Lec6 - Classification

Monday, 8 July 2024 16:15

מטרה: סיווג בינארי
2 אפשרויות/קבוצות

דוגמה: נתונים דו-ממדיים (x_1, x_2)

$x_1, x_2,$	y
⋮	0
⋮	0
⋮	⋮
×	1
⋮	1
	1
	1

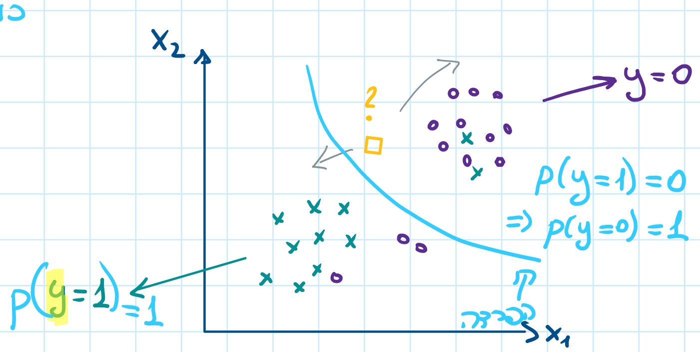
train dataset

מודל סיווג:

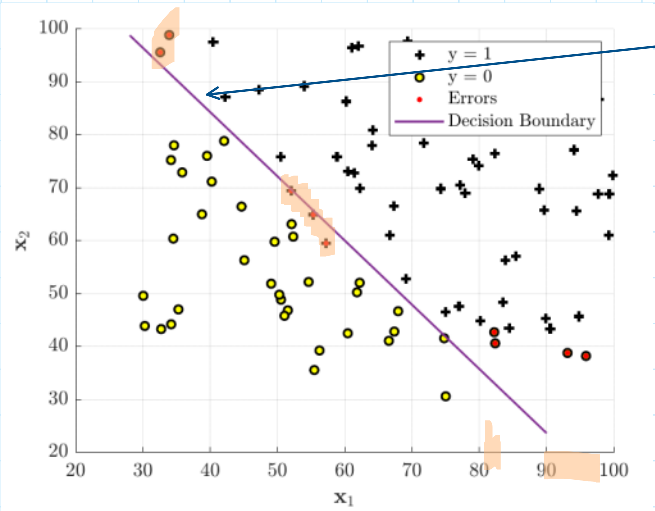
$$\hat{y} = f_{\underline{w}}(x_1, x_2)$$

תחזית מודל — פרמטרים של מודל — כניסה

פרשנות של מודל:
* קו הפרדה (ישר) או (עקום) בין הקבוצות



משפחה של מודלים: Generalized Binary Linear Classification Models



שאלה: מה משוואת קו ישר במקרה דו-ממדי:

$$\omega_1 x_1 + \omega_2 x_2 + \omega_0 = 0$$

$$\Rightarrow \omega_2 = -\frac{\omega_1}{\omega_2} x_1 - \frac{\omega_0}{\omega_2}$$

החלטה: מעל הקו או מתחת לקו ע"י מכפלה וקטורית

$$\mathbf{x}^T \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos(\theta)$$

