Assignment 3

<u>שאלה 1</u>

c. The classification error for kmeans with k=10 and a random sample of size 1000 was:0.465

Cluster Size	common label	% of common label
61	6	70.49%
67	0	83.58%
38	0	92.11%
99	1	68.69%
119	4	45.38%
160	3	35.62%
115	2	28.7%
148	1	39.86%
124	7	60.48%
69	6	79.71%

d. The classification error for singlelinkage with k=10 and a random sample of size 1000 was:0.869

Cluster Size	common label	% of common label
991	7	12.31%
1	6	100.0%
1	9	100.0%
1	2	100.0%
1	0	100.0%
1	8	100.0%
1	8	100.0%
1	5	100.0%
1	3	100.0%
1	9	100.0%

קל לראות שkmeans עשה עבודה משמעותית יותר טובה עבור המקרה הזה.

e. עבור kmeans :

The classification error for kmeans with k=6 and a random sample of size 1000 was:0.526

Cluster Size	common label	% of common label
224	1	50.89%
154	4	44.81%
161	6	56.52%
184	3	41.3%
86	0	87.21%
191	7	30.37%

עבור singlelinkage:

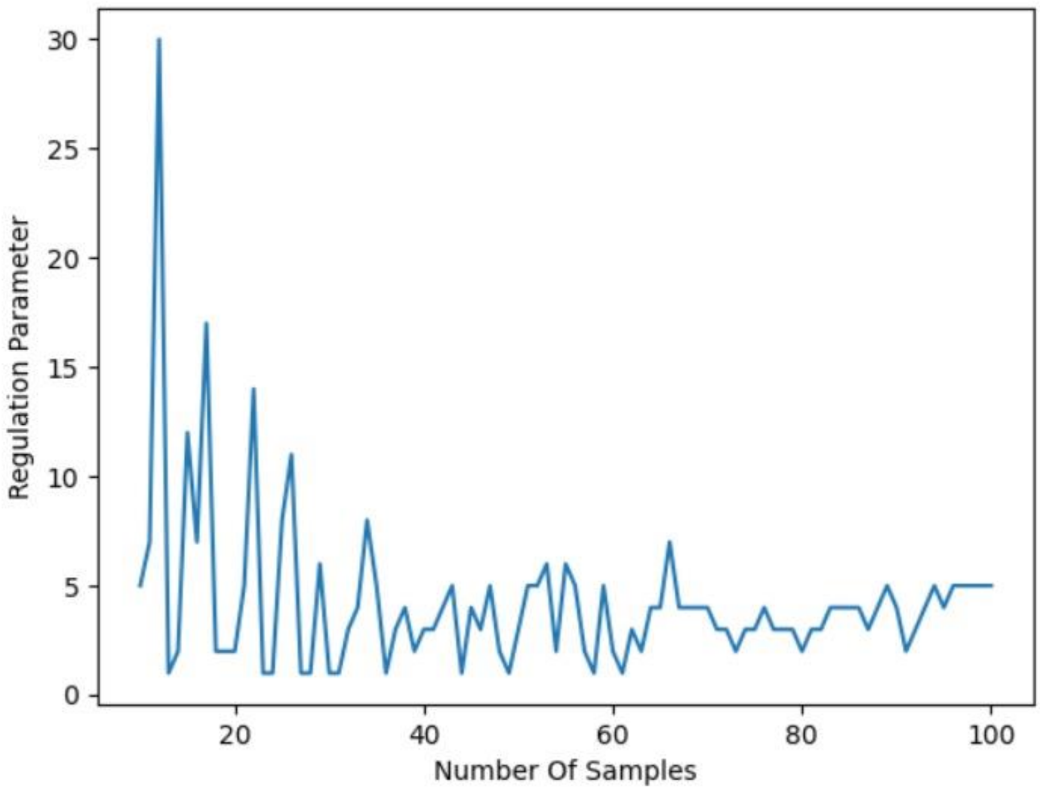
The classification error for singlelinkage with k=6 and a random sample of size 1000 was:0.882

