סווג ורגרסיה לוגיסטית
(Logistic Regression)

- נעבור עתה לדבר על בעיית הסווג
- בעייה דומה לבעיית הרגרסיה, מלבד זאת ש- y יכול לקבל מספר קטן של ערכים בדידים.
- בעיית סווג בינארית – $y \in \{0,1\}$
  - (אפשר להכליל למקרה הרב-מחלקתי).

דוגמא 1

- **.spam classifier for e-mail**
- **y** – מקבל את הערך 1 אם זהו spam ו – 0 אחרת.
- **0** – negative class “**-**”
- 1 – positive class “**+**”

סווג ורגרסיה לוגיסטית (Logistic Regression)

בהינתן $x^{(i)}$ (וקטור תכונות):

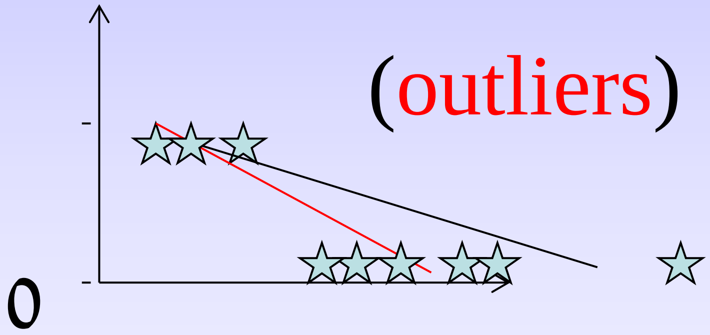
ה- $y^{(i)}$ המתאים נקרא ה**תגית** (label) עבור דוגמת האימון.

לדוגמא (עבור בעיית סווג בינארית:

$$x^{(i)} = \begin{pmatrix} x^{(i)}{}_1 \\ x^{(i)}{}_2 \\ : \\ : \\ x^{(i)}{}_n \end{pmatrix}, \quad y^{(i)} = 0, \ \ or \ \ y^{(i)} = 1$$

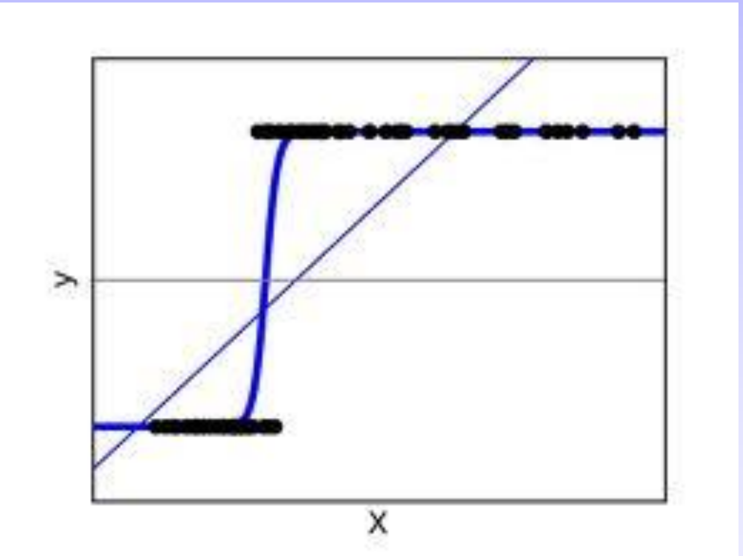
רגרסיה לוגיסטית

- אפשר לגשת לבעיית הסווג ולהתעלם מכך ש- y הינו בדיד, ולהשתמש ברגרסיה לינארית כדי לחזות את הערך של y בהינתן x.
- קל לבנות דוגמאות שיראו שהשיטה הזאת תהיה גרועה למדי.
- א. לדוגמא: בעיית החריגים (outliers)



סווג ורגרסיה לוגיסטית (Logistic Regression)

- נתעלם מכך ש- y הינו בדיד, ונשתמש ברגרסיה לינארית כדי לחזות את הערך של y בהינתן x.



באופן אינטואיטיבי: לא הגיוני ש- $h_\theta(x)$ יקבל ערכים גדולים מ- 1 או קטנים מ- 0 כאשר ידוע ש- $y \in \{0,1\}$

סווג ורגרסיה לוגיסטית (Logistic Regression)

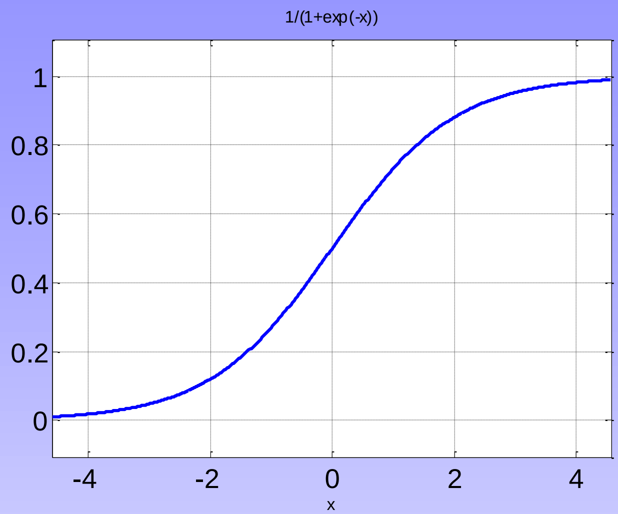
כדי לתקן את זה – נשנה את צורת ההיפותזה $h_\theta(x)$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

הפונקציה: $$g(z) = \frac{1}{1+e^{-z}}$$

נקראת **פונקציה סיגמואידית** או **פונקציה לוגיסטית** **(sigmoid or logistic function)**.

פונקציה לוגיסטית



הערות:

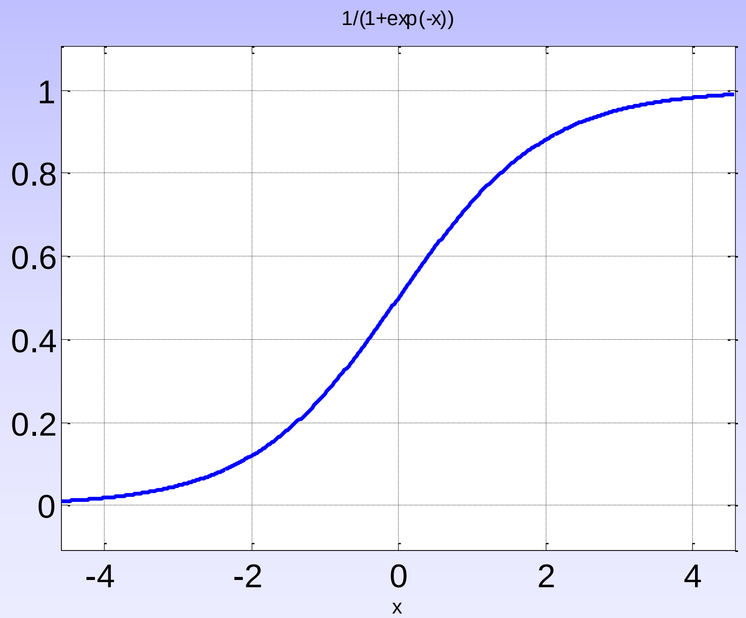
1. $g(z)$ שואף ל- 1 כאשר z שואף ל- ∞ .
2. $g(z)$ שואף ל- 0 כאשר z שואף ל- $-\infty$
3. יותר מכך, $g(z)$ ולכן $h_\theta(x)$ חסומים תמיד בין 0 ל-1.