וקטור מאונך קו ישר — נקודה — החלטה לפי סימן הזוית

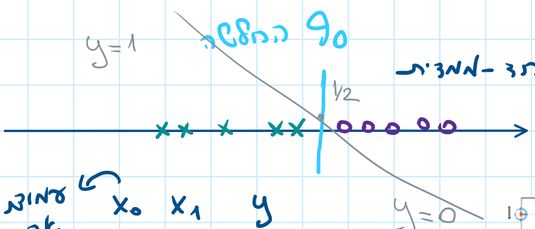
$$x^T y \lessgtr 0$$

מודל בסיסי: linear regression

$$\tilde{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

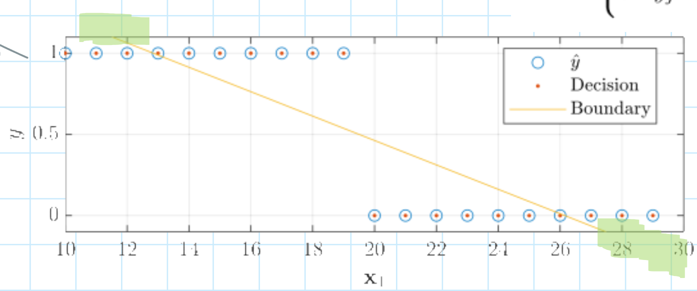
$$\hat{y}_j = \begin{cases} 1 & \tilde{y}_j > \frac{1}{2} \\ 0 & \tilde{y}_j < \frac{1}{2} \end{cases}$$

דוגמה מספרית חד-ממדית

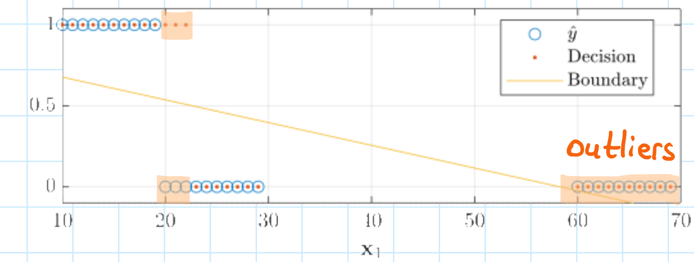


עמודת אחדות

x_0	x_1	y
1	10	1
⋮	⋮	⋮
1	19	1
1	20	0
⋮	⋮	⋮
1	29	0

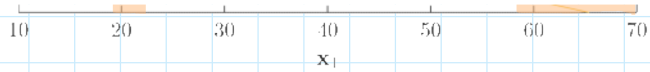


* ערכים יכולים להיות גדולים מאחד באופן שרירותי.



סקלה

* רגישות משמעותית מאוד ל- outliers



Logistic Model

- Linear model
- Outliers handling
- Probabilistic interpretation

$\hat{y} \sim p(y=1)$

סקלר

משפחה: $\hat{y}_i = g(\mathbf{w}^T\mathbf{x}_i)$ ← תוצאה

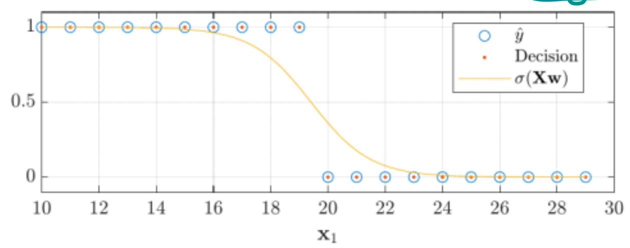
סיווג   פרמטרים   נתון לסיווג

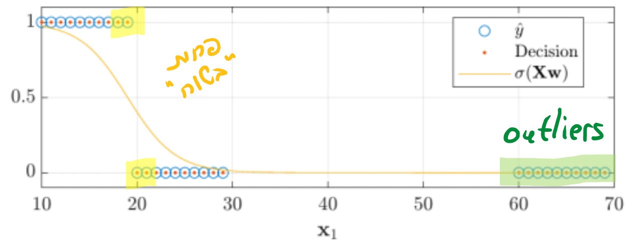
linear regression   מודל - $g(x) = x$ basic linear model

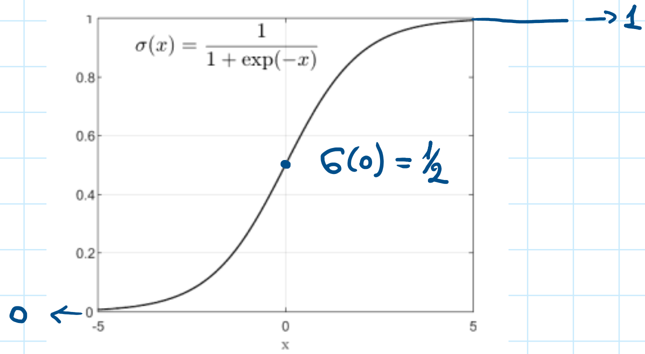
logistic regression $g(x) = \frac{1}{1-\exp(x)}$ sigmoid $= \sigma(x)$

