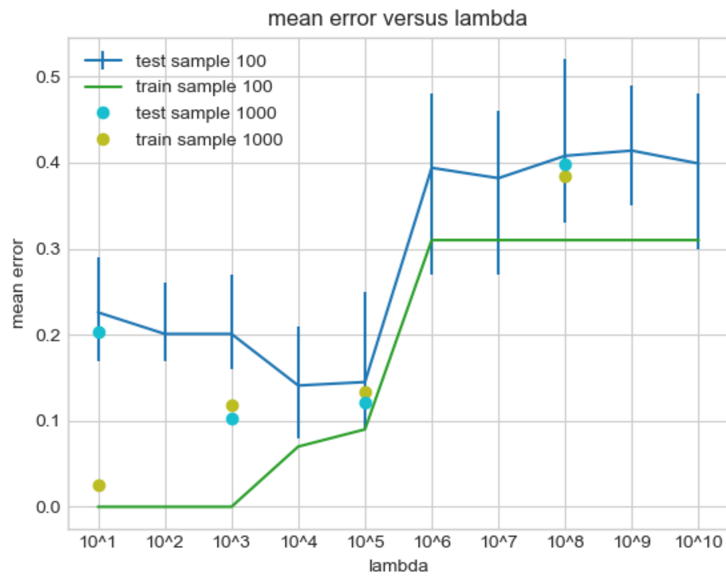
Assignment 2

שאלה 2

.b



.c

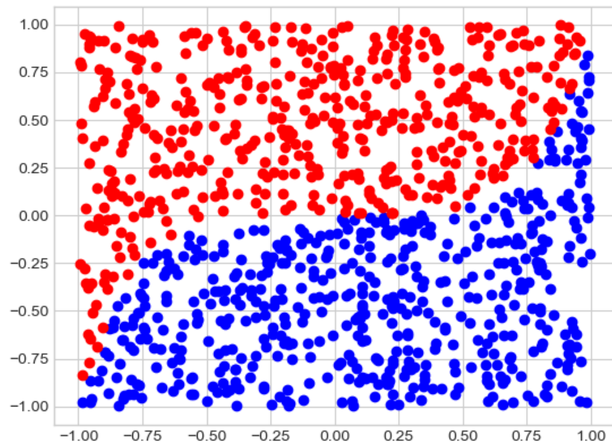
- כפי שניתן לראות בגרף, הtraining error של גודל מדגם 100 יצא נמוך משל גודל מדגם 1000. דבר זה הגיוני מכיון שככל שהמדגם גדל, הסיכוי שיהיה דרך טובה להפריד את המדגם ולקבל שגיאה נמוכה קטנה.
- מכיוון שהאימון נעשה על מדגם גדול יותר, אנו מצפים שפונקציית הניבוי תהיי נכונה יותר ותוכל לסווג מדגם בדיקה כמו הtest בצורה טובה יותר. זה אכן מה שאנחנו רואים בגרף, בכל המקרים הtest error של גודל מדגם 1000 נמוך מזה של גודל מדגם 100.
- הטרנד של הtest error כפונקציה של λ אמור להראות כסוג של פרבולה. כלומר- מתחיל גבוהה כאשר ה λ נמוכה (אין הרבה עונש על גודל w גבוהה מה שיכול לגרום לoverfitting) דבר שאפשר לראות בגרף לפי ההבדל בין הtest error לtrain error.

  יורד עד שמגיעים למצב של λ אופטימלי כלשהי, בו אין overfitting. רואים זו בגרף באזור $\lambda = 10^5$ ששם ההבדל בין הerror לtrain error נמוך.

  ולבסוף עולה שוב כאשר ה λ גבוה מכיון שאנחנו מענישים יותר מידי על גודל w ולכן מפספסים אופציות שיכולות להיות טובות לסיווג. ניתן לראות שגם ה train error גבוהה מאוד, כלומר קשה לסווג עם λ כזו גבוהה.

