Introduction to machine learning 67577
שחר נחום 313586877
תרגיל 5:

שאלה 1:

ראינו בכיתה כי מתקיים לכל היפותזה h ולכל $\delta \in (0,1)$:

$$\mathbb{P}\left(|L_{S_{all}}(h) - L_D(h)| \leq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}\right) \geq 1 - \delta$$

ובפרט נוכל להסיק זאת לכל $h_i \in \mathcal{H}_k$:

$$\mathbb{P}\left(|L_{S_{all}}(h_i) - L_D(h_i)| \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right) \leq \frac{\delta}{|\mathcal{H}_k|}$$

נשתמש בחסם האיחוד:

$$\mathbb{P}\left(\exists i :\ |L_{S_{all}}(h_i) - L_D(h_i)| \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}\right) \leq |\mathcal{H}_k| \frac{\delta}{|\mathcal{H}_k|} \leq \delta$$

כלומר, בהסתברות $1 - \delta$ נקבל כי לכל $h_i \in \mathcal{H}_k$ מתקיים:

$$|L_{S_{all}}(h_i) - L_D(h_i)| \leq \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}$$

ולכן, בהסתברות $1 - \delta$ נקבל כי לכל $h_i \in \mathcal{H}_k$ מתקיים:

$$L_D(h^*) \leq L_{S_{all}}(h^*) + \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}$$

$$\leq L_{S_{all}}(h_i) + \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}$$

$$\leq L_D(h_i) + 2\sqrt{\frac{\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}}$$

$$\leq L_D(h_i) + \sqrt{\frac{4\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{2m}} = L_D(h_i) + \sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}$$

כנדרש.

<u>שאלה 2:</u>

כעת, בעבור *model selection*, נשתמש בסעיף הקודם, ובטענות שניתנו לנו ברמז, ונסיק:

$$\mathbb{P}\left(L_D(h^*) \leq \min_{\mathrm{h}\in\mathcal{H}} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}\right) \geq 1 - \frac{\delta}{2}$$

כעת, לכל $i \in [k]$

$$\mathbb{P}\left(L_D(h_i) \leq \min_{\mathrm{h}\in\mathcal{H}_\mathrm{i}} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}{(1-\alpha)m}}\right) \geq 1 - \frac{\delta}{2}$$

כלומר, בהסתברות של לפחות $1-\delta$:

$$L_D(h^*) \leq L_D(h_j) + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}$$

$$\leq \min_{\mathrm{h}\in\mathcal{H}_\mathrm{j}} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}$$

$$= \min_{\mathrm{h}\in\mathcal{H}_\mathrm{k}} L_D(h) + \sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}$$

כנדרש.

<u>שאלה 3:</u>

נשים לב בנוסחאות הנ"ל שבמקרה שבו $j = k$, מתקיים ששימוש ב $standard\ method$ תביא לנו חסם הדוק יותר, ואילו במקרה ש $|\mathcal{H}_i| = 2^{2^{c_i}}$ לכל $i \in [k]$ ואיזשהו קבוע c , שימוש ב- $standard\ method$ יביא לנו את החסם הבא:

$$L_{\mathcal{D}}(h^*) \leq \min_{\mathrm{h}\in\mathcal{H}_\mathrm{k}} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{m}\ln\left(\frac{2^{2^{c_k+1}}}{\delta}\right)}$$

אך אם נשתמש ב $model\ selection$, נקבל את החסם:

$$L_{\mathcal{D}}(h^*) \leq \min_{\mathrm{h}\in\mathcal{H}_\mathrm{k}} L_{\mathcal{D}}(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\left(\frac{2^{2^{c_k}}+2}{\delta}\right)} + \sqrt{\frac{2}{\alpha m}\cdot\ln\left(\frac{4k}{\delta}\right)}$$

כעת, אם נשווה את שני החסמים, נראה כי כאשר האינדקס j האופטימלי קטן משמעותית מ- k החסם שקיבלנו מה $model\ selection$ יהיה משמעותית יותר טוב:

$$\frac{\epsilon_{est}^{MS}}{\epsilon_{est}^{S}} = \frac{\sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}{(1-\alpha)m}} + \sqrt{\frac{2\ln\left(\frac{4k}{\delta}\right)}{\alpha m}}}{\sqrt{\frac{2\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}{m}}} = \sqrt{\frac{\ln\left(\frac{4k}{\delta}\right)}{\alpha\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}} + \sqrt{\frac{\ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}{(1-\alpha)\ln\left(\frac{2|\mathcal{H}_k|}{\delta}\right)}}$$

נשים לב שבבחירת $|\mathcal{H}_i|$ כך ש: $|\mathcal{H}_i| = 2^{2^{c_i}}$ לכל $i \in [k]$ ואיזשהו קבוע c, מתקיים:

$$\lim_{c\to\infty}\frac{\epsilon_{est}^{MS}}{\epsilon_{est}^{S}} = \infty$$

<u>שאלה 6:</u>

1. נשים לב כי $P(x=y) = 1-\frac{1}{a}$, וגם $P(x=ay) = \frac{1}{a}$.

