## שאלה 1

במודל רגרסיה לוגיסטית שהותאם לתיאור המשתנה $Y=\{0,1\}$ באמצעות המשתנים $X_1,X_2,X_3$ התקבלו המקדמים הבאים:

$$\beta_0=0.66, \beta_1=-1.72, \beta_2=2.84$$

כמו כן ידוע כי עבור **בתוספת** של יחידה אחת לערכו של $X_3$ **קטן** יחס הסיכויים (Odds) ב-30%.

4. ההסתברות $P(Y=1|X_1=2,X_2=0,X_3=1)$ נמצאת באינטרוול:

$$\text{Odds ratio}=\frac{p(Y=1|X)}{P(Y=0|X)}=$$

פתרון:

ראשית נמצא את ערך המקדם $\beta_3$. מן הנתונים ידוע שיחס הסיכויים קטן ב 30% עבור שינוי ביחידה אחת. לפי הנלמד בכיתה המשמעות של נתון זה היא: $e^{\beta_3}=0.7 \Rightarrow \beta_3=\ln(0.7)=-0.357$.

חישוב ההסתברות מתבצע לפי הצבה בנוסחא:

$$P(Y=1|X_1=2,X_2=0,X_3=1)=\frac{1}{1+e^{-0.66+1.72\times2-2.84\times0+0.357\times1}}=\frac{1}{24.03}=0.04$$

## שאלה 2

רני רוצה לחזות את ערכו של המשתנה הרציף $Y$. לרני נתונים על $Y$ ועל שלושה משתנים רציפים נוספים $X_1,X_2,X_3$ שלדעתו יש קשר בינהם לבין $Y$. רני בוחן שלושה מודלים שונים ומקבל את הפלטים הבאים עבור מדגם בן 25 תצפיות:

עבור המודל $Y=\beta_0+\beta_3\cdot X_3$ (מודל I) מתקבל כי $R^2=0.89$

עבור המודל $Y=\beta_0+\beta_1\cdot X_1$ (מודל II) התקבל כי $SSE=138.2$

עבור המודל $Y=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2$ (מודל III) התקבל כי $SSR=1589.3$ ו $SST=1710.6$

כמו כן עבור כל אחד מהמודלים התקבל כי כל המשתנים המסבירים בו מובהקים.

איזה מהמודלים יבחר רני? הסבר תשובתך והצג את חישוביך.

רמז: ערך השברון $F_{(0.95)(1,22)}=4.30$, $F_{(0.99)(1,22)}=7.94$

**פתרון**

ראשית נשלים את הפרמטרים שניתן להשלים. עבור מודל III נתון כי $SST=1710.6$, אך נתון זה לעשה אינו תלוי מודל וממנו ניתן להסיק עבור מודל II כי $SSR=SST-SSE=1572.4$ ומכאן כי $R^2=\frac{SSR}{SST}=0.9192$.

כמו כן עבור מודל III מתקבל כי $SSE=SST-SSR=121.3$ וגם $R^2=0.9291$.

בשלב זה ניתן לבצע השוואה בין ערכי $R^2$ של מודלים I ו-II ולהסיק כי מעדיף את מודל II על פני מודל I.

כעת נרצה לבחור בין מודל II למודל III, נעשה זאת ע"י שימוש במבחן F להשוואת מודלים.

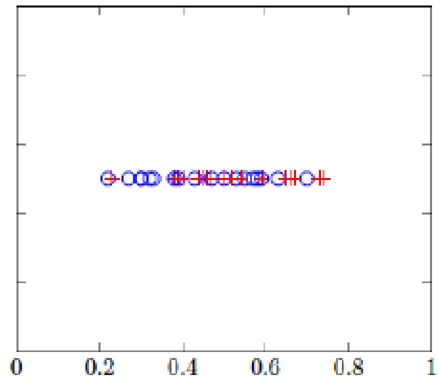
$$\frac{(SSR(p+q)-SSR(p))/q}{SSE(p+q)/(n-p-q-1)}\sim F(q,n-p-q-1)$$

ובמקרה שלנו $\frac{(1589.3-1572.4)/1}{108.3/22}\sim F(1,22)=3.433$

היות וערך זה קטן מערך הסטטיסטי המתאים לשברון ה-0.95 התוספת של $X_2$ במעבר מהמודל השני לשלישי אינה מובהקת. על כן יבחר רני במודל השני לתיאור הנתונים.

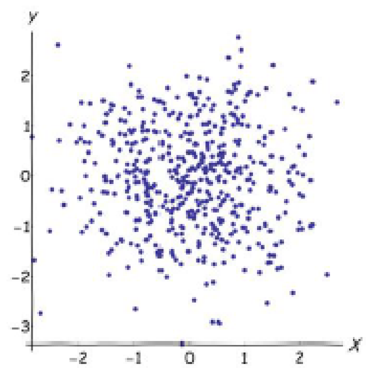
## **שאלה 3**

ב. בתמונה נתונות תצפיות דו-מימדיות משתי מחלקות. האם הפעלת PCA והטלה על הוקטור העצמי הראשון יוריד את מימד הבעיה? הסבר. (5 נק')




מכיוון שהתצפיות ממוקמות כולן על קו ישר, כל השונות נמצאת לאורך הקו הזה וביצוע PCA יחזיר תצפיות ממימד הזהה למימד התצפיות דה-פקטו.

ג. בתמונה נתונות תצפיות דו-מימדיות משתי מחלקות. בהפעלת PCA על הנתונים, מה היחס שהיינו מצפים לראות בתמונה הבאה בין כמות השונות שנתפסה על ידי הרכיב הראשון לבין הרכיב השני? הסבר. (5 נק')



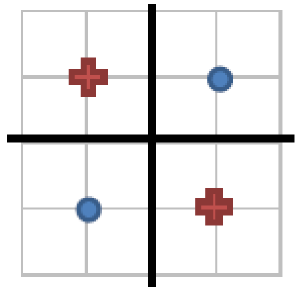

התצפיות מפוזרות באופן שבו השונות בכל אחד מהכיוונים דומה ומכאן ששני הרכיבים צפויים "לתפוס" בערך את אותו אחוז מהשונות ולכן היחס הינו 1.

## שאלה 4

ברצוננו לחשב את פונקציית ה XOR באמצעות SVM. במקום לעבוד עם קלט בינארי נעבור לעבוד עם משתני קלט $X_1, X_2 \in \{-1,1\}$ כאשר 1 נשאר 1 ו-0 ממופה ל 1-.

תזכורת: פונקציית XOR הינה פונקציית "או-אקסקלוסיבי", כלומר. הפונקציה מחזירה 1 כאשר **בדיוק** אחד משני הקלטים שלה הוא 1.