כמו קודם נשמור על הקונבנציה של $x_0 = 1$ כך ש: $\theta^T x = \sum_{j=1}^{n} \theta_j x_j$

פונקציה לוגיסטית

תכונה שימושית של הנגזרת של הפונקציה הסיגמואידית:

$$g'(z) = \frac{d}{dz}(g(z)) = g(z)\,(\,1 - g(z)\,)$$

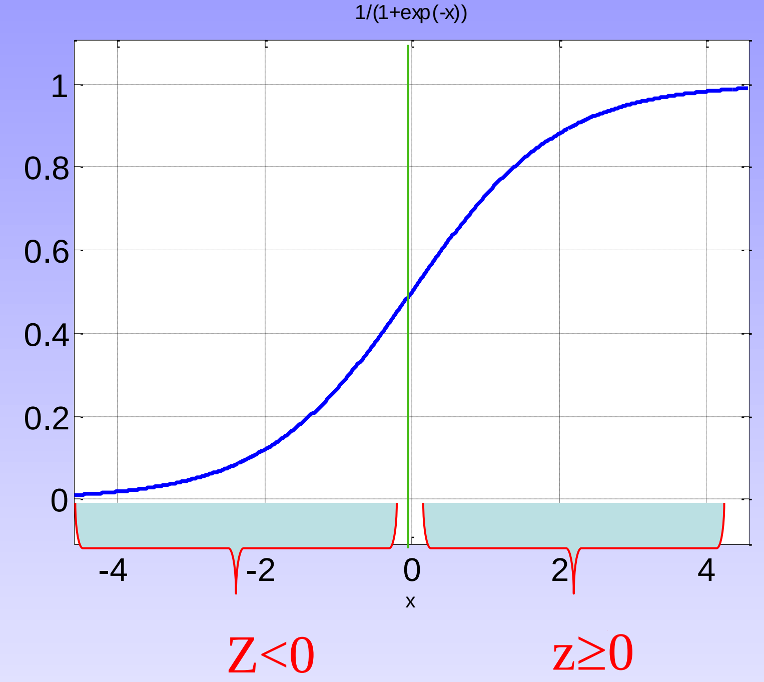


Decision Boundary משטח החלטה

תזכורת: פונקציית ההיפותזה עבור רגרסיה לוגיסטית היא:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

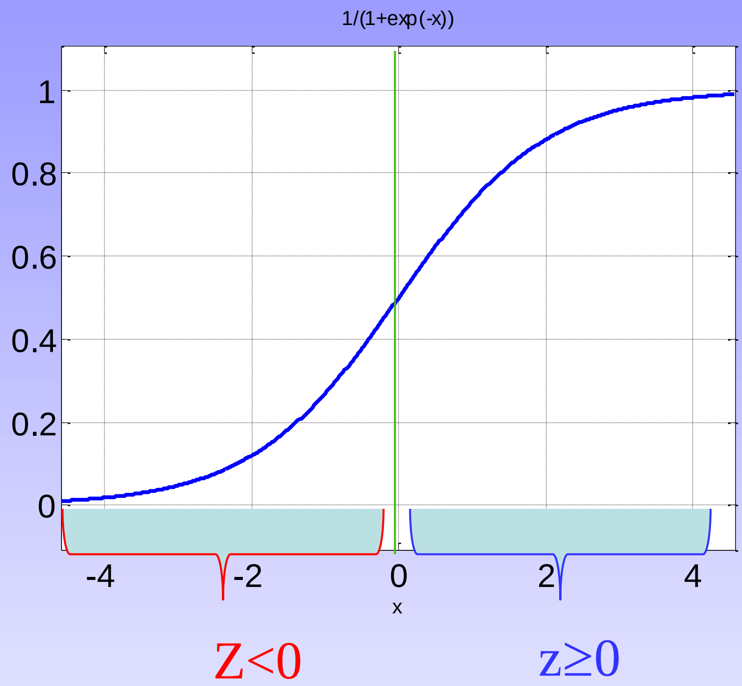
$$g(z) = \frac{1}{1+e^{-z}}$$



נְנסה להבין עבור אלו ערכים ההיפותזה חוזה “y=1” לעומת ערכים עבורם החיזוי הוא “y=0”

Decision Boundary משטח החלטה

נניח כי:



$$h_\theta(x) = g(\theta^T x) = p(y=1 \mid x;\ \theta)$$

$$"y=1" \ \ if \ \ h_\theta(x) \geq 0.5$$

$$"y=0" \ \ if \ \ h_\theta(x) < 0.5$$

$$g(z) = \frac{1}{1+e^{-z}}$$

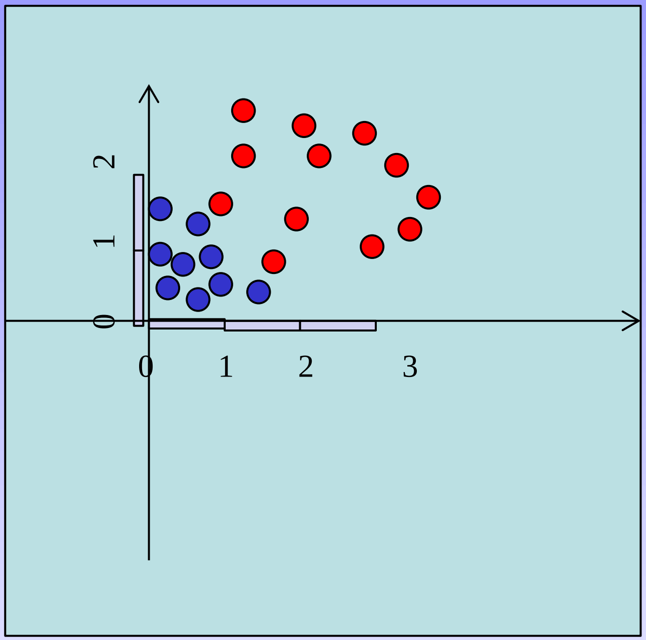
לכן:

$$h_\theta(x) = g(\theta^T x) \geq 0.5 \quad if \quad \theta^T x \geq 0$$

$$h_\theta(x) = g(\theta^T x) < 0.5 \quad if \quad \theta^T x < 0$$

Decision Boundary משטח החלטה

נניח שנתונה קבוצת אימון כמו בדוגמא הבאה:



פונקציית ההיפותזה היא:

$$h_\theta(x) = g(\theta^T x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