Cluster Size	common label	% of common label
995	4	12.16%
1	5	100.0%
1	2	100.0%
1	4	100.0%
1	8	100.0%
1	7	100.0%

עבור אלגוריתם kmeans, בכך שעברנו מk=10 לk=6, האלגוריתם קיבץ מספרים דומים לאותו קלאסטר (אנו משערים שיש כמות גדולה של 8 בתוך הקלאסטר המתויג כ6 מכיוון שהם דומים). לכן קיבלנו 6 קבוצות שבתוכן האחוז עם התווית הנפוצה ביותר היא קטנה יותר מאשר עם k=10.

עבור singlelinkage, גם במקרה הזה אנו מקבלים כי כמעט כל הנקודות קובצו לאותו קלאסטר, ולכן אין הבדל משמעותי וחשוב לדבר עליו במעבר בין k=10 לk=6.

שאלה 2



a.

b. כיוון שככל שיש יותר דגימות יותר קשה לייצר overfitting, ציפינו שככל שמספר הדגימות יעלו ידרש פרמטר רגולציה קטן יותר כדי לא לייצר overfitting על הדטא (כיוון שגם לא נרצה להגיע למצב של underfitting אם הוא יהיה סתם גדול מידי).

c. הטרנד התרחש אבל לא בצורה חלקה, זה יכול לנבוע מהמון סיבות, לדוגמה יכול להיות שהדטא של האימון דומה מאוד לדטא של הטסט ואז קצת overfitting עליו לא ממש פוגע כיוון שמדובר במספר דגימות יחסית קטן. אם היינו מריצים על כמות גדולה יותר של דטא מספר רב של פעמים עם קרוס ולידישן זה כנראה היה פותר את הבעיה.

d. אנו יודעים כי עבור שגיאה אבסולוטית הוא יהיה $h_{bayes}(x) = Median[Y|X = x] = Median[\langle w, x\rangle + \eta|X = x] = Median[\langle w, x\rangle|X = x]$ כיוון שהחציון של η=0
אנו יודעים כי עבור שגיאה ריבועית הוא יהיה $h_{bayes}(x) = E[Y|X = x] = E[\langle w, x\rangle + \eta|X = x] = E[\langle w, x\rangle|X = x] + E[\eta|X = x] = E[\langle w, x\rangle|X = x]$. כיוון שהתוחלת של η=0

שאלה 3

a. כפי שלמדנו, באלגוריתם gradient decent, $w^{(t+1)}(w^{(t)}, \eta) = w^{(t)} - \eta\nabla f(w^{(t)})$.
נחשב את נגזרת $f(w) = \lambda\|w\| + \sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2$.
נגדיר $g(w) = \lambda\|w\|$, ו- $z_i(w) = (\langle w, x_i\rangle - y_i)^2$. נקבל כי

$$f(w) = g(w) + \sum_{i=1}^{m} z_i(w)$$

$$\nabla g(w) = \nabla(\lambda\|w\|) = \nabla\left(\lambda\sqrt{\sum_{i=1}^{d} w_i^2}\right) = \frac{\lambda}{2}\left(\sum_{i=1}^{d} w_i^2\right)^{-\frac{1}{2}} \nabla\left(\sum_{i=1}^{d} w_i^2\right) = \frac{\lambda 2w}{2\|w\|} = \frac{\lambda w}{\|w\|}$$

$$\nabla z_i(w) = \nabla((\langle w, x_i\rangle - y_i)^2) = 2(\langle w, x_i\rangle - y_i)x_i$$

נקבל כי $\nabla \mathrm{f(w)} = \nabla g(w) + \sum_{i=1}^{m} \nabla z_i(w) = \frac{\lambda w}{\|w\|} + \sum_{i=1}^{m} 2(\langle w, x_i\rangle - y_i)x_i$