   יתר על כן, $L_{\mathcal{D}}(w) = \mathbb{E}\left[l\big(w,(x,y)\big)\right]$.

   היות ופונקציית השגיאה היא מיצוע של פונקציה ריבועית, מתקיים:

   $$l\big(w(x,y)\big) = (h(x)-y)^2 = (wx-y)^2$$

   נציב,

   $$L_{\mathcal{D}}(w) = \mathbb{E}\left[l\big(w,(x,y)\big)\right] = \mathbb{E}[(wx-y)^2]$$

   $$= P(x=ay)\cdot(awy-y)^2 + P(x=y)\cdot(wy-y)^2$$

   $$= \frac{(awy-y)^2}{a} + \left(1-\frac{1}{a}\right)(wy-y)^2$$

   מכנה משותף:

   $$\frac{(awy-y)^2}{a} + \frac{(a-1)(wy-y)^2}{a}$$

$$= \frac{(awy-y)^2 + (a-1)(wy-y)^2}{a}$$

כלומר סה״כ קיבלנו:

$$L_{\mathcal{D}}(w) = \frac{(awy-y)^2 + (a-1)(wy-y)^2}{a}$$

2. נחפש את נקודת המינימום של $L_{\mathcal{D}}(w)$:

$$L_{\mathcal{D}}(w) = \frac{1}{a} \cdot \left((awy-y)^2 + (a-1)(wy-y)^2\right)$$

$$= \frac{1}{a}\left((a^2w^2y^2 - 2awy^2 + y^2) + (a-1)(w^2y^2 - 2wy^2 + y^2)\right)$$

$$L'_{\mathcal{D}}(w) = \frac{1}{a}\left((2a^2y^2w - 2ay^2) + (a-1)(2y^2w - 2y^2)\right)$$

נציב $y^2 = 1$

$$= \frac{2a^2w - 2a + (a-1)(2w-2)}{a}$$

$$= \frac{2a^2w - 2a + 2aw - 2a - 2w + 2}{a}$$

$$= \frac{2a^2w + 2aw - 2w - 4a + 2}{a}$$

$$= \frac{(2a^2 + 2a - 2)w - 4a + 2}{a}$$

נשווה לאפס,

$$\frac{(2a^2 + 2a - 2)w - 4a + 2}{a} = 0$$

$$\Rightarrow 4a - 2 = (2a^2 + 2a - 2)w$$

$$w^* = \frac{2a-1}{a^2 + a - 1}$$

כעת נוודא שזה אכן מינימום:

$$L''_{\mathcal{D}}(w) = \frac{2a^2 + 2a - 2}{a}$$

ידוע לנו כי $a > 1$, ולכן נקבל כי $L''_{\mathcal{D}}(w) > 0$, ולכן זוהי אכן נקודת מינימום.

נציב בפונקציה ונקבל:

$$f(w^*)=\frac{\left(ay\left(\frac{2a-1}{a^2+a-1}\right)-y\right)^2+(a-1)\left(\left(\frac{2a-1}{a^2+a-1}\right)y-y\right)^2}{a}$$

כאשר $a\to\infty$, נקבל כי $w^*\to 0$, ו– $f(w^*)\to 1$, שהרי נקבל

$$\lim_{a\to\infty}\frac{\left(ay\left(\frac{2a-1}{a^2+a-1}\right)-y\right)^2+(a-1)\left(\left(\frac{2a-1}{a^2+a-1}\right)y-y\right)^2}{a}$$

$$=\lim_{a\to\infty}\frac{y^2+(a-1)y^2}{a}$$

$$=1$$

3. נחקור את האפשרויות של x וההסתברויות לכך.
בהסתברות $\frac{1}{a}$:

$$x=ay\to sign(ay)\cdot\min(1,|ay|)$$

כעת, היות ו- $a>1$ וגם $|y|=1$, נקבל:

$$x=sign(y)\cdot|a|$$

ולכן, היות ו y נדגם בהתפלגות אחידה, הרי ש

$$P(x=a)=P(x=-a)=\frac{1}{a}\cdot\frac{1}{2}=\frac{1}{2a}$$

כעת, בהסתברות של $1-\frac{1}{a}$:

$$x=y\to sign(y)\cdot\min(1,|y|)$$

והרי ש $|y|=1$, ולכן $\min(1,|y|)=1$ לכל y.