נניח כי קיימת שיטה למציאת הפרמטרים,

ונמצא כי:

$$\theta_0 = -2, \quad \theta_1 = 1, \quad \theta_2 = 1$$

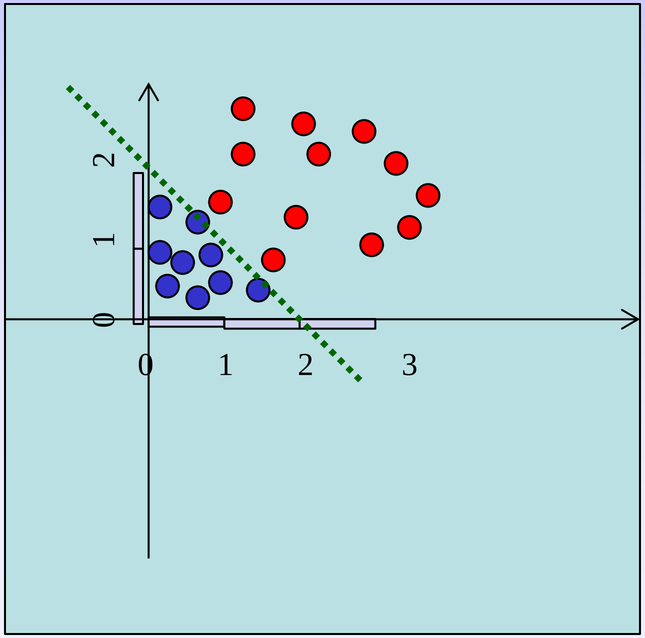
Decision Boundary משטח החלטה

כלומר: $\theta^T = (\theta_0 \quad \theta_1 \quad \theta_2) = (-2 \;\; 1 \;\; 1)$

עבור אלו ערכי x1 ו- x2 ההיפותזה תחזה
y=1, ועבור אלו ערכים y=0?

$$"y=1" \quad if \quad -2+1\cdot x_1 + 1\cdot x_2 \geq 0$$

$$\Rightarrow x_1 + x_2 \geq 2$$



לכל וקטור תכונות $(x_1,x2)$ המקיים תנאי זה ההסתברות לקבל "y=1" גדולה מ- 0.5.

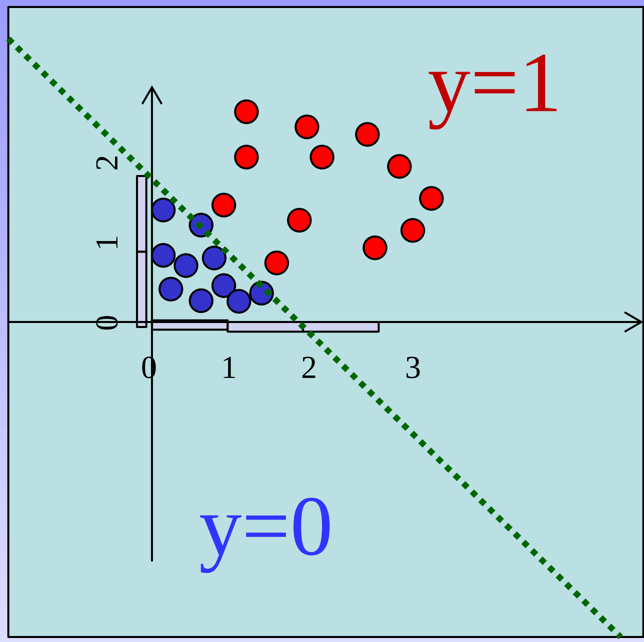
נתבונן במשוואה $x_1 + x_2 = 2$

מהו שיפוע הישר?

מהו ערך החיתוך עם ציר ה- y?

Decision Boundary משטח החלטה

עבור אלו ערכי x1 ו- x2 ההיפותזה תחזה y=1, ועבור אלו ערכים y=0?



$$"y=1" \quad if \quad x_1 + x_2 \geq 2$$

$$"y=0" \quad if \quad x_1 + x_2 < 2$$

הישר המפריד בין שני האזורים נקרא משטח ההחלטה (Decision Boundary).

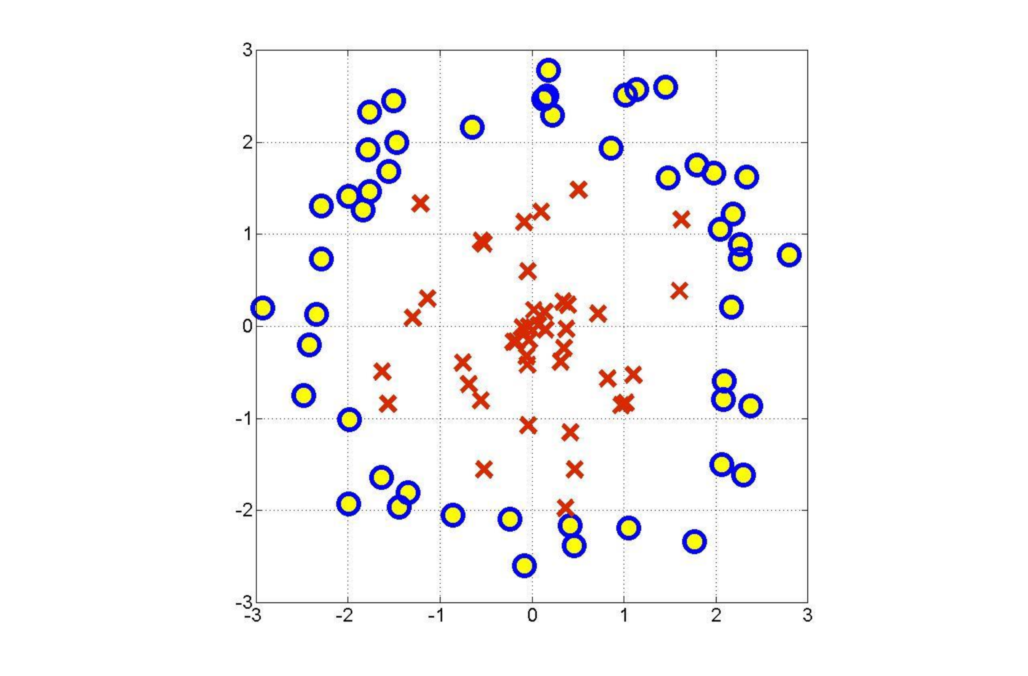
זהו התיאור הגאומטרי של המשוואה $x_1+x_2=2$