א. צייר את ארבע הנקודות המתאימות במרחב ואת הקלסיפיקציה של שתי התוצאות האפשריות. (5 נק')

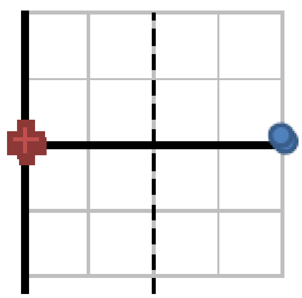


ב. תאר את התוצאה שתתקבל כאשר נריץ SVM על הקלט שציירת בא'. (5 נק')

<u>לא ניתן להפריד את התצפיות באמצעות משטח הפרדה לינארית.</u>

ג. נרצה לבצע טרנספורמציה על הנתונים ולהריץ SVM על התצפיות לאחר הטרנספורמציה. תן דוגמא לפונקציית טרנספורמציה $\phi: R^2 \rightarrow R^2$ שתאפשר יצירת משטח הפרדה לינארית באמצעות SVM. צייר את התצפיות במרחב החדש ואת המשטח שיתקבל ע"י SVM. (5 נק')

$$\phi(x_1, x_2) = ((x_1 + x_2)^2, 0)$$



ד. האם ניתן למדל את פונקציית XOR באמצעות פרספטרון יחיד (רשת נוירונים ללא שכבות נסתרות ובעלת פלט יחיד)? הסבר או הדגם (5 נק')

<u>ניתן לבנות פרספטרון עם פונקציית הפעלה ריבועית ונקבל הפרדה מוחלטת כאשר שתי המשקולות שוות ל-1:</u>

$$y = (x_1 + x_2)^2$$

**שאלה 5**

ברגרסיה לוגיסטית המנבאת משתנה פלט $Y \in \{0,1\}$ באמצעות 3 משתני קלט רציפים $X_1, X_2, X_3$ התקבל וקטור המקדמים הבא: $\beta = (0.5, 1.21, -0.7, 0.3)$. לאחר התאמת המודל מנבאים באמצעותו את 3 התצפיות הבאות:

| מספר תצפית | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| 1 | 1 | 0 | 0.5 |
| 2 | 3 | 5 | 1 |
| 3 | 0.1 | 1 | 7 |

1. נסמן ב $p_i$ את ההסתברות שהתצפית ה-$i$ תקבל ערך 1 (לפי המודל). עבור שלושת התצפיות לעיל מתקיים:

א. $p_1 > p_2 > p_3$

ב. $p_3 > p_2 > p_1$

ג. $p_3 > p_1 > p_2$

ד. $p_2 > p_1 > p_3$

ה. לא ניתן לקבוע את יחס הסדר בין ההסתברויות מבלי להתייחס לערך $\tau$ ספציפי.

**פתרון: ג'**

לפי מודל הרגרסיה הלוגיסטית מתקיים

$$P(Y = 1|x) = \frac{1}{1 + e^{-\beta^T x}}$$

נחשב אם כן את ההסתברות המתקבלת לכל אחת מהתצפיות:

$$p_1 = P(Y = 1|x = (1,0,0.5)) = \frac{1}{1 + e^{-(0.5+1.21\cdot1-0.7\cdot0+0.3\cdot0.5)}} = \frac{1}{1 + e^{-1.86}} = 0.865$$

$$p_2 = P(Y = 1|x = (3,5,1)) = \frac{1}{1 + e^{-(0.5+1.21\cdot3-0.7\cdot5+0.3\cdot1)}} = \frac{1}{1 + e^{-0.93}} = 0.717$$

$$p_3 = P(Y = 1|x = (0.1,1,7)) = \frac{1}{1 + e^{-(0.5+1.21\cdot0.1-0.7\cdot1+0.3\cdot7)}} = \frac{1}{1 + e^{-2.02}} = 0.883$$

ומתקיים $p_3 > p_1 > p_2$

**שאלה 6**

בפני סטודנט נתונה קבוצה של נקודות דו-ממדיות השייכות לשתי מחלקות שונות. הנקודות ברות הפרדה לינארית. סטודנט מתלבט האם להשתמש ב SVM רגיל, או בגרסת ה Soft Margin.

מה העצה הטובה ביותר שתוכל/י לספק לו?

א. מסווג מסוג $SVM$ אינו מתאים לשימוש במקרה של תצפיות דו מימדיות ולכן אף אחד מהכלים לא יעבוד עבור הסטודנט ועליו לחפש פתרון חלופי.

ב. כדאי לנסות להפעיל Soft Margin SVM ולראות אם מקבלים "פרוזדור" גדול יותר בין הנקודות שסווגו נכונה. הדבר יכול למנוע Over-fitting.

ג. אם הנקודות ניתנות להפרדה ע"י קו לינארי, אזי התוצאה של SVM ושל Soft Margin SVM יהיו בהכרח זהות, ולכן לא משנה באיזה שיטה יבחר הסטודנט.

ד. כאשר הנקודות ניתנות להפרדה לינארית, Soft Margin SVM יוביל בהכרח ל under fitting, ולכן כדאי להימנע משימוש בו.

ה. לא ניתן להפעיל Soft Margin SVM כאשר הנקודות ניתנות להפרדה לינארית. בשיטה זו משתמשים רק כאשר אין אפשרות להפרדה כזו.

**פתרון: ב'**

**שאלה 7**

2. נתונות שלוש נקודות במרחב דו-ממדי (1,1), (2,2) ו-(3,3). מחשבים את ה-Principle Component הראשי ומטילים את הנקודות עליו. מה ה-variance של הנקודות המוטלות?

א. 4/3

ב. 2

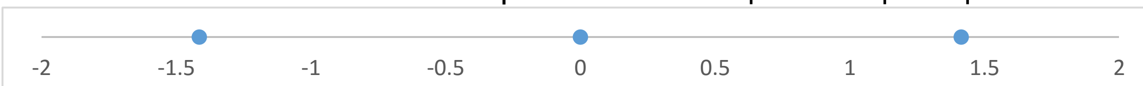
ג. 1.5

ד. $\sqrt{3}$

ה. 0

**פתרון: א', ב'**

לאחר הטלת הנקודות נקבל שלוש נקודות על ציר אחד להלן:

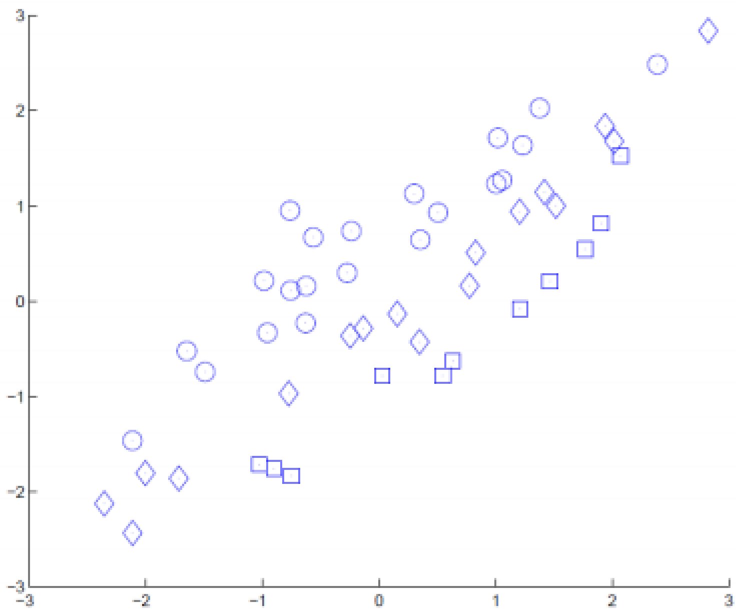



חישוב השונות: $Var = \frac{\left((\sqrt{2}-0)^2+0+(-\sqrt{2}-0)^2\right)}{3} = \frac{4}{3}$