לכן, $\mathrm{w}^{(\mathrm{t+1})}\left(w^{(t)}, \eta\right) = w^{(t)} - \eta\left(\frac{\lambda w^{(t)}}{\|w^{(t)}\|} + \sum_{i=1}^{m} 2\left(\langle w^{(t)}, x_i\rangle - y_i\right)x_i\right)$

b. בstochastic gradient decent, למדנו כי הצעד מתבצע כך על i שנבחר אקראית על המדגם $\{1, \dots, \mathrm{m}\}$

$$\mathrm{w}^{(\mathrm{t+1})}\left(w^{(t)}, \eta\right) = w^{(t)} - \eta\left(\nabla R\left(w^{(t)}\right) + \nabla l\left(w^{(t)}, (x_i, y_i)\right)\right)$$

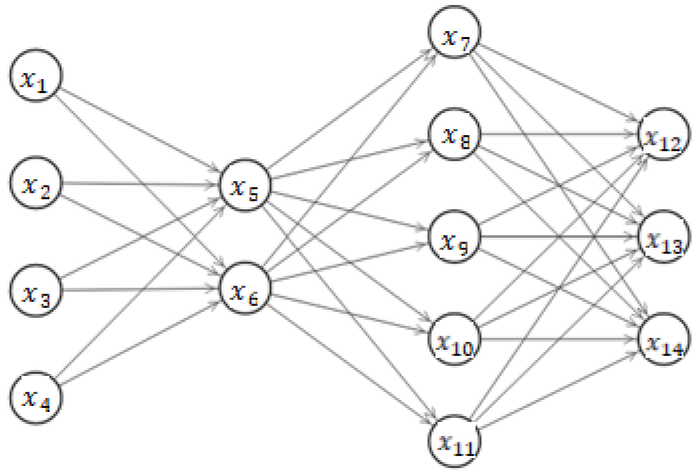
עבור פונקציית המטרה הזו נסמן: $l\left(w, (x_i, y_i)\right) = (\langle w, x_i\rangle - y_i)^2 = z_i(w), R(w) = \lambda\|w\| = g(w)$

לכן מחישוב גרדיאנטים שעשינו סעיף קודם נקבל כי-

$$\mathrm{w}^{(\mathrm{t+1})}\left(w^{(t)}, \eta\right) = w^{(t)} - \eta\left(\frac{\lambda w^{(t)}}{\|w^{(t)}\|} + 2\left(\langle w^{(t)}, x_i\rangle - y_i\right)x_i\right)$$

שאלה 4

.a



.b מרחב הקלט X הינו $\mathbb{R}^4$, מכוון שיש 4 כניסות בשכבה הראשונה.

.c על פי o_i $\psi(o_1, \dots, o_k) = argmax_{i\in[k]}$ שלמדנו בכיתה, התוויות האפשריות הן $output \in \{1,2,3\}$.

.d נגדיר את משקלה של הקשת המחברת בין x_i לx_j להיות $w_{i,j}$.

בנוסף נגדיר את o_I להיות הערך היוצא מx_i.

עבור $i \in \{1,4\}$, $o_i = x_i$

עבור $i \in \{5,6\}$, $o_i = \sigma\left(\sum_{j=1}^{4} w_{j,i} \cdot o_j\right)$

עבור $i \in \{7,11\}$, $o_i = \sigma\left(\sum_{j=5}^{6} w_{j,i} \cdot o_j\right)$

עבור $i \in \{12,14\}$, $o_i = \sigma\left(\sum_{j=7}^{11} w_{j,i} \cdot o_j\right)$

לדוגמה עבור $i = 12$,

$$o_{12} = \sigma\left(\sum_{j=7}^{11} w_{j,12} \cdot \sigma\left(\sum_{k=5}^{6} w_{k,j} \cdot \sigma\left(\sum_{z=1}^{4} w_{z,k} \cdot x_z\right)\right)\right)$$