כל הנקודות על משטח ההחלטה מקיימות

$$h_\theta(x) = g(\theta^T x) = 0.5$$

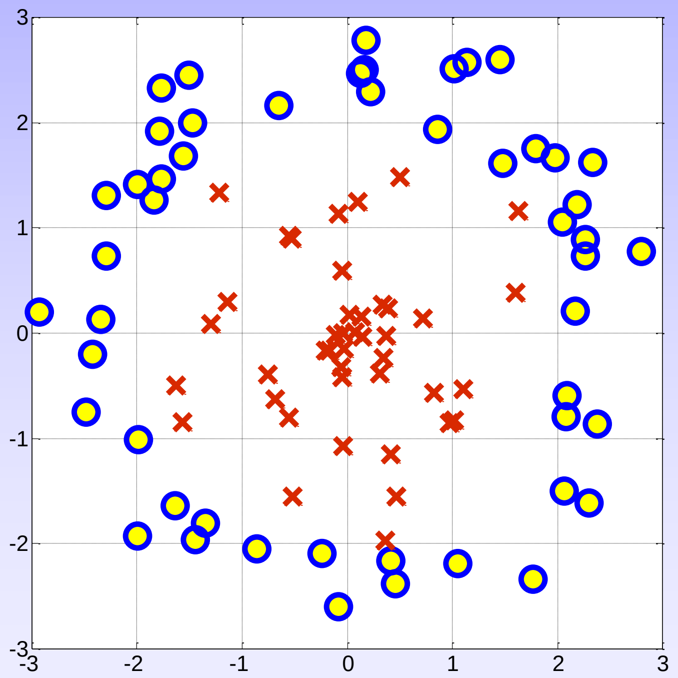
משטח החלטה לא ליניארי Non-Linear

עבור קבוצת האימון בציור, כיצד אפשר ליצור משטח החלטה שיפריד בין הנתונים? כלומר בין הדוגמאות החיוביות "y=1" לבין הדוגמאות השליליות "y=0"?



משטח החלטה לא ליניארי Non-Linear

עבור קבוצת האימון בציור, כיצד אפשר ליצור משטח החלטה שיפריד בין הנתונים? כלומר בין הדוגמאות החיוביות "y=1" לבין הדוגמאות השליליות "y=0"?



משטח החלטה לא ליניארי Non-Linear

פונקציית ההיפותזה במקרה זה היא:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

נניח כי התקבלו הפרמטרים הבאים:

$$\theta^T = (\theta_0 \quad \theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4)^T = (-5 \;\; 0 \;\; 0 \;\; 1 \;\; 1)^T$$

כלומר:

$$h_\theta(x) = g(-5 + 1\cdot x_1^2 + 1\cdot x_2^2)$$

$$"y=1" \quad \Leftrightarrow g(z) \ge 0.5 \Leftrightarrow -5 + 1\cdot x_1^2 + 1\cdot x_2^2 \ge 0$$

$$\Leftrightarrow x_1^2 + x_2^2 \ge 5$$

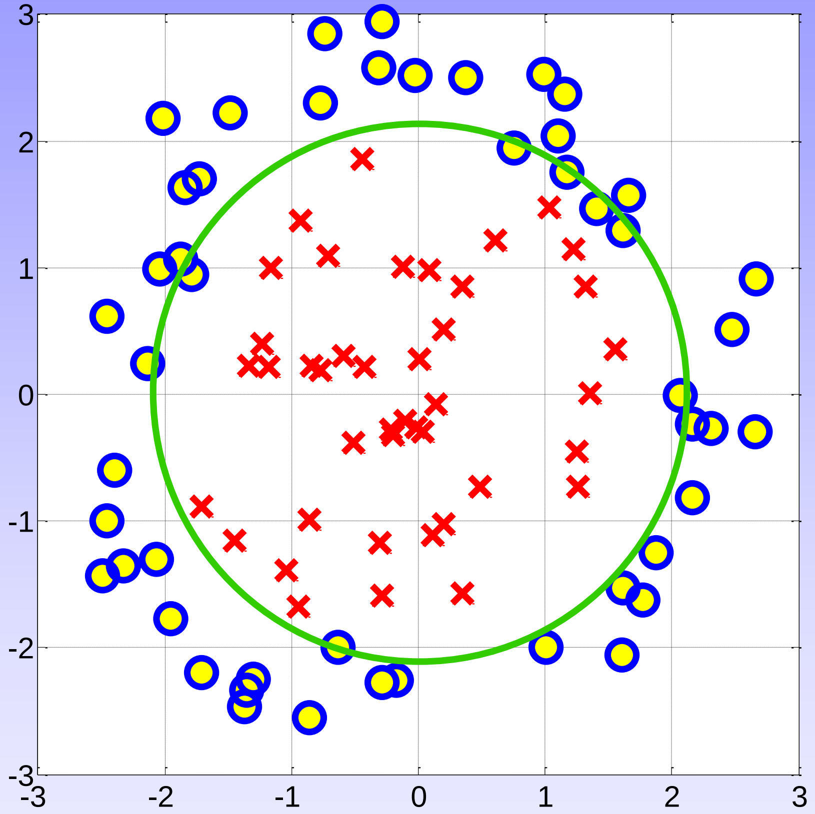
משטח החלטה לא ליניארי Non-Linear

העקום עבור $x_1^{\ 2} + x_2^{\ 2} = 5$

הוא מעגל ברדיוס $\sqrt{5}$ סביב הראשית.

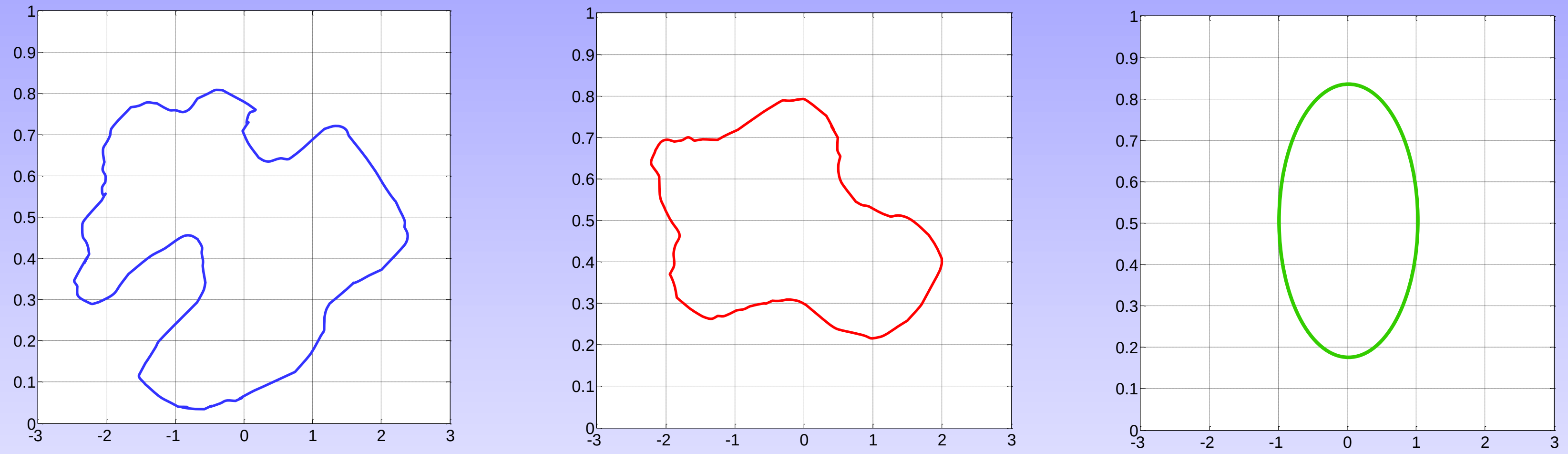
כל הנקודות מחוץ למעגל – "y=1"

כל הנקודות בתוך המעגל – "y=0"



משטח החלטה לא ליניארי Non-Linear

לסיכום : הוספת פולינומים ממעלה גבוהה יותר מאפשרות ליצור משטחי החלטה מורכבים.