$0 \leq \sigma(x) \leq 1$

$$\sigma(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$







Logistic regression:

ערך יחיד $\hat{y}_i = \sigma(\mathbf{w}^T\mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}_i)}$

רישום וקטורי $\hat{\mathbf{y}} = \sigma(\mathbf{Xw})$

איך מוצאים w ?

MSE → $\mathscr{L}(\cdot) = \frac{1}{2M}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$

* אין פתרון אנליטי

* יש הרבה נקודות מינימום מקומי

לא עושים

Cross-entropy loss

פונ' מחיר עם פרשנות הסתברותית

Entropy: For the discrete distribution $P = \{p_i = \Pr[X = x_i]\}$, the entropy is given by

$$H(P) = -\sum_i p_i \log(p_i) \qquad (7.11)$$

ln

$$p_1 = p_2 = \frac{1}{2} \Rightarrow H(p) = -2 \cdot \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

$$p_1 = \frac{1}{10}, p_2 = \frac{9}{10} \Rightarrow H(p) = -\frac{1}{10}\log_2\left(\frac{1}{10}\right) - \frac{9}{10}\log_2\left(\frac{9}{10}\right) \approx 0.4690$$

מדד: כמות של אי-וודאות:

* "יותר" אקראיות ⇐ יותר $H(p)$

$P_i = P_j$ * מקסימום של אקראיות עבור התפלגות אחידה

Cross-entropy: For two discrete distributions, p and q, the cross-entropy is given by

$$H(p,q) = -\sum_i p_i \log(q_i) \qquad (7.12)$$

Binary Cross-Entropy (BCE) Loss

$i = \{0,1\}$

$$H(p,q) = -p_0\log(q_0) - p_1\log(q_1)$$

* מינימום של $H(p,q)$ עבור $p_i = q_i$

$\Rightarrow H(p,q) = H(p)$

* לא מצב אחר $H(p,q)$ יורד

* מונים של נשן - סיווג - בינארי.

* ניתן להסתכל על y כהסתברות

$y = 1 \Rightarrow p(y=1) = 1$ $p(y=1)$ ב

$\Rightarrow H(p,q) = H(p)$

* כל מצב אחר $H(p,q)$ יורד

$y=1 \Rightarrow p(y=1)=1$

$y=0 \Rightarrow p(y=1)=0$

$p(y=1)$ של

רישום הסתברותי של נתונים

$$p_0 = \Pr(y=0) = 1-y$$
$$p_1 = \Pr(y=1) = y$$
$$q_0 = \Pr(\hat{y}=0) = 1 - f_w(\mathbf{x})$$
$$q_1 = \Pr(\hat{y}=1) = f_w(\mathbf{x})$$