לכן מחלקת ההיפותזות שארכיטקטורת הרשת הזו מתארת היא:

$$\mathcal{H} = \{h_w(x), w \in \mathbb{R}^{33}\}$$

$$h_w\big((x_1, x_2, x_3, x_4)\big) = \psi(o_{12}, o_{13}, o_{14}) =$$

$$argmax_{i\in\{12,13,14\}}\left(\sigma\left(\sum_{j=7}^{11} w_{j,i} \cdot \sigma\left(\sum_{k=5}^{6} w_{k,j} \cdot \sigma\left(\sum_{z=1}^{4} w_{z,k} \cdot x_z\right)\right)\right)\right)$$

שאלה 5

a. ראשית נשים לב כי בעץ שעומקו לכל היותר n יש לכל היותר $2^n - 1$ צמתים ועלים שונים.
מכיוון שכל צומת בעץ היא מהצורה $x(i) \geq \theta, \theta \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$, לכל צומת יש 3 אפשרויות שונות עבור כל קואורדינטה של x. מכיוון ש$Y = \{0,1\}$, לכל עלה יש 2 אופציות אפשריות. לכן לכל צומת או עלה יש $3d + 2$ אפשרויות שונות להיות.
לכן לכל עץ יש לכל היותר $2^n - 1$ צמתים ועלים, ולכל אחד יש $3d + 2$ אופציות שונות, לכן בסך הכל נקבל:

$$|\bar{\mathcal{H}}_n| \leq (3d+2)^{2^n - 1} \leq (3d+2)^{2^{n+1}}$$

b. למדנו בכיתה כי לבצע ERM על עצי החלטה זה NP-hard, בפרט אלגוריתם ID3 עם pruning הוא אלגוריתם חמדן שלא בהכרח ממזער את השגיאה על המדגם. לכן אפילו בשימוש בm גדול ככל שנרצה, לא נוכל להשתמש בחסמי PAC שלמדנו ולא נוכל להבטיח כי בהסתברות של $1 - \delta$

$$err(T_S, D) \leq \inf_{T \in \mathcal{H}_k} err(T, D) + \epsilon$$

שאלה 6

a. הנחת Naïve-Bayes לא מתקיימת, נראה כי לא מתקיימת אי תלות בהינתן התווית בין המשתנים.

$$P[x(1)=-1|y=-1]=$$
$$P[x(1)=-1\wedge x(2)=-1|y=-1]+P[x(1)=-1\wedge x(2)=1|y=-1]=$$
$$\frac{5}{60}*\frac{60}{20}+0=\frac{5}{20}$$
$$P[x(2)=1|y=-1]=$$
$$P[x(2)=1\wedge x(1)=-1|y=-1]+P[x(2)=1\wedge x(1)=1|y=-1]=$$
$$0+\frac{4}{60}*\frac{60}{20}=\frac{4}{20}$$

עכשיו נשים לב כי-

$$P[x=(-1,1)|y=-1]=0$$
$$\prod_{i=1}^{2}P[X(i)=x(i)|Y=y]=P[X(1)=-1|Y=-1]\cdot P[X(2)=1|Y=-1]=$$
$$\frac{5}{20}\cdot\frac{4}{20}=\frac{1}{20}\neq 0=P[x=(-1,1)|y=-1)$$

b. כפי שלמדנו בכיתה, האלגוריתם *naïve-bayes* מחזיר –

$$h_{bayes}(x)=\underset{y\in Y}{\operatorname{argmax}}\,P(Y=y)\prod_{i=1}^{d}P[X(i)=x(i)|Y=y]$$

נחשב את הנתונים הנחוצים על ההסתברויות.