שאלה 4

.a



ניתן לראות שלא קיים קו לינארי יחיד שיפריד בצורה טובה את הנקודות. שימוש בקו לא לינארי יוכל לעשות עבודה יותר טובה בהפרדה זו ולכן שימוש בקרנל soft SVM יוכל להועיל.

.b תוצאות ההרצה-

```
results for 5 fold cross validation on soft SVM with poly kernel:
The average validation error for [1 2] is 0.068
The average validation error for [1 5] is 0.05800000000000001
The average validation error for [1 8] is 0.048
The average validation error for [10  2] is 0.069
The average validation error for [10  5] is 0.064
The average validation error for [10  8] is 0.06
The average validation error for [100   2] is 0.069
The average validation error for [100   5] is 0.064
The average validation error for [100   8] is 0.061
The selected parameter is [1 8]
test error for selected parameters: 0.02
results for 5 fold cross validation on soft SVM:
The average validation error for 1 is 0.063
The average validation error for 10 is 0.063
The average validation error for 100 is 0.063
The selected parameter is 1
test error for selected parameters: 0.04
```

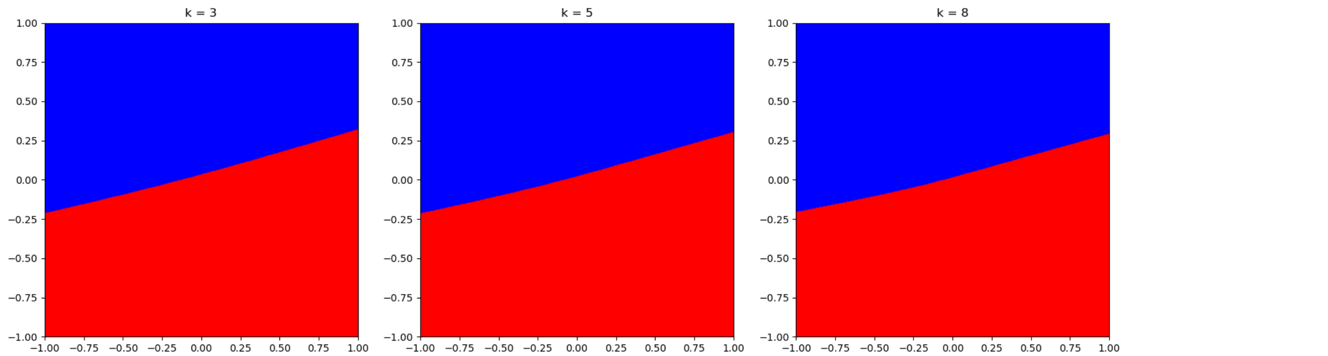
.c כפי שניתן לראות, התוצאות השגיאה עבור soft SVM עם קרנל פולינומי טובה יותר מאשר בלי, כלפי שהנחנו שיהיה המצב בסעיף א.

.d סיבה אחת בעד שימוש SVM פולינומי יכולה להיות זה שבמקרה הכללי יכול להיות התפלגות שסיווג בעזרת מפריד לינארי לא מתאים לה כלל, אבל סיווג בעזרת מפריד פולינומי ייתן תוצאות טובות יותר (כפי שניתן לראות במקרה שלנו).

סיבה ששימוש בSVM פולינומי יכול להוביל לשגיאה גדולה יותר יכול להיות שההתפלגות כן יכולה להיות מופרדת בצורה טובה יחסית עם מפריד לינארי ובכך שאנו מאפשרים

מפריד פולינומי אנו עלולים לגרום למצב של overfitting, יוחזר מסווג שמותאם מאוד למדגם האימון אך לא בהכרח להתפלגות הכללית ובכך להגדיל את השגיאה.

.e



.f עבור k=5 הערך הטוב ביותר שמצאנו בסעיף ב הוא $\lambda = 1$.