התאמת הפרמטרים

בהינתן המודל של הרגרסיה הלוגיסטית, איך נתאים את הפרמטרים עבורו?

הנחות: נתונה קבוצת האימון:

$$\left\{ (x^{(i)}, y^{(i)}); \quad i = 1, 2, \ldots, m \right\}$$

כאשר:

$$x = \left(x_0 \ x_1 \ \ldots \ x_n \right)^T \in R^{n+1} \quad x_0 = 1, \ y \in \left\{ 0, 1 \right\}$$

פונקציית ההיפותזה:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

התאמת הפרמטרים

איך בוחרים את הפרמטרים θ?

ברגרסיה ליניארית השתמשנו בפונקציית המחיר הבאה:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

נשנה מעט את ההגדרה:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{cost}\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

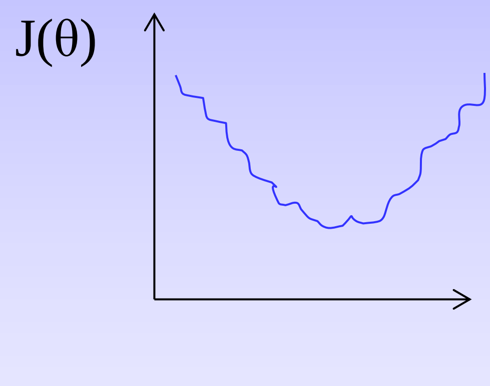
נפשט ונניח שקיימת דוגמא אחת בלבד:

$$\mathrm{cost}(h_\theta(x), y) = \frac{1}{2}\left(h_\theta(x) - y\right)^2$$

(מחצית ריבוע ההפרש)

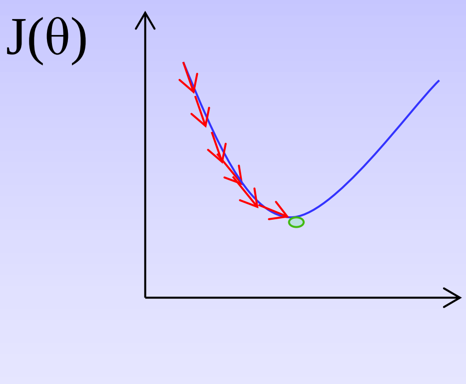
התאמת הפרמטרים

- לו יכולנו למזער את פונקציית המחיר הנ"ל עבור רגרסיה לוגיסטית, אפשר היה לקבל את הפרמטרים המתאימים.
- מתברר שהפונקציה $J(\theta)$ היא לא פונקציה קמורה (convex) ולכן לא מובטח להגיע למינימום גלובלי.



דוגמא לפונקציה לא קמורה

non convex



דוגמא לפונקציה קמורה

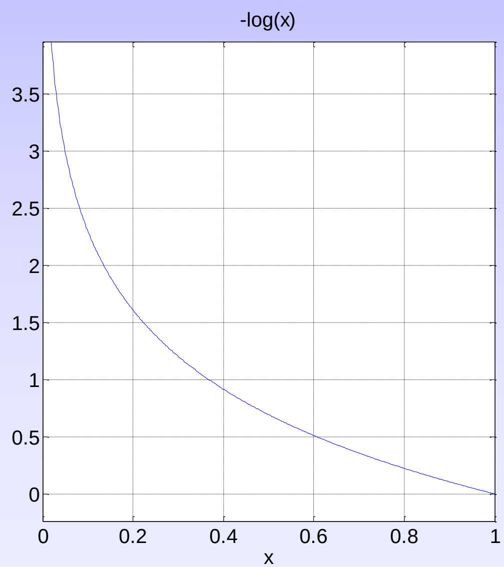
convex

התאמת הפרמטרים

- נגדיר פונקציה אלטרנטיבית לפונקציית המחיר.
- כזאת שמובטח למצוא מינימום גלובלי כי היא קמורה:

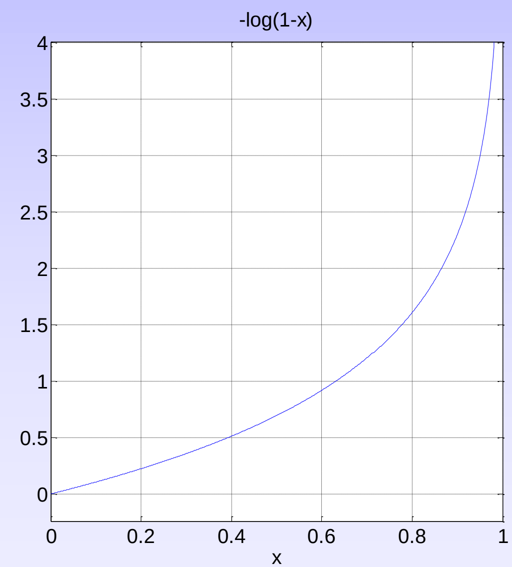
$$\mathrm{cost}(h_\theta(x), y) = \begin{cases} -\log\left(h_\theta(x)\right) & \text{if} \quad y = 1 \\ -\log\left(1 - h_\theta(x)\right) & \text{if} \quad y = 0 \end{cases}$$

$J(\theta)$



y=1 $h_\theta(x)$

$J(\theta)$

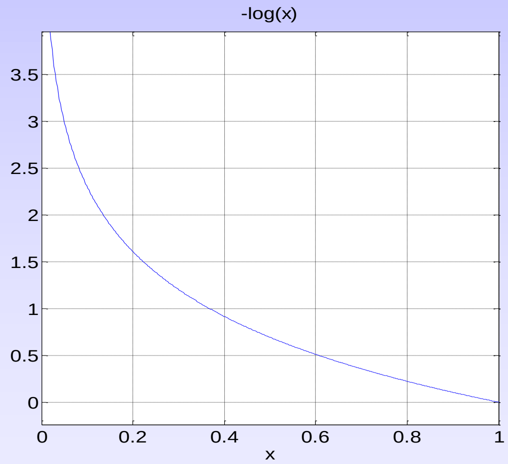


y=0 $h_\theta(x)$

התאמת הפרמטרים

- אם $y=1$ ו- $h_\theta(x)=1$ השגיאה היא 0.
- אם $y=1$ ו- $h_\theta(x)=0$ או קרוב ל- 0 המחיר הוא גבוה מאוד ושואף לאינסוף.
- אם לדוגמא $y=1$ הוא תצפית (spam, חדירה למתקן רגיש וכו') והחיזוי הוא קרוב ל-0 (כלומר ההסתברות לכך ש- $y=1$ היא קרובה ל- 0 כאשר ידוע ש- $y=1$), אזי המחיר צריך להיות גבוה מאוד.
- ה"עונש" של אלגוריתם הלמידה הוא במחיר גבוה מאוד.
-