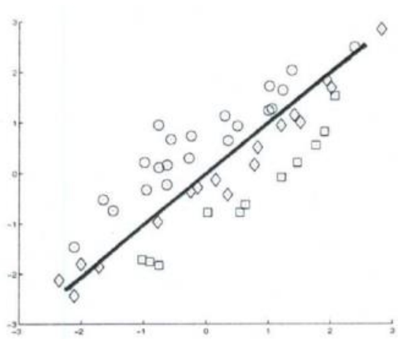
חישוב השונות (ב): $Var = \frac{\left((\sqrt{2}-0)^2+0+(-\sqrt{2}-0)^2\right)}{2} = 2$

**שאלה 8**

1. נתון מדגם אימון עבור בעיית סיווג עם 3 מחלקות. להלן גרף של תצפיות מדגם האימון במישור (כל תצפית מסומנת ע"י צורה אחרת שמסמלת את המחלקה שלה).




הוסיפו לגרף את הקו שכיוונו לפי ה- principal component הראשון.



1(a) First PCA component

**שאלה 9**

נתונות התצפיות הבאות: $(-1,0), (0,1), (1,-1)$.

א. חשבו את ה- principal component הראשון. מה אחוז השונות המוסברת על ידו?

ב. מה הקואורדינטות של שלוש התצפיות לאחר הטרנספורמציה שלהם למרחב חד-ממדי באמצעות הוקטור שחישבתם בסעיף הקודם? מה השונות שלהן במרחב החד-ממדי?

ג. ביחס לתצפיות החדשות שחישבת בסעיף הקודם, עכשיו מנסים לשחזר מהן את התצפיות המקוריות בדו-ממד (reconstruction). מה טעות השחזור (reconstruction error)?

**פתרון**

א. נרשום את מטריצת הנתונים:

$$X = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}$$

ממוצע התצפיות הינו $\bar{x} = \frac{1}{3}(x_1 + x_2 + x_3) = 0$ ולכן אין צורך להחסיר את הממוצע מהנתונים.

מטריצת הקוואריאנס היא:

$$S = X^T X = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

חישוב ערכים עצמיים:

$$|S - \lambda I| = \begin{vmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1 = 0 \;\Rightarrow\; \lambda_1 = 3, \lambda_2 = 1$$

ה-principal component הראשון מתאים לערך העצמי הגבוה יותר $\lambda_1 = 3$:

$$\begin{pmatrix} 2-3 & -1 \\ -1 & 2-3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\Rightarrow\; v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

ולאחר נרמול נקבל:

$$v = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

אחוז השונות המוסברת על ידו היא:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{3}{3+1} = 0.75$$

ב. נמצא את התצפיות במרחב החדש :

$$\widetilde{x_1} = v^T x_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}^T\begin{pmatrix}1\\-1\end{pmatrix} = \sqrt{2}$$

$$\widetilde{x_2} = v^T x_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}^T\begin{pmatrix}0\\1\end{pmatrix} = -\frac{1}{\sqrt{2}}$$

$$\widetilde{x_3} = v^T x_3 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}^T\begin{pmatrix}-1\\0\end{pmatrix} = -\frac{1}{\sqrt{2}}$$

והשונות שלהן היא :

$$\frac{1}{3}\left[\left(\sqrt{2}\right)^2 + \left(-\frac{1}{\sqrt{2}}\right)^2 + \left(-\frac{1}{\sqrt{2}}\right)^2\right] = 1$$

ג. התצפיות המשוחזרות הן :

$$\widehat{x_1} = \widetilde{x_1}v = \sqrt{2}\cdot\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}$$

$$\widehat{x_2} = \widetilde{x_2}v = -\frac{1}{\sqrt{2}}\cdot\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}-0.5\\0.5\end{pmatrix}$$

$$\widehat{x_3} = \widetilde{x_3}v = -\frac{1}{\sqrt{2}}\cdot\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}-0.5\\0.5\end{pmatrix}$$

טעות השחזור היא :

$$error = \frac{1}{3}\sum_{i=1}^{3}\|x_i - \widehat{x_i}\|^2 = \frac{1}{3}[0 + 0.5 + 0.5] = \frac{1}{3}$$

**<u>שאלה 10</u>**

בהתאמת מודל רגרסיה לוגיסטית על סט נתונים בעל 4 משתנים מסבירים - $v_1, v_2, v_3, v_4$ התקבלו המקדמים הבאים: $w_0 = -5.592, w_1 = 1.0189, w_2 = -1.7721, w_3 = -0.2876, w_4 = 0.6547$

א. רשום את הנוסחה המפורשת לחישוב ההסתברות שהמשתנה Y (התלוי) יקבל ערך 1.

**<u>פתרון:</u>**

$$P(Y = 1|v) = \frac{1}{1 + e^{5.592 - 1.0189v_1 + 1.7721v_2 + 0.2876v_3 - 0.6547v_4}}$$

ב. נתונות 6 התצפיות הבאות :

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $v_1$ | 6 | 2 | 4 | 2 | 8 | 3 |
| $v_2$ | 2 | 3 | 2 | 1 | 1.5 | 1 |
| $v_3$ | 1 | 9 | 0 | 1 | 9 | 0 |
| $v_4$ | 7 | 8 | 7 | 10 | 5 | 10 |
| $Y$ | 0 | 1 | 0 | 1 | 0 | 1 |

a. הצג את מטריצת הבלבול עבור ערך $\tau = 0.5$. חשב את ערך מדדי sensitivity ו- specificity.

**<u>פתרון:</u>**

נמצא את הסיווג לכל תצפית:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $Y$ | 0 | 1 | 0 | 1 | 0 | 1 |
| $P(Y=1\|i)$ | 0.781 | 0.002 | 0.383 | 0.718 | 0.642 | 0.904 |

עבור $\tau = 0.5$ נקבל את הסיווגים הבאים:

| | True class positive | True class negative |
|---|---|---|
| Predicted positive | 2 | 2 |
| Predicted negative | 1 | 1 |

מכאן נקבל:

$$sensitivity = \frac{2}{2+1} = \frac{2}{3} \ , \ specificity = \frac{1}{2+1} = \frac{1}{3}$$

b. פי כמה יגדל **יחס הסיכויים** אם נגדיל את ערך $v_1$ של תצפית מס' 2 בשתי יחידות? הסבר.