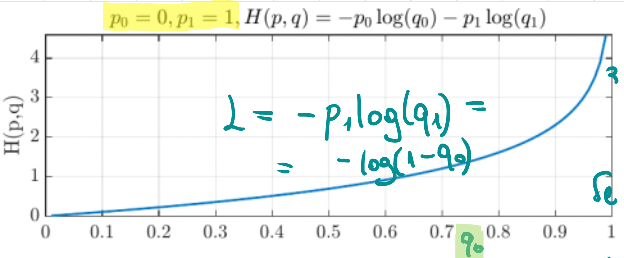
מוצא המודל

מצב רצוי

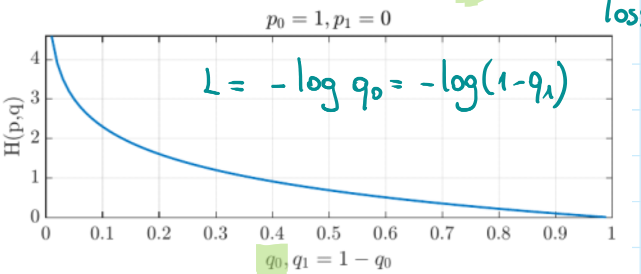
$f_{\underline{w}}(x) \to 1$

מקבל מינימום עבור $p_i = q_i$

$$H(p,q) = -p_0\log(q_0) - p_1\log(q_1)$$
$$= -(1-y)\log\big(1-f_w(\mathbf{x})\big) - y\log\big(f_w(\mathbf{x})\big)$$



$L = -p_1\log(q_1) =$
$= -\log(1-q_0)$



$L = -\log q_0 = -\log(1-q_1)$

BCE loss: Binary cross-entropy (BCE) loss function

ערך בודד loss

$$\mathscr{L}(y,\hat{y}) = -(1-y)\log(1-\hat{y}) - y\log(\hat{y}) \qquad (7.14)$$

ממוצע של ערכי loss

$$\mathscr{L} = -\frac{1}{M}\sum_{j=1}^{M}(1-y_i)\log(1-\hat{y}_i) + y_i\log(\hat{y}_i)$$
$$= -\frac{1}{M}\left[(1-\mathbf{y})\log(1-\hat{\mathbf{y}}) + \mathbf{y}\log(\hat{\mathbf{y}})\right]$$

* גזיר
* רציף
* מינימום גלובלי יחיד

פוני מחיר הכי נפוצה בשיטות סיווג

סיכום

פוני מחיר

$$\mathscr{L} = \frac{1}{M}\left[-\mathbf{y}^T\log\big(\sigma(\mathbf{X}\mathbf{w})\big) - (1-\mathbf{y})^T\log\big(1-\sigma(\mathbf{X}\mathbf{w})\big)\right]$$

$$\nabla_{\mathbf{w}}\mathscr{L}(\mathbf{w}) = \frac{1}{M}\mathbf{X}^T\left(\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y}\right)$$

$\nabla_{\underline{w}} L = 0$

אין פתרון אנליטי

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \alpha\nabla_{\mathbf{w}}\mathscr{L}(\mathbf{w})$$

Gradient descent

חישוב מינימום ע"י נגזרת

* מינימום גלובלי יחיד (BCE)
* כמון תומך regularization

Probabilistic prediction:

$$p(y=1|\mathbf{x},\mathbf{w}) = \sigma(\tilde{\mathbf{x}}\mathbf{w})$$
$$p(y=0|\mathbf{x},\mathbf{w}) = 1 - \sigma(\tilde{\mathbf{x}}\mathbf{w}) \qquad (7.16)$$

Classification decision: $\hat{y} \lessgtr \frac{1}{2}$

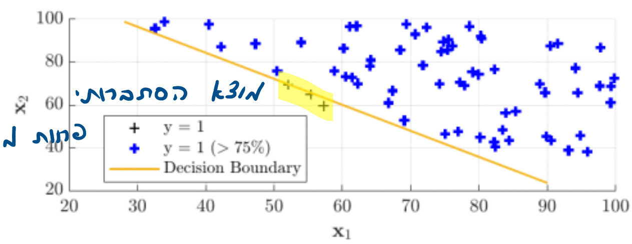
Another way:

$$\hat{\mathbf{y}} = \begin{cases} 1 & \mathbf{x}^T\mathbf{w} \geq 0 \\ 0 & \mathbf{x}^T\mathbf{w} < 0 \end{cases}$$