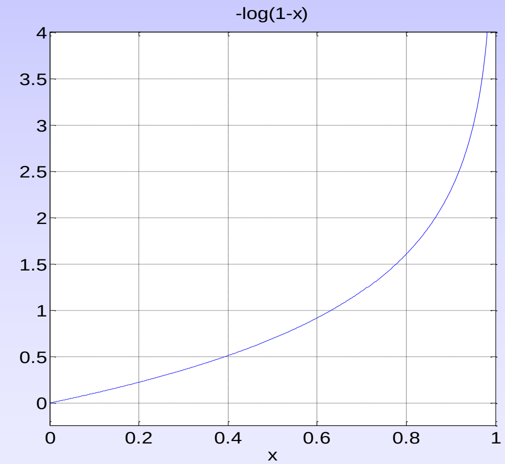
$J(\theta)$



$h_\theta(x)$

$y=1$

$J(\theta)$



$h_\theta(x)$

$y=0$

פונקציית המחיר עבור רגרסיה לוגיסטית

פונקציית המחיר תראה אם-כן כך:

(בהמשך נראה כי פונקציית המחיר הנ"ל נובעת מעיקרון ה- ML)

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{cost}(h_\theta(x^{(i)}),y^{(i)})$$

כאשר:

$$\mathrm{cost}(h_\theta(x),y)=\begin{cases}-\log\left(h_\theta(x)\right) & if \quad y=1\\ -\log\left(1-h_\theta(x)\right) & if \quad y=0\end{cases}$$

מאחר ו- y הוא בינארי ומקבל את הערכים 0 או 1, אפשר לרשום:

$$\mathrm{cost}(h_\theta(x),y)=-y\log\left(h_\theta(x)\right)-\left(1-y\right)\log\left(1-h_\theta(x)\right)$$

פונקציית המחיר עבור רגרסיה לוגיסטית

פונקציית המחיר תראה אם-כן כך:

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{cost}(h_\theta(x^{(i)}),y^{(i)}) \quad =$$

$$=\frac{1}{m}\sum_{i=1}^{m}-y^{(i)}\log\left(h_\theta(x^{(i)})\right)-\left(1-y^{(i)}\right)\log\left(1-h_\theta(x^{(i)})\right)$$

כדי להתאים את הפרמטרים, נרצה למצוא את המינימום של פונקציית המחיר או פונקציית העלות $J(\theta)$:

$$\min_\theta J(\theta)$$

פונקציית המחיר עבור רגרסיה לוגיסטית

נמצא את המינימום של פונקציית המחיר באמצעות אלגוריתם Gradient Descent (אפשר להראות כי לפונקציית המחיר מינימום יחיד, גלובלי):

כזכור, האלגוריתם הוא:

$$\text{Repeat} \quad \{$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\}$$

מעדכנים באופן סימולטני את כל הפרמטרים θ_j.

פונקציית המחיר עבור רגרסיה לוגיסטית

$$\frac{\partial}{\partial\theta_j}J(\theta)=\frac{\partial}{\partial\theta_j}\left(\frac{1}{m}\sum_{i=1}^{m}-y^{(i)}\log\left(h_\theta(x^{(i)})\right)-\left(1-y^{(i)}\right)\log\left(1-h_\theta(x^{(i)})\right)=\right)$$

$$=\frac{1}{m}\sum_{i=1}^{m}\frac{-y^{(i)}}{h_\theta(x^{(i)})}\frac{\partial}{\partial\theta_j}\left(h_\theta(x^{(i)})\right)-\frac{1-y^{(i)}}{1-h_\theta(x^{(i)})}\frac{\partial}{\partial\theta_j}\left(1-h_\theta(x^{(i)})\right)=$$

$$\ldots=\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)})-y^{(i)}\right)x_j^{(i)}$$

פונקציית המחיר עבור רגרסיה לוגיסטית

כלל העדכון אם-כן הוא:

$$\text{Repeat} \ \{$$
$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$
$$\}$$

מעדכנים באופן סימולטני את כל הפרמטרים θ_j.

פונקציית המחיר עבור רגרסיה לוגיסטית

האלגוריתם נראה זהה לזה של רגרסיה ליניארית,

$$\text{Repeat } \{$$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

$$\}$$

ההבדל הוא שפונקציית ההיפותזה במקרה של רגרסיה לוגיסטית היא:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

פונקציית המחיר עבור רגרסיה לוגיסטית

מימוש: כדי לודא שהאלגוריתם מתכנס, נצייר את פונקציית המחיר כתלות במספר האיטרציה.
אפשר לעדכן את כל הפרמטרים על-ידי שימוש בלולאה, או על-ידי וקטוריזציה.

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_1^{(i)}$$

$$\vdots$$

$$\theta_n := \theta_n - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_n^{(i)}$$

התאמת הפרמטרים

בהינתן המודל של הרגרסיה הלוגיסטית, איך נתאים את הפרמטרים עבורו?
עתה נראה כיצד מגיעים לפונקציית המחיר בה השתמשנו.

בעקבות הפיתוח שעשינו עבור מודל הרגרסיה הלינארית, אותו קיבלנו לאחר שהנחנו מספר הנחות הסתברותיות פשוטות,
נניח גם כאן מספר הנחות הסתברותיות ואז נתאים את הפרמטרים באמצעות סבירות מירבית (maximum likelihood).

התאמת הפרמטרים

- נניח כי:

$$p(y=1\,|\,x;\ \theta)=h_\theta(x)$$

$$p(y=0\,|\,x;\ \theta)=1-h_\theta(x)$$

- (הערה: הכוונה כאן היא p(y=1|x; θ) ההסתברות לקבל y=1 בהינתן x כאשר θ – הפרמטרים)
- אפשר לרשום את זה באופן יותר קומפקטי על-ידי:

$$p(y\,|\,x;\ \theta)=(h_\theta(x))^{y}(1-h_\theta(x))^{1-y}$$

התאמת הפרמטרים

- נניח כי m דוגמאות האימון נוצרות באופן בלתי תלוי, כלומר שהן בלתי-תלויות סטטיסטית, אפשר אז לרשום את הסבירות של הפרמטרים על-ידי:

$$L(\theta) = p(y \mid X;\ \theta) =$$

$$= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\ \theta) =$$

$$= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

התאמת הפרמטרים

- כמו ברגרסיה הלינארית, יהיה קל יותר לבצע מקסימיזציה ללוג הסבירות (Log likelihood)

$$l(\theta) = \log(L(\theta)) =$$

$$= \sum_{i=1}^{m} \log(p(y^{(i)} \mid x^{(i)};\ \theta))$$

$$= \sum_{i=1}^{m} \log((h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}})$$

$$\sum_{i=1}^{m} y^{(i)} \log((h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

סבירות מירבית

- כיצד נביא את הסבירות למקסימום?
- בדומה למה שעשינו ברגרסיה לינארית, אפשר להשתמש ב-gradient ascent (או כפי שכבר ראינו על-ידי gradient descent).
- בכתיבה וקטורית, חוק ה- gradient ascent:

$$\theta := \theta + \alpha \cdot \nabla_{\theta} l(\theta)$$

סבירות מירבית

$$\theta := \theta + \alpha \cdot \nabla_\theta l(\theta)$$

נניח שיש לנו רק דוגמת אימון אחת: (x, y)