**<u>פתרון:</u>**

יחס הסיכויים לפני ההגדלה:

$$\frac{P(Y=1|v)}{P(Y=0|v)} = \frac{\left(1+e^{-w^Tv}\right)^{-1}}{1-\left(1+e^{-w^Tv}\right)^{-1}} = \frac{\left(1+e^{-w^Tv}\right)^{-1}}{e^{-w^Tv}\left(1+e^{-w^Tv}\right)^{-1}} = e^{w^Tv}$$

לאחר שנגדיל את $v_1$ ב-2 (נסמן את התצפית עם השינוי ב-$v^*$) נשים לב כי:

$$e^{w^T v^*} = e^{w^T v} \cdot e^{2w_1}$$

מכאן נקבל כי יחס הסיכויים יגדל פי $e^{2w_1}$.

## שאלה 11

1. בשאלה זאת נראה כי פיצול של הנתונים לפי מאפיינים אינה תמיד הגישה הכי טובה לסיווג. במהלך השאלה, הניחו כי ניתן לפצל כל מאפיין מספר פעמים בבניית עץ ההחלטה. נתונים $n$ נקודות בריבוע היחידה $(x_i, y_i) \in [0,1] \times [0,1]$, כל אחת מסומנת ב-'+' או '-'.

א. הראו כי קיים עץ החלטה בעומק לכל היותר $\log_2 n$ שמסווג נכון את כל $n$ הנקודות. בכל צומת העץ יבצע פיצול בינארי (לפי רכיב ה-$x$ של הנקודה או רכיב ה-$y$).

ב. תארו (בצורה מתמטית או מילולית) אוסף נתונים שמכיל $n$ נקודות בריבוע היחידה, כולל הסיווג שלהם ל-'+' או '-', כך שלכל עץ החלטה שמסווג את הנקודות ללא טעות נדרשים לפחות $n-1$ פיצולים.

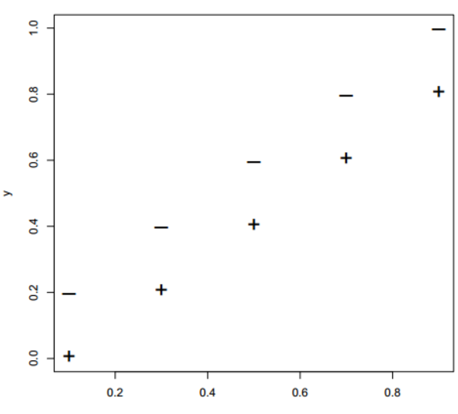
ג. ענו שוב על סעיף ב', אך הפעם על אוסף הנתונים לקיים את האילוץ שניתן להפריד את הנקודות שמסווגות כ-'+' מן הנקודות שמסווגות כ-'-' ע"י קו ישר (מסווג לינארי).

## פתרון

א. הטענה נכונה פשוט מכיוון שבכל פיצול אנו יכולים להוריד את כמות הנקודות פי 2 עד אשר יהיו $n$ עלים, עלה לכל נקודה, שיבטיחו סיווג נכון של כל $n$ הנקודות.

ב. נסתכל על אוסף הנקודות הבא: $x_i = \frac{i}{n},\ y_i = 0,\ i = 1 \dots n$. כל הנקודות עם $i$ אי-זוגי שייכות למחלקה '+' וכל הנקודות עם $i$ זוגי שייכות למחלקה '-'. עבור סט נקודות זה, כדי להשיג 0 טעות בסיווג, ברור כי צריך להפריד (לפצל) בין כל 2 נקודות סמוכות, ולכן יידרשו $n-1$ פיצולים.

ג. נסתכל על אוסף הנקודות הבא: $x_i = \frac{2\lceil \frac{i}{2} \rceil - 1}{n}, y_i = \frac{2\lfloor \frac{i}{2} \rfloor}{n},\ i = 1 \dots n$. כל הנקודות עם $i$ אי-זוגי שייכות למחלקה '+' וכל הנקודות עם $i$ זוגי שייכות למחלקה '-'. דוגמה עם $n = 10$ מוצגת בגרף. הנקודות הללו בבירור ניתנות להפרדה לינארית באמצעות שיטות שראינו כמו SVM ורגרסיה לוגיסטית, אבל עץ החלטה שמפצל בכל פעם לפי קואורדינטה אחת אינו יעיל כאן וידרוש $n-1$ פיצולים.



**<u>שאלה 12</u>**

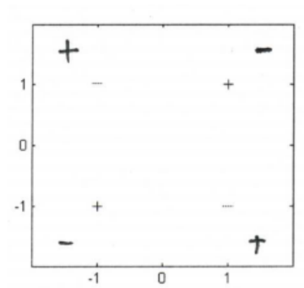
א. נתונה בעיית סיווג בינארית דו-ממדית. התצפיות (1,1) ו-(-1,-1) הן '+', והתצפיות (1,-1) ו-(-1,1) הן '-'. מאמנים SVM עם פונקציית קרנל $\phi(x) = [x_1, x_2, x_1x_2]$ כך ש- $f(x) = sign\ (w^T\phi(x) + b)$.

i. מצאו את הערכים של $w$ ו-$b$ האופטימליים שיתקבלו.

**<u>פתרון:</u>** $w = (0,0,1),\ \ b = 0$.

ii. הוסיפו תצפית אימון נוספת למדגם כך שהוא לא יהיה ניתן להפרדה לינארית במרחב $\phi(x)$.

**<u>פתרון:</u>** להלן 4 אפשרויות לנקודות (ויש עוד...)



ב.

i. איך משפיע מספר העלים בעץ החלטה על ה- bias-variance tradeoff?

**<u>פתרון:</u>** ככל שמספר העלים קטן יותר, כך מודל העץ הוא פשוט יותר ונקבל הטיה גדולה יותר אך שונות נמוכה.

ii. מה הבעיה בהערכת מסווג לפי ביצועיו על מדגם האימון? הסבירו איך שימוש בשיטת Cross Validation פותר בעיה זאת.

**<u>פתרון:</u>** יש סכנה ל-Overfitting, ביצועים על מדגם אימון לא מייצגים את הביצועים של המסווג על מדגם מבחן. Cross Validation עוזר כי אנו אומנם נעזרים במדגם האימון בלבד, אך בכל איטרציה אנו בודקים את ביצועי המסווג על קבוצת תצפיות שלא הייתה בסט האימון של אותה איטרציה. כך אנו מקבלים אמד לטעות שעקבי עם טעות מדגם המבחן.

**<u>שאלה 13</u>**

בשאלה זו, עליכם לפתח אלגוריתם EM כדי לבצע Model-Based Clustering על מסמכי טקסט.

כדי לייצג מסמך נשתמש במודל מקובל הקרוי Bag of Words (הרצאה 1 שקפים 19-22). במודל זה, מסמך $d$ (Document) הוא אוסף של מילים, כך שאין משמעות לסדר המילים במסמך. כל מילה לקוחה מתוך מילון $V$ (Vocabulary) שהינו קבוצה סופית של מילים.