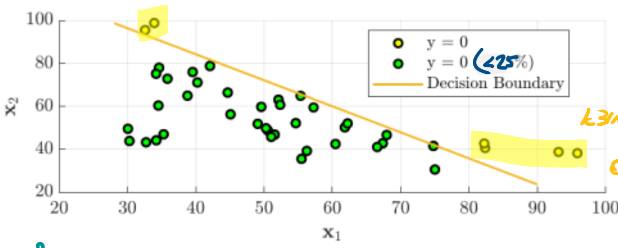
סימן $\cos(\theta)$

- Regularization can be applied,

$$\mathscr{L}_{new} = \mathscr{L}(\mathbf{y},\hat{\mathbf{y}}) + \frac{\lambda}{2M}\sum_{i=1}^{N}w_i^2 \qquad (7.21)$$



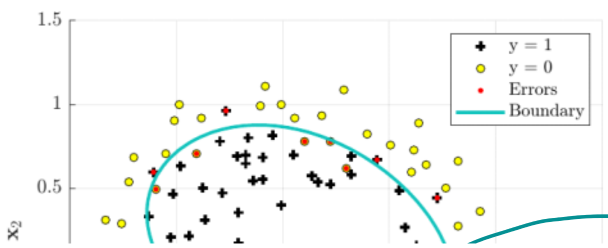
מוצא הסתברותי פחות מ. 0.75



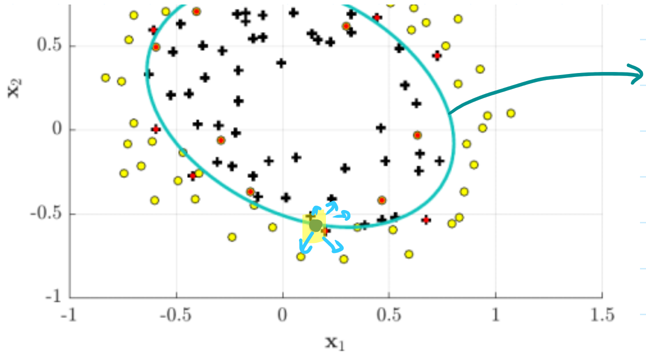
מוצא מעל 0.25

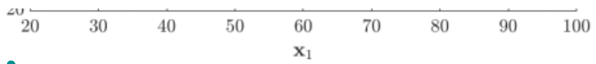
* תומך מיפוי פולינומיאלי וכו'

$$\varphi(x_1,x_2) = \langle 1, x_1, x_1^2, x_2, x_2^2, x_1x_2, x_1^2x_2, x_1x_2^2, x_1^2x_2^2\rangle$$



אליפסה





אליפסה

k-NN

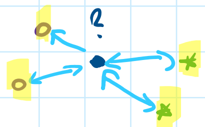
nearest neighbors

שיטה: החלטה ע"פ רוב "השכנים"
מתוך k הכי קרובים
k=1 "שכן" הכי הקרוב (שכן יחיד)

* שיטה למדידת מרחק:
היפר-פרמטר של השיטה

* תיקו: 2 שכנים סיווג 0
2 " 1
- החלטה אקראית
- ע"פ שכן הכי קרוב

* עובד על מספר שרירותי של קבוצות סיווג



בידועים
⇐ ב-train בקלות מגיעים ל-100% עבור 1-NN

- Euclidean distance metric,

$$d(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_M-b_M)^2} \quad (7.22)$$

- City block (Manhattan) distance

$$d(\mathbf{a},\mathbf{b}) = \sum_{j=1}^{M}|a_j-b_j| = |a_1-b_1|+\cdots+|a_M-b_M| \quad (7.23)$$

- Minkowski distance with (hyper) parameter p,

p=1
p=2

$$d(\mathbf{a},\mathbf{b}) = \sqrt[p]{\sum_{j=1}^{M}|a_j-b_j|^p} \quad (7.24)$$

* שיטה משמשת לתוצאות ראשוניות
baseline performance

* סיבוכיות חישוב:
עבור כל נקודה ב test
נדרש חישוב M מרחקים
של train

* לא תומך מתמודד ב outliers

* חשוב לנרמל נתונים