כעת, היות ו y נדגם בהתפלגות אחידה, נקבל כי:

$$P(x=1)=P(x=-1)=\frac{1}{2}\cdot\left(1-\frac{1}{a}\right)=\frac{1}{2}-\frac{1}{2a}=\frac{a-1}{2a}$$

היות ופונקציית השגיאה היא מיצוע של פונקציה ריבועית, מתקיים:

$$l\big(w(x,y)\big)=(h(x)-y)^2=(wx-y)^2$$

נציב,

$$L_{\mathcal{D}}(w)=\mathbb{E}\left[l\big(w,(x,y)\big)\right]=\mathbb{E}[(wx-y)^2]$$

$$=P(x=1)\cdot(w-y)^2+P(x=-1)\cdot(-w-y)^2+P(x=a)\cdot(wa-y)^2+P(x=-a)\cdot(-wa-y)^2$$

$$=\frac{a-1}{2a}\left((w-y)^2+(-w-y)^2\right)+\frac{1}{2a}\left((wa-y)^2+(-wa-y)^2\right)$$

$$= \frac{a-1}{2a}(w^2 - 2wy + y^2 + w^2 + 2wy + y^2)$$
$$+ \frac{1}{2a}(w^2a^2 - 2awy + y^2 + w^2a^2 + 2awy + y^2)$$

נציב $y^2 = 1$:

$$= \frac{a-1}{2a}(2w^2 + 2) + \frac{1}{2a}(2w^2a^2 + 2)$$

$$= \frac{a-1}{a}(w^2 + 1) + \frac{1}{a}(w^2a^2 + 1)$$

נגזור:

$$L'_{\mathcal{D}}(w) = \frac{a-1}{a}(2w) + \frac{1}{a}(2wa^2)$$

נשווה ל – 0:

$$\frac{2(a-1)w + 2a^2w}{a} = 0$$

$$2a = 2w - 2a^2w$$
$$a = (1-a^2)w$$

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

נבדוק לפי הנגזרת השנייה שזה אכן מינימום:

$$L''_{\mathcal{D}}(w) = \frac{2(a-1)}{a} + 2a > 0$$

שהרי $a > 1$.

כעת,

$$f(w^*) = \frac{a-1}{2a}\left(2\left(\frac{a}{1-a^2}\right)^2 + 2\right) + \frac{1}{2a}\left(2\left(\frac{a}{1-a^2}\right)^2 a^2 + 2\right)$$

$$= \frac{a-1}{a}\left(\frac{a^2}{(1-a^2)^2} + 1\right) + \frac{1}{a}\left(\frac{a^4}{(1-a^2)^2} + 1\right)$$

ונקבל שכאשר $a \to \infty$ מתקיים:

$$\lim_{a\to\infty} f(w^*) = \lim_{a\to\infty} \frac{a-1}{a}\left(\frac{a^2}{(1-a^2)^2} + 1\right) + \frac{1}{a}\left(\frac{a^4}{(1-a^2)^2} + 1\right)$$

$$\lim_{a\to\infty}(a-1)\left(\frac{a}{(1-a^2)^2}\right) + \left(\frac{a^3}{(1-a^2)^2}\right) = 0$$

ואילו

$$\lim_{a\to\infty} w^* = \lim_{a\to\infty} \frac{a}{1-a^2} = 0$$

4. נגדיר כי y מתפלג אחיד על הערכים $\{\pm 1\}$ וכאשר $y = 1$ נגדיר $x = \alpha$ וכאשר $y = -1$ נגדיר $x = \beta$.
נחפש את α, β שיקיימו את אי השוויון הנדרש.