למשל, אם נתונים לנו 2 המסמכים הבאים:

$d_1$: John likes to watch movies. Mary likes too.

$d_2$: John also likes to watch football games.

המילון אשר ייבנה הוא:

$$V = \{\text{1-John, 2-likes, 3-to, 4-watch, 5-movies, 6-also, 7-football, 8-games, 9-Mary, 10-too}\}$$

בייצוג BoW, כל מסמך $d$ הוא וקטור באורך $|V|$, כך שהאיבר ה-$j$ של $d$ הינו מספר הפעמים שהמילה ה-$j$ מופיעה במסמך $d$. למשל, עבור 2 המסמכים מלמעלה, ייצוגם במודל BoW הוא:

$d_1 = (1,2,1,1,1,0,0,0,1,1)$

$d_2 = (1,1,1,1,0,1,1,1,0,0)$

נניח כי התפלגות מסמך הינה מודל צירופי (Mixture model), כלומר:

$$p(d) = \sum_{k=1}^{K} \pi_k p(d|\mu_k) \quad , \quad \pi_k \geq 0 \ , \quad \sum\nolimits_{k=1}^{K} \pi_k = 1$$

בנוסף, נניח כי ההסתברות שמילה $j$ תופיע במסמך הינה $\mu_{kj}$, באופן ב"ת במילים אחרות במסמך. מכאן שהתפלגות מסמך יחיד תחת הרכיב ה-$k$ הינה מולטינומית, כלומר:

$$p(d|\mu_k) = \prod_{j=1}^{|V|} \mu_{kj}^{d_j}$$

כאשר $d_j$ זה מספר המופעים של המילה ה-$j$ במסמך $d$. כמו-כן $\mu_{kj} \geq 0$ ו- $\sum_{j=1}^{|V|} \mu_{kj} = 1$.

נתונים $N$ מסמכים $D = \{d_1, d_2, \ldots, d_N\}$, בלתי תלויים ושווי התפלגות, תחת מודל הצירוף המולטינומי שהגדרנו לעיל (Mixture of Multinomials).

א. תארו במדויק אלגוריתם EM המחשב אמדי MLE לפרמטרים $\{\pi_k, \mu_k\}_{k=1}^{K}$. הציגו בפירוש את הפיתוח של ה- E step וה- M step באלגוריתם.

ב. נתון אוסף המסמכים הבא המוגדר מעל מילון המכיל 8 מילים :

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $d_1$ | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 1 |
| $d_2$ | 2 | 1 | 2 | 2 | 2 | 0 | 1 | 0 |
| $d_3$ | 1 | 1 | 0 | 3 | 1 | 1 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $d_4$ | 0 | 1 | 1 | 2 | 7 | 3 | 0 | 1 |
| $d_5$ | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |

מריצים EM מעל אוסף המסמכים הנתון (לפי האלגוריתם שפיתחתם בסעיף הקודם) כאשר $K = 2$, ומקבלים לבסוף את אמדי ה-MLE הבאים :

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\mu_1$ | 0.268 | 0.085 | 0.155 | 0.155 | 0.268 | 0.033 | 0.011 | 0.025 |
| $\mu_2$ | 0.12 | 0.044 | 0.055 | 0.23 | 0.3 | 0.11 | 0.025 | 0.116 |

$$\pi_1 = 0.25, \qquad \pi_2 = 0.75$$

חלקו את אוסף המסמכים לפי Model-Based Clustering כאשר $K = 2$.

<u>תזכורת :</u> מסמך $d$ ישויך לאשכול (Cluster) $C_k$ אם $k = \arg\max_{j=1...K} P\left(d \in C_j \middle| d, \pi, \mu\right)$.

<u>הערה :</u> במקום לחשב מכפלה ארוכה של הסתברויות, עדיף לחשב את סכום ה-log של ההסתברויות כדי לא להגיע לערכים יותר מדי קטנים. השינוי הזה לא ישפיע על התוצאה הסופית עקב המונוטוניות של פונקציית ה-log.

**<u>פתרון</u>**

א. נגדיר $r_{ik}$ משתנה מקרי בינארי ששווה ל-1 אם המסמך ה-$i$ לקוח מהרכיב ה-$k$ , ו-0 אחרת.

כעת נוכל למצוא את פונקציית הנראות המשותפת של המשתנים $d$ ו- $r$:

$$p(d,r|\mu,\pi)=\prod_{k=1}^{K}\left[\pi_k\prod_{j=1}^{|V|}\mu_{kj}^{d_j}\right]^{r_k}$$

פונקציית ה-Dataset log-likelihood :

$$\ln p(D,R|\mu,\pi)=\sum_{i=1}^{N}\sum_{k=1}^{K}r_{ik}\left\{\ln\pi_k+\sum_{j=1}^{|V|}d_{ij}\ln\mu_{kj}\right\}$$

נגדיר $\mathbb{E}(r_{ik})=\gamma(r_{ik})=P(r_{ik}=1|d_i,\mu,\pi)$, ונמצא את תוחלת הנראות:

$$Q(\mu,\pi)=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma(r_{ik})\left\{\ln\pi_k+\sum_{j=1}^{|V|}d_{ij}\ln\mu_{kj}\right\}$$

שלב ה-E-step:

$$\gamma(r_{ik})=P(r_{ik}=1|d_i,\mu,\pi)=\frac{\pi_k p(d_i|\mu_k)}{\sum_{s=1}^{K}\pi_s p(d_i|\mu_s)}=\frac{\pi_k\prod_{j=1}^{|V|}\mu_{kj}^{d_{ij}}}{\sum_{s=1}^{K}\pi_s\prod_{j=1}^{|V|}\mu_{sj}^{d_{ij}}}$$

שלב ה- M-step:

נמצא מקסימום ל-$Q(\mu,\pi)$ ביחס ל-$\mu_k$. נשים לב כי צריך לדרוש ש- $\sum_{j=1}^{|V|}\mu_{kj}=1$, ולכן ניעזר בכופלי לגרנז':

$$\frac{\partial}{\partial\mu_{kj}}\left\{Q(\mu,\pi)+\lambda\left(\sum_{j=1}^{|V|}\mu_{kj}-1\right)\right\}=\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}{\mu_{kj}}+\lambda=0$$

נכפול ב-$\mu_{kj}$ ונסכום על כל $j$:

$$\sum_{j=1}^{|V|}\left(\left(\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}\right)+\lambda\mu_{kj}\right)=\left(\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}\right)+\lambda=0$$

$$\Rightarrow\ \lambda=-\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}$$

נציב במשוואה המקורית ונקבל:

$$\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}{\mu_{kj}}+\lambda=\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}{\mu_{kj}}-\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}=0$$

$$\Rightarrow\ \mu_{kj}=\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}{\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}$$

ובאופן כללי:

$$\mu_{k}=\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{i}}{\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}}$$