ונשתמש בנגזרות כדי לקבל את חוק העדכון של ה- gradient ascent:

$$\frac{\partial}{\partial \theta_j} l(\theta) = \left(y \frac{1}{h_\theta(x)} - (1 - h_\theta(x)) \frac{1}{1 - h_\theta(x)}\right) \frac{\partial}{\partial \theta_j} h_\theta(x)$$

סבירות מירבית

$$\theta_j := \theta_j + \alpha \cdot (\, y^{(i)} - h_\theta(x^{(i)})\,)\, x_j^{(i)}$$

אם משווים את הכלל הזה לכלל ה- LMS, הם נראים זהים, אבל זהו לא אותו אלגוריתם, כי עכשיו $h_\theta(x^{(i)})$ מוגדר על-ידי פונקציה לא לינארית של $\theta^T x^{(i)}$

אלגוריתם לימוד הפרספטרון

נניח שמשנים את שיטת הרגרסיה הלוגיסטית, ו"מכריחים" אותה להוציא ערכי פלט של 0 או 1 בלבד:

$$g(z) = \begin{cases} 1 & if \quad z \geq 0 \\ 0 & if \quad z < 0 \end{cases}$$

אם מניחים ל- $h_\theta(x)$ להיות $g(\theta^T x)$:

$$h_\theta(x) = g(\theta^T x)$$

אבל משתמשים בהגדרה החדשה של g ובכלל העדכון:

$$\theta_j := \theta_j + \alpha \cdot (\, y^{(i)} - h_\theta(x^{(i)})\,)x_j^{(i)}$$

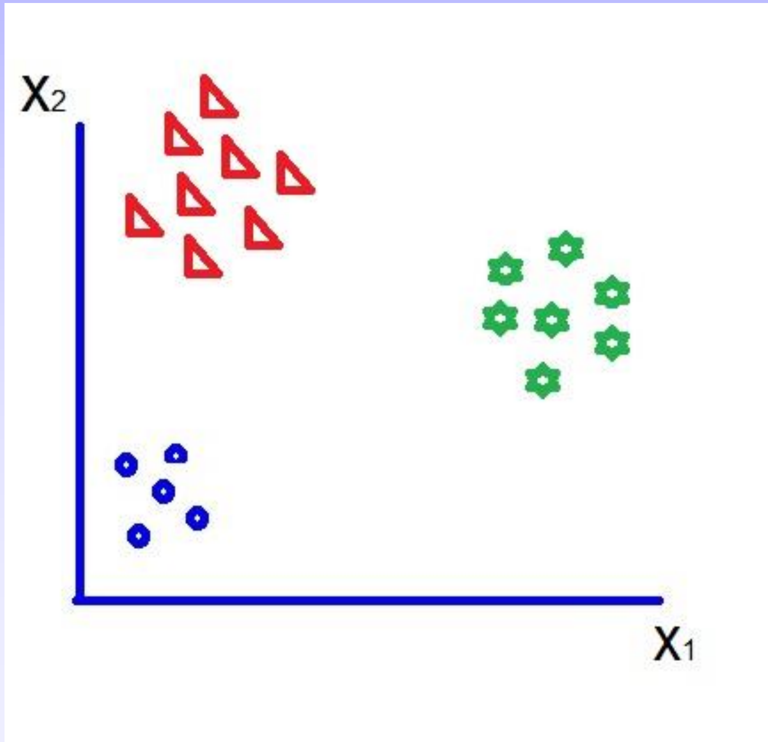
מקבלים את אלגוריתם לימוד הפרספטרון.

סווג רב-מחלקתי

- סווג רב-מחלקתי multiclass classification
- דוגמא:
- מזג האוויר: שמשי (y=1), מעונן חלקית (y=2), מעונן (y=3), גשום (y=4), מושלג (y=5).
- מצב קוגניטיבי: תקין, MCI, פגוע (y=0,1,2)

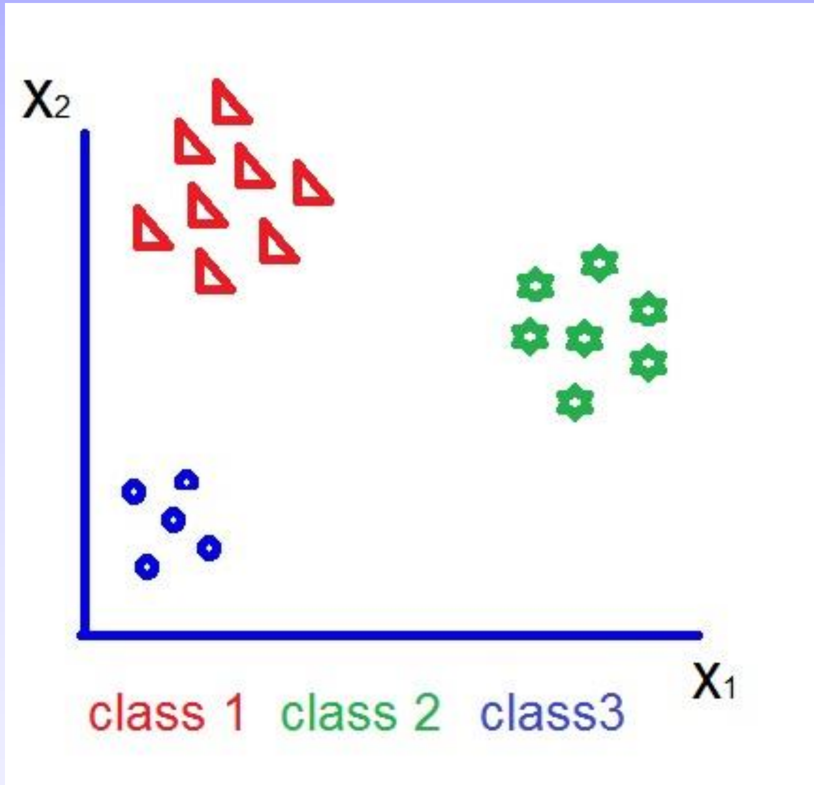
סווג רב-מחלקתי

- עבור סווג בינארי – אפשר לבצע את ההפרדה על-ידי רגרסיה ליניארית.
- כיצד נבצע את ההפרדה עבור המקרה הרב-מחלקתי?
- נדגים סווג ל -3 מחלקות



סווג רב-מחלקתי

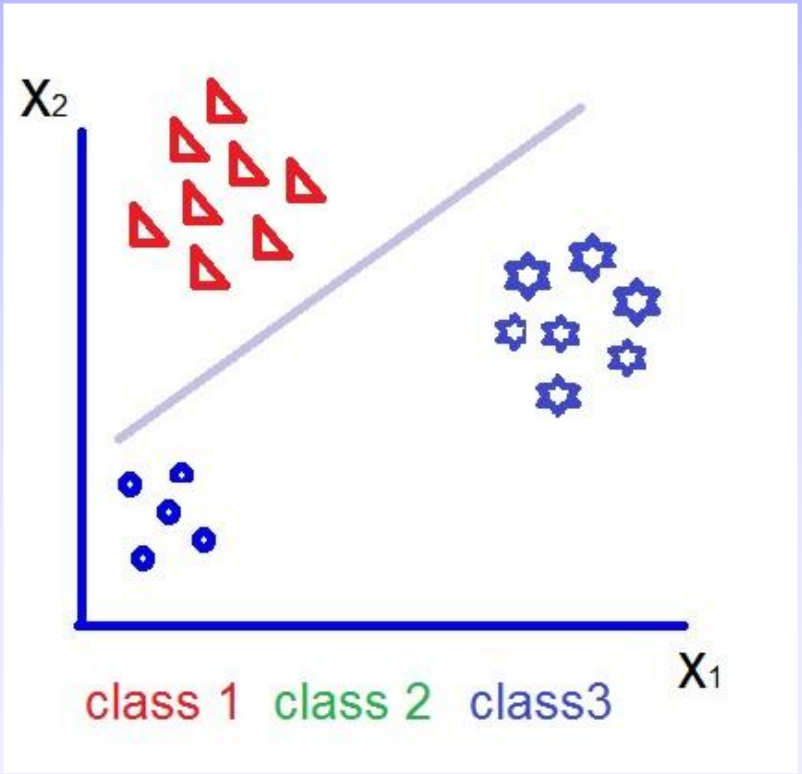
נשתמש בשיטה הנקראת סווג אחד כנגד השאר,
או one versus all classification



סווג רב-מחלקתי

one versus all classification

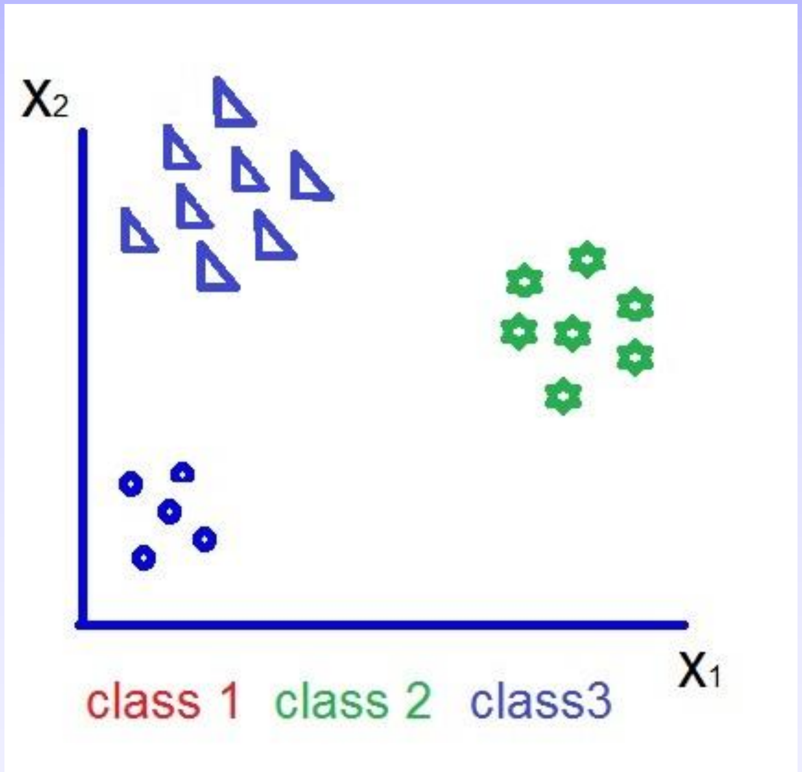
נאחד את מחלקות 2 ו-3, ונסמן אותן כמחלקה (0) או (-)
ואת דוגמאות מחלקה 1 כמחלקה (1) או (+).
נתאים לדוגמאות אלה מסווג בינארי, לדוגמא באמצעות רגרסיה
לוגיסטית $h_\theta^1(x) = p(y=1|\, x;\ \theta)$



סווג רב-מחלקתי

one versus all classification

נאחד את מחלקות 1 ו-3, ונסמן אותן כמחלקה (0) או (-)
ואת דוגמאות מחלקה 2 כמחלקה (1) או (+).
נתאים לדוגמאות אלה מסווג בינארי, לדוגמא באמצעות רגרסיה לוגיסטית $h_\theta^{\,2}(x) = p(y=2\,|\,x;\ \theta)$



סווג רב-מחלקתי

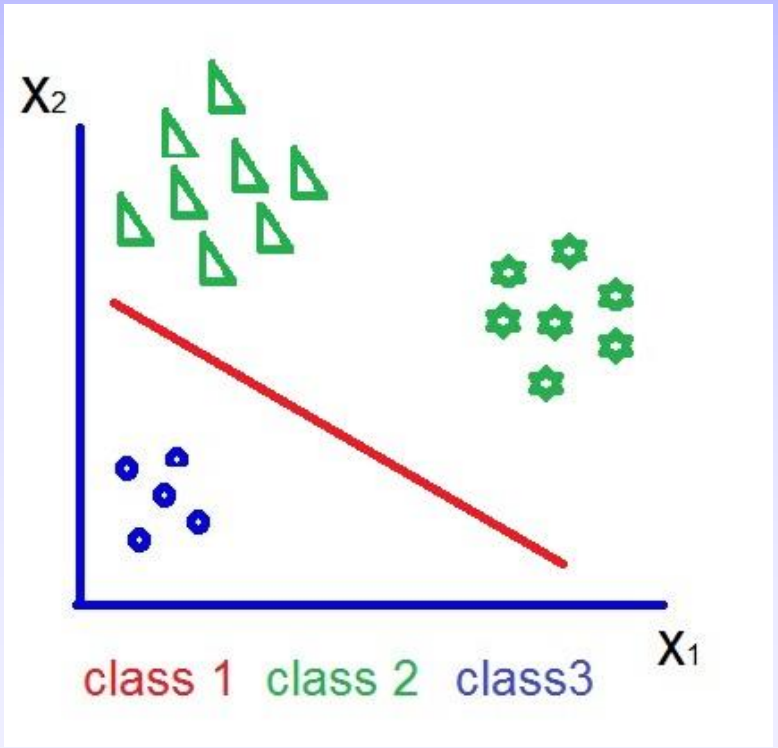
one versus all classification

נאחד את מחלקות 1 ו-2, ונסמן אותן כמחלקה (0)

ואת דוגמאות מחלקה 3 כמחלקה (1) או (+).

נתאים לדוגמאות אלה מסווג בינארי באמצעות רגרסיה לוגיסטית

$$h_\theta^{3}(x) = p(y=3 \mid x;\ \theta)$$



סווג רב-מחלקתי

one versus all classification

התאמנו 3 מסווגים המשערכים את ההסתברות לקבל את y=i
i=1,2,3 בהינתן הדוגמאות X: $h_\theta^{i}(x) = p(y=i \mid x;\ \theta)$

כדי לבצע חיזוי עבור קלט חדש x בחרו את המחלקה עבורה מתקבל המקסימום של הביטוי:

$$\max_i h_\theta^{(i)}(x) = \max_i (p(y=i \mid x;\ \theta))$$