.i כדי לקבל את w מα שהוחזר השתמשנו בנוסחה-

$$\mathrm{w} = \sum_{i=1}^{m} \alpha_i * \psi(x_i)$$

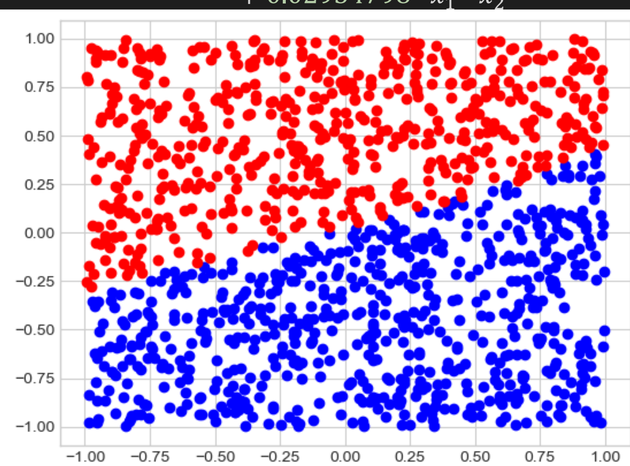
.ii הקואורדינטות של w הינן-

```
[  0.01351729 -0.3078727  -0.00106905 -0.15673627 -0.00225819 -0.02410338
   0.09152261  0.00465646 -0.02252933 -0.00274381 -0.00744335  0.01390728
  -0.12513128  0.00060624 -0.02453628  0.10939883  0.00724383 -0.00143707
   0.00755525 -0.01444307  0.02954798]
```

.iii עבור המכפלה $< w, \psi(x) >$ נקבל את הmultivariate polynomial הבא:

$$0.01351729 \cdot x_1^0 \cdot x_2^0 - 0.3078727 \cdot x_1^0 \cdot x_2^1 - 0.00106905 \cdot x_1^0 \cdot x_2^2 - 0.15673627 \cdot x_1^0 \cdot x_2^3 - 0.00225819 \cdot x_1^0 \cdot x_2^4 - 0.02410338 \cdot x_1^0 \cdot x_2^5 + 0.09152261 \cdot x_1^1 \cdot x_2^0 + 0.00465646 \cdot x_1^1 \cdot x_2^1 - 0.02252933 \cdot x_1^1 \cdot x_2^2 - 0.00274381 \cdot x_1^1 \cdot x_2^3 - 0.00744335 \cdot x_1^1 \cdot x_2^4 + 0.01390728 \cdot x_1^2 \cdot x_2^0 - 0.12513128 \cdot x_1^2 \cdot x_2^1 + 0.00060624 \cdot x_1^2 \cdot x_2^2 - 0.02453628 \cdot x_1^2 \cdot x_2^3 + 0.10939883 \cdot x_1^3 \cdot x_2^0 + 0.00724383 \cdot x_1^3 \cdot x_2^1 - 0.00143707 \cdot x_1^3 \cdot x_2^2 + 0.00755525 \cdot x_1^4 \cdot x_2^0 - 0.01444307 \cdot x_1^4 \cdot x_2^1 + 0.02954798 \cdot x_1^5 \cdot x_2^0$$

.iv



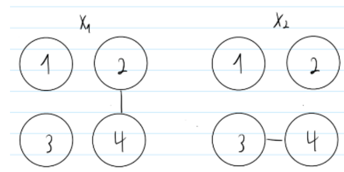
שאלה 5

a. נראה כי ההתפלגות אינה realizable עבור $\mathcal{H}$. נתבונן בדוגמה הבאה:
נגדיר התפלגות D בה עבור כל $x \epsilon X$ כך שדרגת קודקוד 1 היא 0, תוויתו תהיי 1, אחרת 0. כלומר-

$$D(x) = \begin{cases} 1 & g(x)_1 = 0 \\ 0 & \text{אחרת} \end{cases}$$

ונבחין בגרפים $x_1, x_2 \epsilon X$ הבאים:

$$g(x_1) = (0,1,0,1)$$
$$g(x_2) = (0,0,1,1)$$



לא קיים $v \in \mathbb{N}^n$ יחיד כך ש $v = (0,1,0,1)$ וגם $v = (0,0,1,1)$ ולכן לא קיים h_v כך ש

$$h_v(x_1) = h_v(x_2) = 1$$

לכן הבעיה אינה *realizable*. נשתמש בחסם *PAC* האגנוסטי –

$$m \geq \frac{2\log(|\mathcal{H}|) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2}$$

קיבלנו תלות ריבועית בϵ.

b. נחשב את גודל מחלקת ההיפותזות. מכיוון שX היא מחלקת הגרפים הלא מכוונים עם קדקודים עם דרגה 7 לכל היותר, כל $v \in \mathbb{N}^n$ שבו כאורדינטה כלשהי גדולה מ7 תוביל לכך ש$h_v = 0$, ודבר זה אינו במחלקת ההיפותזות.
לכן גודל המחלקה היא כמות ה$v \in \mathbb{N}^n$ השונים כך שאף כאורדינטה לא גדולה מ7.

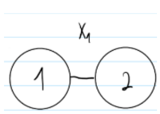
$$|\mathcal{H}| = 8^n$$

נציב זו בחסם –

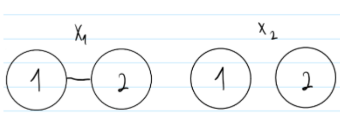
$$m \geq \frac{2\log(8^n) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = \frac{2n\log(8) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2}$$

קיבלנו תלות לינארית בn.

c. נמצא את הVC-dimension של $\mathcal{H}$.



עבור דוגמה יחידה $x_1 \in X$ כמובן שניתן לנתץ,
עבור תווית $y_1 = 1$ ניקח $\text{v} = (1,1)$
עבור תווית $y_1 = 0$ ניקח $\text{v} = (0,0)$



עבור זוג דוגמאות $x_1, x_2 \in X$ נראה כי אי אפשר לנתץ.
עבור תוויות $y_1 = 1, y_2 = 1$ לא קיים v שייתן סיווג זה

לכן $VC(\mathcal{H}) = 1$. נשתמש במידע זה כדי לשפר את החסם

$$m = \Theta\left(\frac{VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right) = \Theta\left(\frac{1 + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$$

החסם השתפר, אין תלות בn.

שאלה 6

פסודו- קוד:

Perceptron_steps
Input: A training sample S
Output: upper bound on number of steps Perceptron would take or -1

1. w ← hard_svm(S)
2. If hard_svm fails return -1
3. $R \leftarrow \max_i \|x_i\|$
4. $\gamma \leftarrow \frac{1}{R}\min_i \frac{|<w,x_i>|}{\|w\|}$
5. Return $\frac{1}{\gamma^2}$

אנו יודעים כי אם hard_svm יכשל אז המדגם לא פריד ולכן אנו מחזירים 1- כנדרש, אחרת אנו יודעים כי החסם העליון למספר העדכונים של ה-Perceptron הוא $\frac{1}{\gamma^2}$, מה-hard_svm נחלץ את הרווח הגדול בעזרת ה-w שחזר.

שאלה 7

a. נמזער את הבעיה הבאה שערכה שווה ל $Minimize\ \lambda\|w\|^2 + \sum_{i=1}^{m}[\ell^h(w,(x_i,y_i))]^2$

$$Minimize\ \lambda\|w\|^2 + \sum_{i=1}^{m}\xi_i^2$$

$$s.t. \forall i, y_i\langle w, x_i\rangle \geq 1-\xi_i\ and\ \xi_i \geq 0.$$

ערכיהם זהים בדיוק מאותן סיבות שנראו בכיתה, רק שבפונקציה שאותה צריך למזער אנו מעלים את ξ בריבוע.

b. בתוכנית הריבועית הזו אנו נמזער וקטור $z \in \mathbb{R}^{d+m}$ כאשר $z = \{w_1, \dots, w_d, \xi_1, \dots, \xi_m\}$

$$H = \begin{pmatrix} 2\lambda I_d & 0 \\ 0 & 2I_m \end{pmatrix}$$

$$u\ = \{0, \dots, 0\}$$

$$v = \{0, ..m..0\ , 1, ..m..1\}$$

$$A = \begin{pmatrix} 0 & I_m \\ y_i x_i & I_m \end{pmatrix}$$

כאשר $y_i x_i$ היא מטריצה בגודל *dXm* בצורה הבאה -

$$y_i x_i = \begin{pmatrix} y_1 \cdot \{x_1(1), \dots, x_1(d)\} \\ \vdots \\ y_m\ \cdot\ \{x_m(1), \dots, x_m(d)\} \end{pmatrix}$$

שאלה *8*:

A) נניח בשלילה כי קיימת פונקציה $f: \mathbb{R} \to \mathbb{R}$ מונוטונית לא יורדת המקיימת עבור כל $x \in \mathbb{R}^d$ –

$$f(\|x\|_2) = \|x\|_1$$

נתבונן ב$x_1 = (6,1), x_2 = (4,4)$.

$$f(\|x_1\|_2) = f(6.08) = \|x_1\|_1 = 7$$

$$f(\|x_2\|_2) = f(5.65) = \|x_2\|_1 = 8$$

נשים לב כי $\|x_2\|_2 > \|x_1\|_2$ אך $\|x_2\|_1 < \|x_1\|_1$, סתירה למונוטוניות של f. לכן לא קיימת פונקציה כזו ולכן תנאי ה-$representer\ theorem$ לא יכולים להתקיים עבור בעיית המזעור הזו.

B) נוכל להסיק שאין דרך לייצג את *w*, מבלי לציין את כל קואורדינטות שלו, ולכן לא נוכל להשתמש בקרנל טריק עבור בעיה האופטימיזציה הזאת.

שאלה 9:

עבור סעיפים *a, b* נניח בשלילה כי *K* היא פונקציית קרנל ונראה סתירה.

A) אנו יודעים כי:

$$K(x,x') = \langle\psi(x),\psi(x')\rangle = \big(x(7)+x(3)\big)\cdot x'(1) = x(7)x'(1) + x(3)x'(1)$$

עבור המקרה בו $x = x'$ חייבת להתקיים תכונת האי שליליות של המכפלה הפנימית עבור $\psi(x)$, אולם לדוגמה עבור $x(3), x(7) > 0 \wedge x(1) < 0$ מתקיים $\langle\psi(x),\psi(x')\rangle = \langle\psi(x),\psi(x)\rangle < 0$.

B) אנו יודעים כי:

$$K(x,x') = \langle\psi(x),\psi(x')\rangle = 3 - \big(x(1)-x(2)\big)\cdot(x'(1)-x'(2))$$

עבור המקרה בו $x = x'$ חייבת להתקיים תכונת האי שליליות של המכפלה הפנימית עבור $\psi(x)$, אולם לדוגמה עבור $x(2)=0, x(1) > \sqrt{3}$ מתקיים $\langle\psi(x),\psi(x')\rangle = \langle\psi(x),\psi(x)\rangle < 0$.

C)

אנו חושבים כי $\psi(x)\colon \mathbb{R}^5 \to \mathbb{R}^3$, כאשר $\psi(x) = (x(1)^4, e^{x(3)+x(5)}, \frac{1}{x(1)})$

נציין כי:

$$f(x,x') = \big(x(1)x'(1)\big)^4 + e^{x(3)+x(5)+x'(3)+x'(5)} + 1/(x(1)x'(1))$$

לכן עבור מפת הפיצ'רים שבחרנו מתקיים:

$$\langle\psi(x),\psi(x')\rangle = x(1)^4 * x'(1)^4 + e^{x(3)+x(5)} * e^{x'^{(3)}+x'^{(5)}} + \frac{1}{x(1)} * \frac{1}{x'^{(1)}} =$$

$$= \big(x(1)x'(1)\big)^4 + e^{x(3)+x(5)+x'(3)+x'(5)} + \frac{1}{x(1)x'(1)} = f(x,x')$$