עבור המקסימום ביחס ל-$\pi_k$ אנו צריכים להבטיח ש- $\sum_{k=1}^{K}\pi_k=1$. לכן שוב ניעזר בכופלי לגרנז':

$$\frac{\partial}{\partial\pi_k}\left\{Q(\mu,\pi)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)\right\}=\frac{\sum_{i=1}^{N}\gamma(r_{ik})}{\pi_k}+\lambda=0$$

וכפי שראינו בתרגול, נקבל:

$$\pi_k=\frac{\sum_{k=1}^{K}\gamma\left(r_{ij}\right)}{N}$$

לסיכום, אלגוריתם EM לאמידת הפרמטרים הינו:

1. אתחל את ערכי הפרמטרים $\{\pi_k,\ \mu_k\}_{k=1}^{K}$ לערך התחלתי אקראי.
2. (E-step) חשב את ערכי "האחריות":

$$\gamma(r_{ik})=\frac{\pi_k\prod_{j=1}^{|V|}\mu_{kj}^{d_{ij}}}{\sum_{s=1}^{K}\pi_s\prod_{j=1}^{|V|}\mu_{sj}^{d_{ij}}}$$

3. (M-step) אמוד את הפרמטרים באמצעות ערכי "האחריות" החדשים:

$$\mu_{k}=\frac{\sum_{i=1}^{N}\gamma(r_{ik})d_{i}}{\sum_{j=1}^{|V|}\sum_{i=1}^{N}\gamma(r_{ik})d_{ij}},\qquad\pi_k=\frac{\sum_{k=1}^{K}\gamma\left(r_{ij}\right)}{N}$$

4. אם פונקציית ה- log-likelihood או הפרמטרים התכנסו, סיים. אחרת חזור לשלב 2.

**<u>שאלה 14</u>**

נכון/לא נכון: אלגוריתם k-NN עבור בעיית סיווג (k השכנים הקרובים ביותר), הינו מוצלח במיוחד לצורך סיווג תצפיות חדשות כאשר יש הרבה מאוד תצפיות במדגם הלמידה. זה נובע מכך שבשיטה זאת אין צורך להתאים מודל מסובך.

<u>פתרון</u>: לא נכון.
כאשר יש הרבה מאוד נתונים, יש בעיה של אחסון כל התצפיות ושל חישוב הניבוי של תצפית חדשה כי חייבים להתחשב בכל התצפיות שיש במדגם הלמידה.

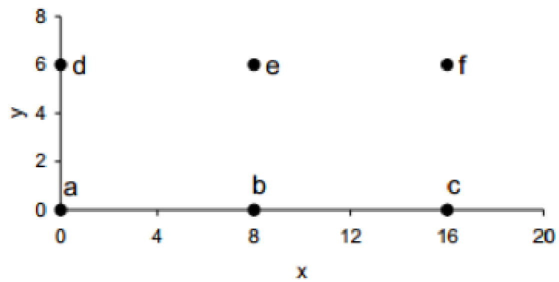
**<u>שאלה 15</u>**

נניח כי עשינו הקבצה על אוסף נתונים עם $N$ תצפיות באמצעות שני אלגוריתמי הקבצה שונים: k-Means ו- Gaussian Mixtures. בשני המקרים קיבלנו 5 אשכולות (clusters), ובשני המקרים המרכזים של האשכולות זהים. האם 3 תצפיות שמסווגות לאשכולות שונים בפתרון שנתן k-Means, יכולות להיות מסווגות לאותו אשכול בפתרון שנתן Gaussian Mixtures? אם לא, הסבירו. אם כן, ציירו דוגמה או נמקו מילולית.

<u>פתרון</u>: כן.
k-Means מסווג כל תצפית לאשכול ייחודי בהסתמך על המרחק שלה ממרכז האשכול. Gaussian Mixtures נותן השמה "רכה" (הסתברותית) לכל תצפית. לכן, גם אם מרכזי האשכולות זהים בשתי השיטות, אם לרכיבים הגאוסיינים יש שונויות גדולות, תצפיות בגבולות בין אשכולות עלולות לקבל סיווג שונה בפתרון של Gaussian Mixtures.

## שאלה 16

נתון אוסף התצפיות הבא במישור:




מריצים על התצפיות את אלגוריתם *k-means* עם $k = 3$. האלגוריתם משתמש בפונקציית מרחק אוקלידית כדי לשייך כל תצפית למרכז הקרוב אליה ביותר. תיקו נפתר לטובת מרכז בכיוון שמאל-למטה.

נגדיר **קונפיגורציה $k$-התחלתית** להיות תת-קבוצה בגודל $k$ של התצפיות שתהווה אתחול ל-$k$ המרכזים באלגוריתם. למשל, $\{a, b, c\}$ הינה קונפיגורציה 3-התחלתית שקובעת כי מרכז ה-cluster הראשון נמצא בנקודה $a$ , מרכז ה-cluster השני בנקודה $b$ ומרכז ה-cluster השלישי בנקודה $c$.

כמה קונפיגורציות 3-התחלתיות קיימות כך שריצת *k-means* עליהן עם $k = 3$ תניב את החלוקה:

$$\{a, b\}, \{d\}, \{c, e, f\}$$

א. 1

ב. 2

ג. 4

ד. 8

ה. לא קיימות קונפיגורציות כאלו

**פתרון**

לא קיימות

## **שאלה 17**

אתם עובדים בתור סוקרים בכנס בינלאומי בנושא כריית נתונים, ונדרשים לעבור על מאמרים ולהחליט האם לדחות או לקבל כל מאמר בהתאם לנכונות הניסויים שהוא מציג והסקת המסקנות שלו. אילו מבין המאמרים הבאים הייתם מוכנים לקבל?

א. "האלגוריתם שלי הכי טוב! הוא השיג טעות חיזוי מאוד נמוכה על מדגם האימון"

ב. "האלגוריתם שלי הכי טוב! הוא השיג טעות חיזוי מאוד נמוכה על מדגם המבחן. (התוצאות מוצגות ביחס לפרמטר $\lambda$ הטוב ביותר שנבחר על פי מדגם המבחן)"

ג. "האלגוריתם שלי הכי טוב! הוא השיג טעות חיזוי מאוד נמוכה על מדגם המבחן. (התוצאות מוצגות ביחס לפרמטר $\lambda$ הטוב ביותר שנבחר על פי 10-fold CV על מדגם האימון)"

ד. סעיפים ב' ו-ג' נכונים

ה. אף תשובה אינה נכונה

**פתרון** : ג

ב' לא מתאים מכיוון שלא משתמשים במדגם המבחן לצורך בחירת פרמטרים לבניית המודל.

## **שאלה 18**

לפניך מספר טענות הקשורות לאלגוריתם Support Vector Machine (SVM). הקף את הטענה הנכונה :

א. SVM, בדומה לרגרסיה לוגיסטית, מחזיר את ההסתברות למחלקה בהינתן תצפית

ב. סביר להניח כי הוקטורים התומכים יישארו אותו דבר כאשר נעבור מקרנל לינארי לקרנל פולינומי עם דרגה גבוהה יותר.