היות ופונקציית השגיאה היא מיצוע של פונקציה ריבועית, מתקיים:

$$l\big(w(x,y)\big) = (h(x) - y)^2 = (wx - y)^2$$

נציב,

$$L_{\mathcal{D}}(w) = \mathbb{E}\big[l\big(w,(x,y)\big)\big] = \mathbb{E}[(wx - y)^2]$$

$$= P(x = \alpha) \cdot (\alpha w - 1)^2 + P(x = \beta) \cdot (\beta w + 1)^2$$

$$= P(y = 1) \cdot (\alpha w - 1)^2 + P(y = -1) \cdot (\beta w + 1)^2$$

$$= \frac{1}{2} \cdot (\alpha w - 1)^2 + \frac{1}{2} \cdot (\beta w + 1)^2$$

$$= \frac{\alpha^2 w^2 - 2\alpha w + 1 + \beta^2 w^2 + 2\beta w + 1}{2}$$

$$= \frac{\alpha^2 w^2 - 2\alpha w + \beta^2 w^2 + 2\beta w + 2}{2}$$

נגזור:

$$L'_{\mathcal{D}}(w) = \frac{2\alpha^2 w - 2\alpha + 2\beta^2 w + 2\beta}{2}$$

$$= \alpha^2 w - \alpha + \beta^2 w + \beta$$

נשווה ל-0:

$$(\alpha^2 + \beta^2)w = (\alpha - \beta)$$

$$w^* = \frac{(\alpha - \beta)}{(\alpha^2 + \beta^2)}$$

כעת, נציב ב $L_{\mathcal{D}}(w)$:

$$\frac{\alpha^2\left(\frac{(\alpha-\beta)}{(\alpha^2+\beta^2)}\right)^2-2\alpha\left(\frac{(\alpha-\beta)}{(\alpha^2+\beta^2)}\right)+\beta^2\left(\frac{(\alpha-\beta)}{(\alpha^2+\beta^2)}\right)^2+2\beta\left(\frac{(\alpha-\beta)}{(\alpha^2+\beta^2)}\right)+2}{2}$$

נשים לב שבחירת $\alpha < \beta$ תעזור לנו בהקטנת הערך.

נציב $\alpha = 1, \beta = 2$:

$$L_{\mathcal{D}}(w^*) = \frac{1\cdot\frac{-1}{25}-2\cdot\frac{-1}{25}+\frac{-1}{25}+2\cdot\frac{-1}{25}+2}{2} = \frac{-\frac{2}{25}+2}{2} = 1-\frac{2}{50} < 1$$

נבצע *clipping*:

$$L_{\mathcal{D}}(w) = \mathbb{E}\left[l\left(w,(x,y)\right)\right] = \mathbb{E}[(sign(x)\cdot\min(1,|x|))^2]$$

$$= \frac{1}{2}(w\cdot sign(\alpha)\cdot\min(1,|\alpha|)-1)^2+\frac{1}{2}(w\cdot sign(\beta)\cdot\min(1,|\beta|)+1)^2$$

נציב $\alpha = 1, \beta = 2$:

$$\frac{1}{2}(w-1)^2+\frac{1}{2}(w+1)^2$$

$$L_{\mathcal{D}}(w) = \frac{1}{2}(w^2-2w+1+w^2+2w+1) = w^2+1$$

נגזור:

$$L'_{\mathcal{D}}(w) = 2w$$

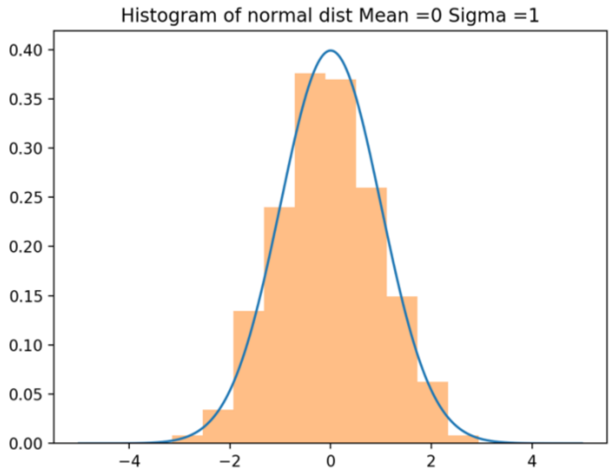
נשווה ל-0, ונקבל $w = 0$.

נציב ונקבל: $L_{\mathcal{D}}(0) = 1$.

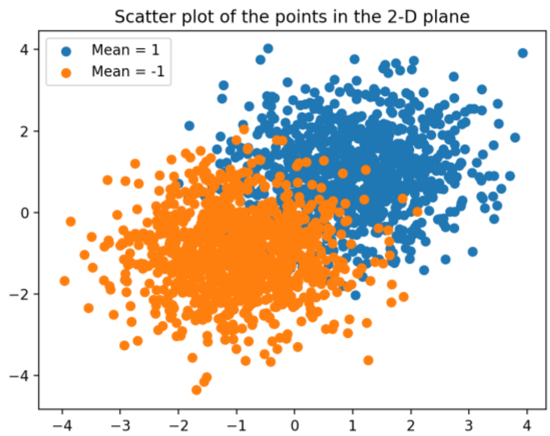
<u>ולכן קיבלנו כי התוצאה לפני ה- *clipping* קטנה מהתוצאה אחרי.</u>

שאלה 7:

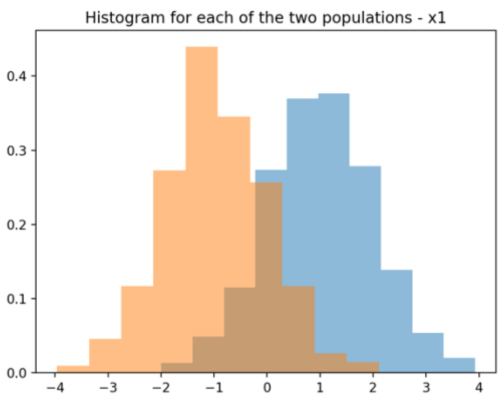
1. גרף:

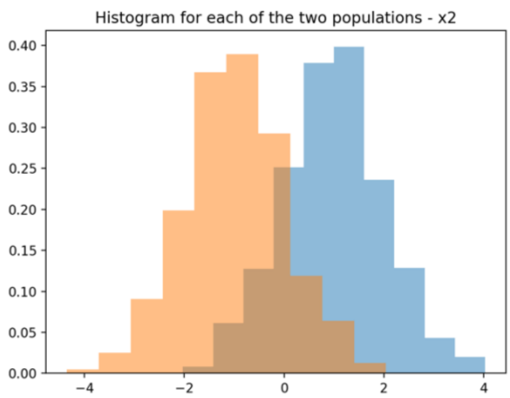


2. גרף:

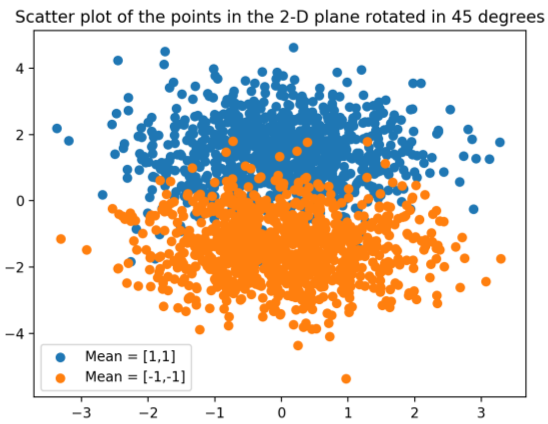


3. גרפים:

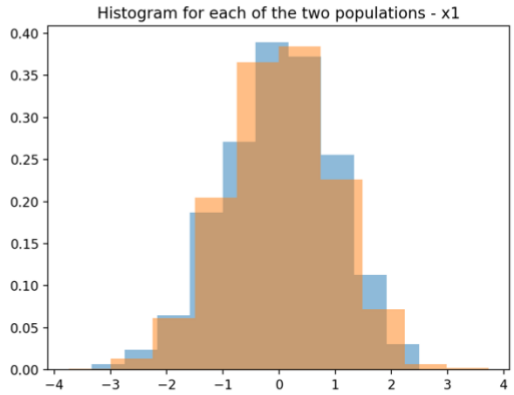


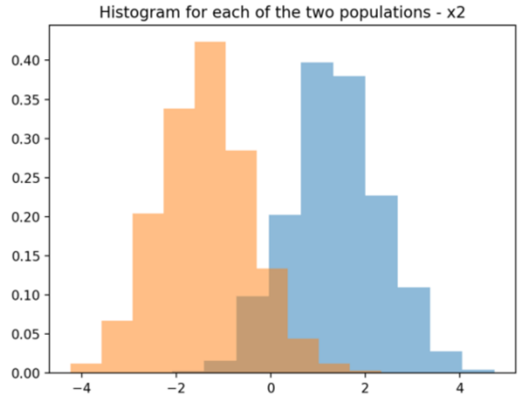


4. גרף:



גרפים:





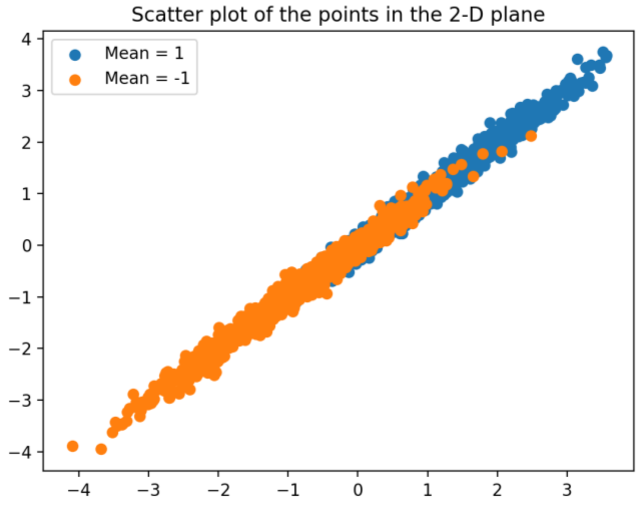
נשים לב כי הסיבוב של הדאטא מצד אחד שיבש את ההפרדה של x_1 , מצד שני היטיב עם ההפרדה של x_2.