$$P(Y=1)=\frac{6}{60}+\frac{24}{60}+\frac{2}{60}+\frac{8}{60}=\frac{2}{3}\ ,P(Y=-1)=1-\frac{2}{3}=\frac{1}{3}$$
$$P[X(1)=1|Y=1]=\frac{5}{20}\ ,P[X(2)=1|Y=1]=\frac{16}{20}$$
$$P[X(1)=-1|Y=1]=\frac{15}{20}\ ,P[X(2)=-1|Y=1]=\frac{4}{20}$$
$$P[X(1)=1|Y=-1]=\frac{15}{20}\ ,P[X(2)=1|Y=-1]=\frac{4}{20}$$
$$P[X(1)=-1|Y=-1]=\frac{5}{20}\ ,P[X(2)=-1|Y=-1]=\frac{16}{20}$$

עכשיו נחשב עבור כל *x* את ערך הפונקציה h_{bayes} .

$$h_{bayes}^{S}(1,1)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{3}*\frac{5}{20}*\frac{16}{20},\frac{1}{3}*\frac{15}{20}*\frac{4}{20}\right)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{15},\frac{1}{20}\right)=1$$
$$h_{bayes}^{S}(-1,1)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{3}*\frac{15}{20}*\frac{16}{20},\frac{1}{3}*\frac{5}{20}*\frac{4}{20}\right)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{5},\frac{1}{60}\right)=1$$
$$h_{bayes}^{S}(1,-1)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{3}*\frac{5}{20}*\frac{4}{20},\frac{1}{3}*\frac{15}{20}*\frac{16}{20}\right)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{1}{30},\frac{1}{5}\right)=-1$$
$$h_{bayes}^{S}(-1,-1)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{2}{3}*\frac{15}{20}*\frac{4}{20},\frac{1}{3}*\frac{5}{20}*\frac{16}{20}\right)=\underset{y\in\{1,-1\}}{\operatorname{argmax}}\left(\frac{1}{10},\frac{1}{15}\right)=1$$

לכן הפונקציה שנקבל ממדגם שמקיים בדיוק את ההסתברויות האלו היא:

$$h_{naive-bayes}^{S}(x)=\begin{cases}1, & x=(1,1)\\ 1, & x=(-1,1)\\ -1, & x=(1,-1)\\ 1, & x=(-1,-1)\end{cases}$$

שאלה 7

א

אנו יודעים כי $x_t(3) = 3x_t(1) + x_t(2),\ \ x_t(4) = 2x_t(2) - 4x_t(3) = -2x_t(2) - 12x_t(1)$

לכן אם היינו פרושים את תוצאות הניסוי במרחב שנפרש על ידי $x_t(1), x_t(2)$, היינו רואים כי דרגת המטריצה A היא 2, ולכן שתי הערכים העצמיים הנמוכים שלה יהיו אפסים.

כלומר ניתן לפרוש את המידע של תוצאות ניסוי זה רק בעזרת $x_t(1), x_t(2)$, כלומר במימד 2, מבלי לאבד מידע. שזוהי הגדרת העיוות, כלומר העיוות יהיה 0.

ב

נבחר את עמודות 1,2:

X1	*X2*
3	*1*
4	*2*
5	*8*
1	*4*

עבורן המטריצה X היא (בהתאם לקשרים שהוגדרו):

x1	x2	x3	x4
3	1	10	49
4	2	24	400
5	8	537	283024
1	4	65	4096

לכן X^T היא:

$$\begin{pmatrix} 3 & 4 & 5 & 1 \\ 1 & 2 & 8 & 4 \\ 10 & 24 & 537 & 65 \\ 49 & 400 & 283024 & 4096 \end{pmatrix}$$

נחשב את A:

$$\begin{pmatrix} 3 & 4 & 5 & 1 \\ 1 & 2 & 8 & 4 \\ 10 & 24 & 537 & 65 \\ 49 & 400 & 283024 & 4096 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 & 10 & 49 \\ 4 & 2 & 24 & 400 \\ 5 & 8 & 537 & 283024 \\ 1 & 4 & 65 & 4096 \end{pmatrix} = \begin{pmatrix} 51 & 55 & 2876 & 1420963 \\ 55 & 85 & 4614 & 2281425 \\ 2876 & 4614 & 293270 & 152260218 \\ 1420963 & 2281425 & 152260218 & 80119524193 \end{pmatrix}$$

כעת נחשב את הע"ע של A:

Eigenvectors for the matrix A:

- $v \approx \begin{pmatrix} -1153.557 \\ 10100.241 \\ -666.775 \\ 1 \end{pmatrix}$, eigenvalue $\lambda_1 \approx +0.000$
- $v \approx \begin{pmatrix} 12526.861 \\ 1386.876 \\ -663.888 \\ 1 \end{pmatrix}$, eigenvalue $\lambda_2 \approx 18.103$
- $v \approx \begin{pmatrix} -23.706 \\ -37.394 \\ -525.420 \\ 1 \end{pmatrix}$, eigenvalue $\lambda_3 \approx 3940.358$
- $v \approx \begin{pmatrix} +0.000 \\ +0.000 \\ 0.002 \\ 1 \end{pmatrix}$, eigenvalue $\lambda_4 \approx 80119813640.540$

ניתן לראות כי שני הערכים הנמוכים שלה הם: 0,18.103 לכן העיוות יהיה $0 + 18.103 = 18.103 > 0$.

<u>שאלה 8</u>

<u>א</u>

אנו יודעים כי:

$$\Theta_0 = \{\theta \in \Theta \mid \theta(1) = 3\theta(2)\}, r\, \theta \in \Theta, \theta(1) + \theta(2) + \theta(3) = 1.$$

לכן $\theta(3) = 1 - 4\theta(2)$

נחשב: $P_{S'D}[S' = S] = \prod_{i=1} P_{xD}[X = x] = \prod_{i=1} (\theta(1)I[x_i = 0] + \theta(2)I[x_i = 1] + \theta(3)I[x_i = 2]) =$

$$\prod_{i=1} (3\theta(2)I[x_i = 0] + \theta(2)I[x_i = 1] + (1 - 4\theta(2))I[x_i = 2]) =$$

$$\prod_{i=1} (3\theta(2)^{I[x_i=0]} * \theta(2)^{I[x_i=1]} * ((1 - 4\theta(2))^{I[x_i=2]})) =$$

נמיר ללוג L:

$$\sum_{i=1} I[x_i = 0] * \log\big(3\theta(2)\big) + \sum_{i=1} I[x_i = 1] * \log\big(\theta(2)\big) + \sum_{i=1} I[x_i = 2] * \log\big(1 - 4\theta(2)\big)$$

נגזור לפי θ:

$$\sum_{i=0} \frac{I[x_i = 0]}{\theta(2)} + \frac{I[x_i = 1]}{\theta(2)} - \frac{4I[x_i = 2]}{1 - 4\theta(2)}$$

ונשווה ל-0 כנדרש:

$$\sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{\theta(2)} - \frac{4I[x_i = 2]}{1 - 4\theta(2)} = 0 = \sum_{i=0} I[x_i = 0 \lor x_i = 1] * \big(1 - 4\theta(2)\big) - 4I[x_i = 2] * \theta(2)$$

$$\to \sum_{i=0} I[x_i = 0 \lor x_i = 1] * \big(1 - 4\theta(2)\big) = \sum_{i=0} 4I[x_i = 2] * \theta(2) \to$$

$$\hat{\theta}(2) = \sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{4(I[x_i = 0 \lor x_i = 1] + I[x_i = 2])}$$

לכן, על פי הקשרים שהראנו:

$$\hat{\theta}(1) = 3(\sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{4(I[x_i = 0 \lor x_i = 1] + I[x_i = 2])}), \hat{\theta}(3) = 1 - 4(\sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{4(I[x_i = 0 \lor x_i = 1] + I[x_i = 2])})$$

לכן:

$$\hat{\theta} = \left(3\left(\sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{4(I[x_i = 0 \lor x_i = 1] + I[x_i = 2])}\right), \sum_{x_i=0} \frac{x_i}{3} + \sum_{x_i=1} x_i - \sum_{x_i=2} \frac{x_i}{4}, 1 - 4\left(\sum_{i=0} \frac{I[x_i = 0 \lor x_i = 1]}{4(I[x_i = 0 \lor x_i = 1] + I[x_i = 2])}\right)\right)$$

ב

נזכור כי עבור $N(\mu, \sigma^2)$, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

עבור ערבוב ההתפלגויות נקבל התפלגות - $X \sim D\big(p_1, \dots, p_k, \sigma_1^2, \dots, \sigma_k^2\big)$, כאשר $\sum_{i=1}^k p_i = 1$.
בעלת פונקציית הצפיפות הבאה:

$$f_{(p_1,\dots,p_k,\sigma_1^2,\dots,\sigma_k^2)}(x) = \sum_{i=1}^k p_i \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{\frac{-(x-1)^2}{2\sigma_i^2}}$$

$$\Theta = \{p_1, \dots, p_k, \sigma_1^2, \dots, \sigma_k^2 \big| \sum_{i=1}^k p_i = 1 \land \ \forall i \in [1, \dots, k], 0 \le p_i \le 1 \land \sigma_1^2 > 0\}$$

עבור צפיפות זו S, ו $\theta \in \Theta$, הlog-likelihood הינו-

$$L(S;\theta) = \sum_{i=1}^{m} \log\left(\sum_{i=1}^{k} p_i \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{\frac{-(x-1)^2}{2\sigma_i^2}}\right)$$

עבור z נתון, הצפיפות המאוחדת היא

$$g_\theta(x,z) = p_z \cdot \frac{1}{\sqrt{2\pi\sigma_z^2}} \cdot e^{\frac{-(x-1)^2}{2\sigma_z^2}}$$

עבור צפיפות זו S, Z, ו $\theta \in \Theta$, הlog-likelihood הינו-

$$L(S,Z;\theta) = \sum_{i=1}^{m} \log\left(p_{z_i} \cdot \frac{1}{\sqrt{2\pi\sigma_{z_i}^2}} \cdot e^{\frac{-(x_j-1)^2}{2\sigma_{z_i}^2}}\right) = \sum_{i=1}^{m} \log(p_{z_i}) - \log\left(\sqrt{2\pi\sigma_{z_i}^2}\right) - \frac{(x_j-1)^2}{2\sigma_z^2}$$

נגדיר פונקציות אינדיקטורים - $\forall 1 \le i \le k: \mathrm{I}_i(j) = \begin{cases} 1, & Z(j) = i \\ 0, & \text{אחרת} \end{cases}$

כעת נגזור את $L(S,Z;\theta)$ לפי σ_i^2:

$$\frac{\partial L(S,Z;\theta)}{\partial \sigma_i^2} = \frac{\partial \sum_{j=1}^{m} \left(\log\left(p_{z_j}\right) - \log\left(\sqrt{2\pi\sigma_{z_j}^2}\right) - \frac{(x_j-1)^2}{2\sigma_{z_j}^2}\right) * \mathrm{I}_i(j)}{\partial \sigma_i^2} =$$

(נסמן את כמות הפעמים שz_i מופיע בZ, כלומר $\eta_i = \sum_{j,\mathrm{I}_i(j)=1} 1$)

$$= \frac{\eta_i}{2\sigma_i^2} - \frac{\sum_{j,\mathrm{I}_i(j)=1} (x-1)^2}{2\sigma_i^4}$$

נשווה ל0

$$\frac{\eta_i}{2\sigma_i^2} - \frac{\sum_{j,\mathrm{I}_i(j)=1} (x-1)^2}{2\sigma_i^4} = 0$$

$$\sigma_i^2 = \frac{\sum_{j,\mathrm{I}_i(j)=1} (x-1)^2}{\eta_i}$$