ג. למשטח ההפרדה עם שוליים מרביים (max-margin decision boundary) ש-SVM בונה יש את **טעות המבחן** (generalization error) הקטנה ביותר מבין כל המסווגים הלינאריים.

ד. מריצים SVM פעמיים על אותו מדגם אימון, כל פעם עם פונקציית קרנל שונה. אזי ניתן לדעת איזה מודל יצליח יותר על **מדגם המבחן** לפי ערכי גודל השוליים שנקבל משני המודלים.

ה. אף תשובה אינה נכונה

**פתרון** : ה

## שאלה 19

ידוע כי במבחן ב"כריית נתונים" ההסתברות שסטודנט יקבל ציון בין 80 ל-100 היא $1/2$, בין 60 ל-80 היא $\mu$, בין 40 ל-60 היא $2\mu$, ובין 0 ל-40 היא $1/2-3\mu$. לאחר המבחן התגלה כי $c$ סטודנטים קיבלו בין 40 ל-60 ו- $d$ סטודנטים בין 0 ל-40. לא ידוע כמה בדיוק סטודנטים קיבלו ציון בין 80 ל-100 (נסמן ב-$a$) וכמה קיבלו בין 60 ל-80 (נסמן ב-$b$), אך כן ידוע שמספר הסטודנטים שציונם נע בין 60 ל-100 הוא $h$. כלומר, $a, b$ הם ערכים נסתרים שמקיימים $a+b=h$.
המטרה היא להיעזר באלגוריתם EM כדי למצוא אמד נראות מרבית ל-$\mu$.

E-step – מה הנוסחה שמחשבת את התוחלת של $a, b$ בהינתן $\mu$?

א. $\hat{b}=\frac{\mu}{1/2+h}\mu \quad \hat{a}=\frac{1/2}{1/2+h}\mu$

ב. $\hat{b}=\frac{\mu}{1/2+\mu}h \quad \hat{a}=\frac{1/2}{1/2+\mu}h$

ג. $\hat{b}=\frac{1/2}{1/2+\mu}h \quad \hat{a}=\frac{\mu}{1/2+\mu}h$

ד. $\hat{b}=\frac{\frac{1}{2}}{1+\mu^2}h \quad \hat{a}=\frac{\mu}{1+\mu^2}h$

ה. $\hat{b}=\frac{1/2}{1/2+h^2}\mu \quad \hat{a}=\frac{\mu}{1/2+h^2}\mu$

**פתרון** : ב

$$E(a|\mu,h,c,d)=h\cdot P(grade\in[80{,}100]|grade\in[60{,}100])=h\cdot\frac{1/2}{1/2+\mu}$$

$$E(b|\mu,h,c,d)=h\cdot\frac{\mu}{1/2+\mu}$$

M-step – מה הנוסחה שמחשבת את אמד הנראות המרבית של $\mu$ בהינתן התוחלות של $a, b$?

א. $\hat{\mu}=\frac{h-a+c}{6(h-a+c+d)}$

ב. $\hat{\mu}=\frac{h-a+d}{6(h-2a-d)}$

ג. $\hat{\mu}=\frac{h-a}{3(h-2a+c)}$

ד. $\hat{\mu}=\frac{2(h-a)}{3(h-a+c+d)}$

ה. $\hat{\mu}=\frac{2(h-a+d)}{3(h-a+c+d)}$

**פתרון** : א

פונקציית (לוג) נראות:

1. $\ell(\mu)=\log\left(\frac{1}{2}\right)^a(\mu)^{h-a}(2\mu)^c\left(\frac{1}{2}-3\mu\right)^d=a\log\frac{1}{2}+(h-a)\log\mu+c\log 2\mu+d\log\left(\frac{1}{2}-3\mu\right)$

2. $\frac{\partial\ell(\mu)}{\partial\mu}=\frac{h-a}{\mu}+\frac{c}{\mu}-\frac{3d}{\frac{1}{2}-3\mu}=0\ \Rightarrow\ \hat{\mu}=\frac{h-a+c}{6(h-a+c+d)}$

## שאלה 20

בשאלה זו נשווה בין שתי שיטות הסיווג: k-NN ועץ החלטה (decision tree). לצורך השאלה הניחו מדגם אימון בעל 5000 תצפיות ועץ החלטה אשר נגזם להכיל 20 פיצולים.

מה מהבאים נכון:

א. בעץ החלטה תהליך בניית המודל על מדגם האימון ארוך יותר אך תהליך הסיווג של תצפית חדשה קצר יותר

ב. ב-k-NN תהליך בניית המודל על מדגם האימון ארוך יותר אך תהליך הסיווג של תצפית חדשה קצר יותר

ג. בעץ החלטה תהליך בניית המודל על מדגם האימון ארוך יותר וגם תהליך הסיווג של תצפית חדשה ארוך יותר

ד. ב-k-NN תהליך בניית המודל על מדגם האימון ארוך יותר וגם תהליך הסיווג של תצפית חדשה ארוך יותר

ה. לא ניתן לקבוע על סמך הנתונים.

**פתרון**

א – ב-k-NN אין תהליך בניית מודל ולכן בנייתו קצרה יותר. עם זאת בסיווג תצפית חדשה יש לחשב את המרחקים מכל 5000 תצפיות האימון לצורך מציאת k השכנים לעומת בדיקה של לכל היותר 20 תנאים לסיווג תצפית בעץ החלטה.

## שאלה 21

נתונים שני מטבעות בעלי הסתברויות לא ידועות לקבלת "עץ" – למטבע הראשון הסתברות $p$ ולמטבע השני הסתברות $q$. המטבע הראשון נבחר בהסתברות $\pi$ והשני בהסתברות $1-\pi$.

נתון אוסף נתונים בעל $N$ תצפיות $X=\{x_1,x_2,\dots x_N\}$. כך תצפית היא תוצאת הטלה אחת של אחד המטבעות. כלומר $x_i \in \{0,1\}$, כאשר "עץ" = 1.