5. נשים לב שהסיבוב כן מיטיב עם שימוש בשני הפיצ׳רים שכן במצב זה נראה כי בפיזור החיתוך בין הקבוצות קטן.

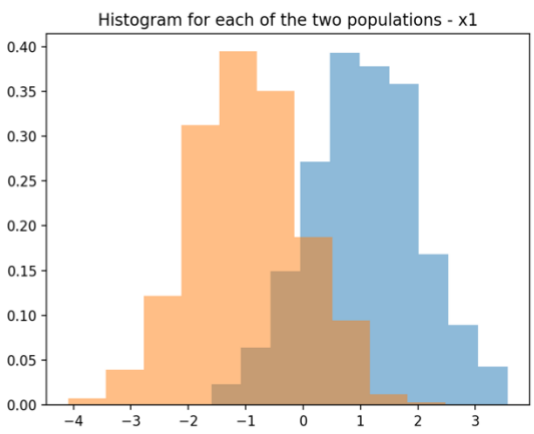
שאלה 8:

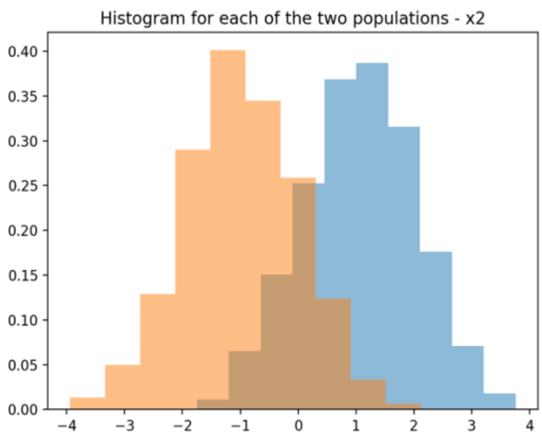
1. פיזור:

הערכים העצמיים הם $2, 0.01$ (עם חשיבות לסדר).
התוחלת:
$blue: Mean = [1,1]$
$0range: Mean = [-1,-1]$



היסטוגרמות:

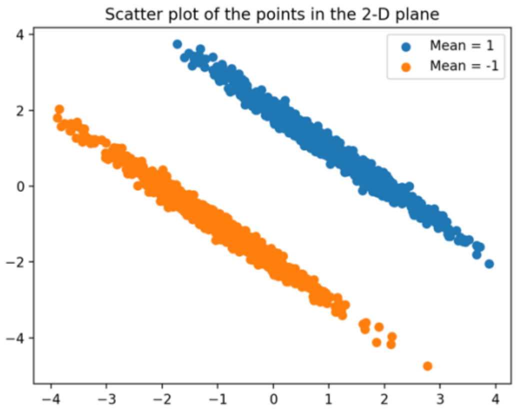




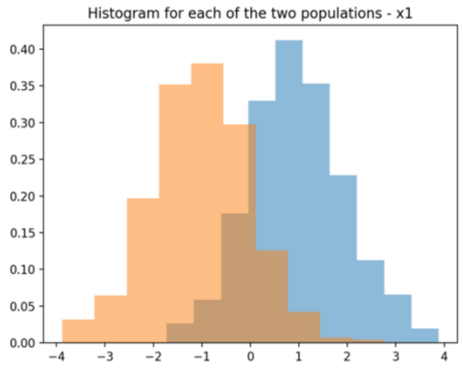
2. פיזור:

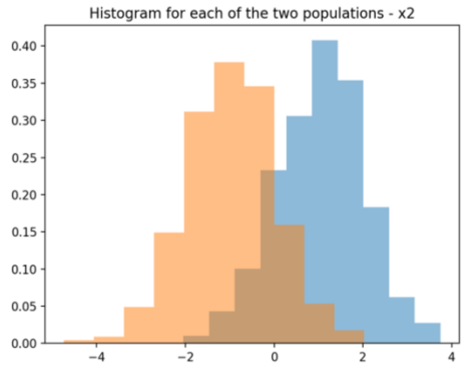
כעת הערכים העצמיים הם $0.01, 2$. (בסדר ההפוך מסעיף א׳)
התוחלת:

$$blue: Mean = [1,1]$$
$$0range: Mean = [-1,-1]$$



היסטוגרמות:



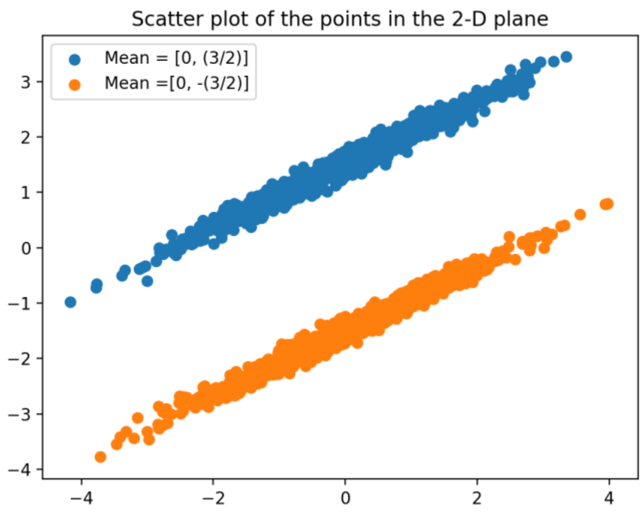


3. נשים לב כי את הדוגמה הראשונה קשה יותר להפריד מאשר את הדוגמא השנייה אשר ניתנת להפרדה מושלמת בדו מימד ע״י שני פיצ׳רים (כמו שניתן לראות בפיזור).

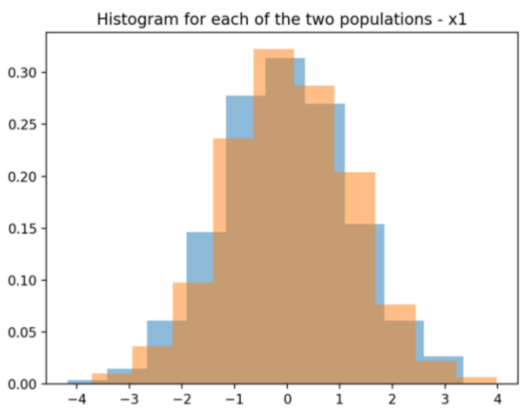
על אף שעל פניו אם מסתכלים על ההיסטוגרמות של x_1, x_2 בנפרד לא נראית לעין הפרדה בולטת מבין שני הסעיפים, כאשר שני הפיצ׳רים עובדים יחד (כפי שרואים בפיזור הנקודות רואים בבירור כי בסעיף ב׳ קיימת הפרדה ברורה בdata. ולכן, בהינתן פונקציה אשר מנקדת כל פיצ׳ר אינדיבידואלית ובאופן בלתי תלוי, לא נוכל להסתבך על התוצאות הללי כי ככל נראה היינו מקבלים כי הפיצ׳רים בשני הסעיפים היו מקבלים תוצאות דומות, ולא היינו יכולים להסיק כי קיים צירוף שנותן לנו הפרדה מלאה לעומת הצירוף השני שאינו נותן.

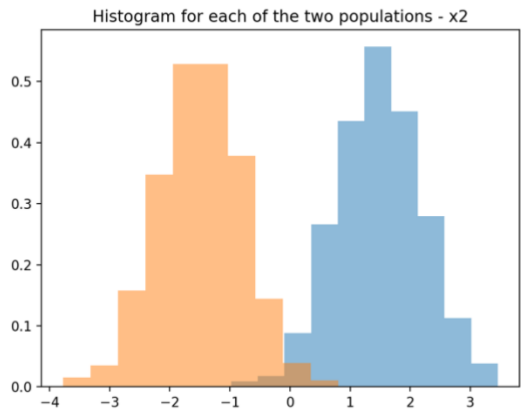
<u>שאלה 9:</u>

1. פיזור:



היסטוגרמות:



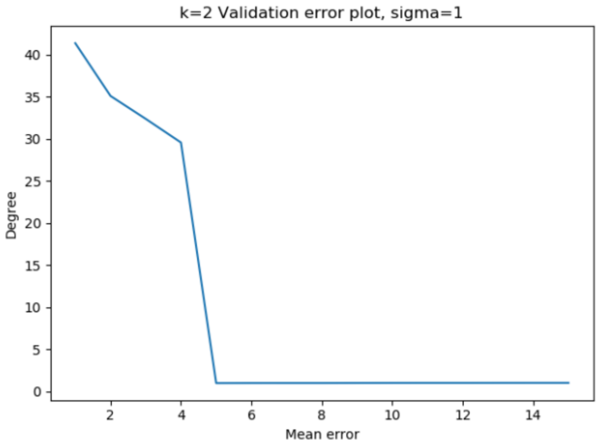


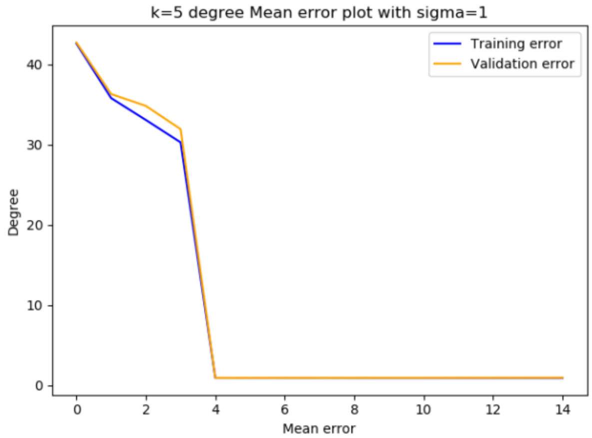
2. הסבר:

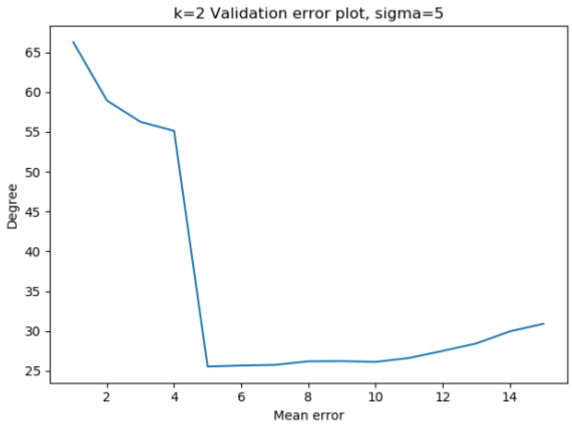
נשים לב כי בפיצ׳ר x_1 לא קיימת הפרדה כלל, ולכן אין הוא תורם לנו כלל ($completley\ useless$), מצד שני היות וב x_2 קיים חיתוך מסוים, מסתמן שכן נזדקק לשניהם ע״מ לנסות להפריד בצורה הטובה ביותר, שכן אחרת כל אחד מהם בנפרד לא יצליח לספק הפרדה מלאה.

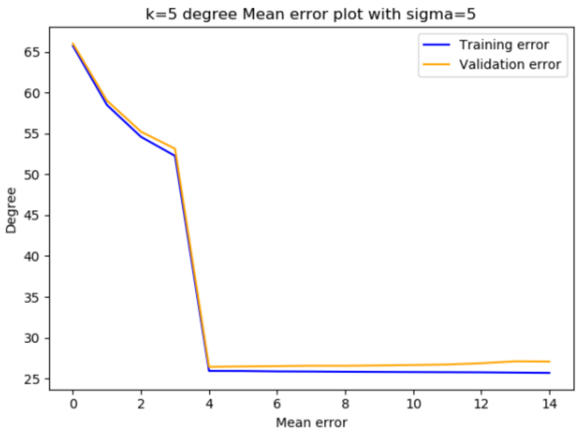
ואכן לפי הגרף בסעיף א׳ ניתן לראות שע״י שימוש בשני הפיצ׳רים נוכל לקבל הפרדה מלאה בין השניים, ולכן נקבל כי בעבור שניהם הצירוף שלהם יחד שיפר אותם (ונקבל כי x_1 השתפר נפלאות בפרט).

שאלה 10:









סעיף g:

נשים לב כי דרגת הפולינום אשר נותנת שגיאה מינימלית היא 5, ואילו יצא לי שהטעות שנמדדת בעבור ה-test set קטנה יותר (אבל לא בצורה משמעותית) מאשר הטעות שנמדדה לפני.
הסיבה לכך היא ככל הנראה שבtest ישנו פחות data, ועל כן גם פחות רעש.

סעיף f:

נשים לב שכאשר $\sigma = 5$, ישנו הבדל יותר משמעותי בין ה-validation error והה- train error, ככל הנראה בגלל שהגדלנו את השונות של הרעש, וכאשר ישנו יותר רעש אנחנו חשופים יותר ל-overfitting.