לצורך חלוקת התצפיות לשתי מחלקות השתמשו באלגוריתם EM באופן הבא:

הוצג משתנה חבוי $Z$ שהינו אינדיקטור למטבע הנבחר:

$r_i = 1$ – המטבע הראשון נבחר (הסתברות ל"עץ" = $p$)

$r_i = 0$ – המטבע השני נבחר (הסתברות ל"עץ" = $q$)

הנוסחא לחישוב ערך האחריות $\gamma(r_i)$ בשלב E הינה:

א. $$\frac{\pi p^{x_i}(1-p)^{(1-x_i)}}{\pi p^{x_i}(1-p)^{(1-x_i)}+(1-\pi)q^{x_i}(1-q)^{(1-x_i)}}$$

ב. $$\frac{\pi p^{x_i}(1-p)^{x_i}}{\pi p^{x_i}(1-p)^{x_i}+(1-\pi)q^{x_i}(1-q)^{x_i}}$$

ג. $$\frac{q^{x_i}(1-q)^{(1-x_i)}}{p^{x_i}(1-q)^{(1-x_i)}+p^{x_i}(1-p)^{(1-x_i)}}$$

ד. $$\frac{\pi q^{x_i}(1-q)^{x_i}}{\pi q^{x_i}(1-q)^{x_i}+(1-\pi)p^{x_i}(1-p)^{x_i}}$$

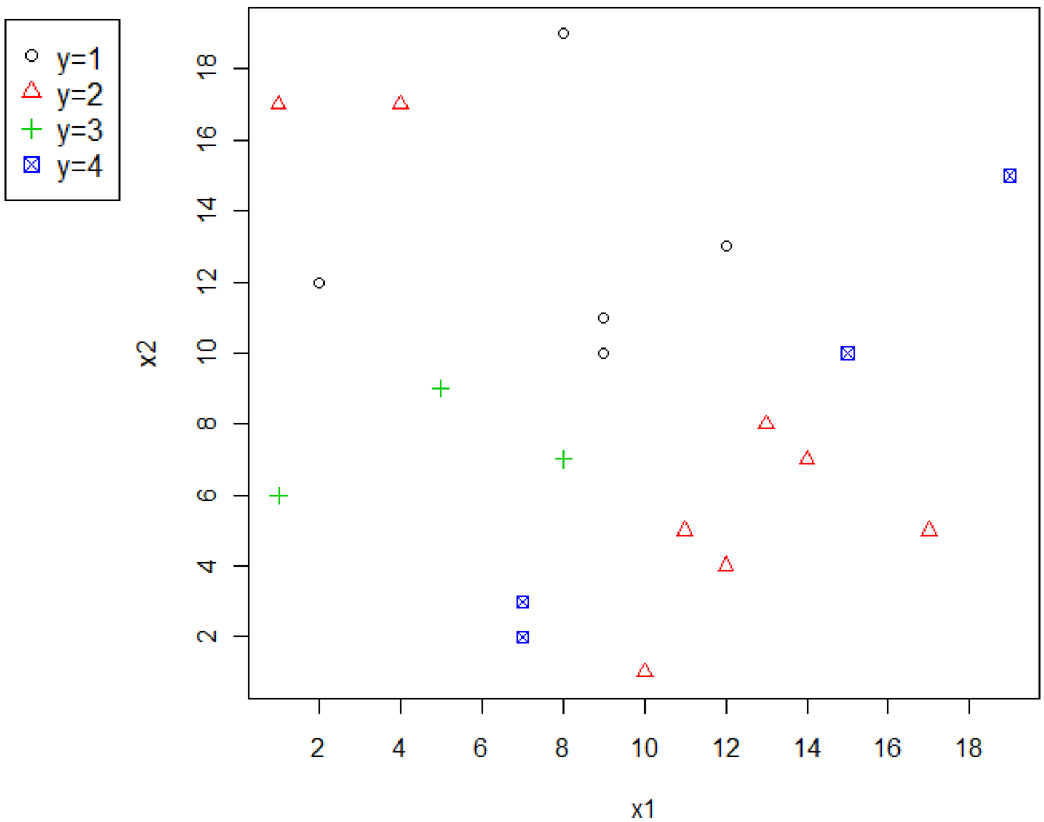
ה. $$\frac{p^{x_i}(1-p)^{x_i}}{p^{x_i}(1-p)^{x_i}+q^{x_i}(1-q)^{x_i}}$$

**פתרון:** א'

$$\gamma(r_i)=P(r_i=1|x_i)=\frac{P(x_i|r_i=1)P(r_i=1)}{P(x_i)}=\frac{\pi p^{x_i}(1-p)^{(1-x_i)}}{\pi p^{x_i}(1-p)^{(1-x_i)}+(1-\pi)q^{x_i}(1-q)^{(1-x_i)}}$$

## שאלה 22

להלן גרף המתאר 20 תצפיות. לכל תצפית שני משתנים מסבירים $X_1$ ו-$X_2$ נומריים ומשתנה מוסבר קטגוריאלי $Y$ שיכול לקבל את הערכים $\{1,2,3,4\}$:




1. בשימוש באלגוריתם CART להתאמת עץ סיווג לנתונים, הוחלט לגדל עץ בעל מספר פיצולים <u>מינימלי</u> אשר מדד הזיהום (Impurity) ע"פ Gini שלו הוא 0. מספר הפיצולים בעץ זה הינו:

א. 5

ב. 6

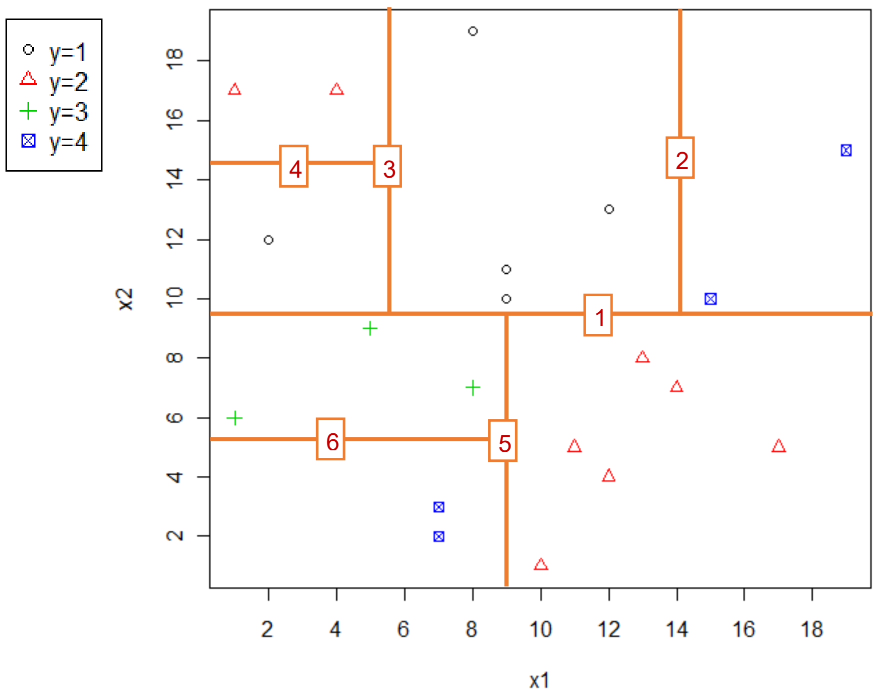
ג. 7

ד. 8

ה. אף אחת מהתשובות לעיל.

**<u>פתרון</u>**

נשרטט על הגרף קוים ישרים מקבילים לצירים שיהוו את הפיצולים בעץ וננסה להגיע למצב שבו בכל "תא שטח" יש רק תצפיות מסוג אחד (ולכן מידת הזיהוי היא 0)




בנינו עץ בעל 6 פיצולים ורמת זיהום 0. ניתן לראות כי לא ניתן להגיע לרמת זיהום 0 ב-5 פיצולים או פחות.

העץ שהתקבל: