מסווגים ליניאריים – הרצאה מספר 6

- תזכורת: אנו עוסקים במסווגים ליניאריים העוברים דרך הראשית:

$$f(x,\theta) = sign(\underline{\theta}^T \underline{x})$$

- כאשר

$$\theta \in R^d$$

- הפרמטרים אותם רוצים לשערך על-פי נקודות האימון: (לדוגמא תמונות, אתרי אינטרנט, וכו').

הפרספטרון – התכנסות והכללה

נקודות האימון: $x^1, x^2, \ldots, x^m$

והתגיות המתאימות: $y^1, y^2, \ldots, y^m$

$"+1"$

$"+1"$

$"+1"$

$"-1"$

$"-1"$

נשתמש באלגוריתם הפרספטרון כדי לפתור את בעיית השערוך (estimation problem).

לימוד מדוגמאות

בתמונה עשויות להופיע פנים שונות, והמטרה היא למצוא וקטור של מקדמים כך שהפרספטרון יוכלו למצוא את הפנים הרצויות

לצורך הלימוד נתון אוסף תמונות $x^{i}\ {}_{i=1}^{m}$

שבחלקן מופיע האדם הרצוי, $y^{i}=+1$

ובחלקן פנים של אנשים אחרים $y^{i}=-1$

המתכנן של המערכת קובע את ערכי y^{i}

לימוד מדוגמאות

- סכימת הלימוד היא לימוד מדוגמאות (לימוד אינדוקטיבי).
- אוסף הדוגמאות הוא
$$\{x^1, y^1\}, \{x^2, y^2\}, \ldots, \{x^m, y^m\}$$
- $x^1, x^2, \ldots, x^m$ הם וקטורי הכניסה או תבניות הכניסה
- $y^1, y^2, \ldots, y^m$ הם התגיות או התוויות (labels) המתאימות
- כאשר: $y^i \in \{-1, 1\}$

הפרספטרון – התכנסות והכללה

- נסמן: k - מספר העדכונים של וקטור הפרמטרים θ שבצענו.
  $\theta^{(k)}$ - וקטור הפרמטרים לאחר k עדכונים.
- באופן התחלתי: עבור $k=0$ $\theta^{(k)}=0$
- האלגוריתם עובר על כל הדוגמאות ומעדכן את הפרמטרים רק בתגובה לשגיאות, כלומר כאשר התבנית חזויה באופן שגוי.

לימוד מדוגמאות

- איטרציה:

$$\theta^{(k+1)} = \theta^{(k)} + y^t \underline{x}^t$$

- כאשר:

$$y^t (\theta^{(k)})^T \underline{x}^t < 0$$

- אחרת הפרמטרים נשארים ללא שינוי.

הפרספטרון – התכנסות והכללה

עתה נוכיח שאלגוריתם הפרספטרון מתכנס עם מספר סופי של עדכונים.

הנחה: לכל התבניות (תמונות, אתרי אינטרנט וכו') נורמה אאוקלידית חסומה:

$$\| x^t \| < R$$

לכל t ועבור R סופי כלשהו.

(זהו כמובן המקרה עבור כל תמונה ספרתית או תבנית (וקטור תכונות) עם ערכי עצמה חסומים).

אנליזה דומה תראה כיצד המסווג הליניארי מתכנס עבור תמונות או תבניות חדשות שלא נכללו בדוגמאות האימון.

הפרספטרון – התכנסות והכללה

- הנחה ראשונה: $\| x^t \| < R$
- הנחה שניה: קיים מסווג ליניארי עבור דוגמאות האימון שלנו, עם ערכי פרמטר סופיים, המסווג נכונה את כל דוגמאות האימון.
- כלומר במלים אחרות מניחים כי קיים ערך γ כך ש:

$$y^t (\theta^{(*)})^T \underline{x}^t \geq \gamma$$

- התוספת $\gamma > 0$ היא כדי להבטיח כי כל דוגמא מסווגת נכונה עם שוליים סופיים (finite margin)

הפרספטרון – התכנסות והכללה

הוכחה: נצרף שתי תוצאות:

1) המכפלה הפנימית $(\theta^{(*)})^T \theta^{(k)}$

גדלה לפחות באופן ליניארי עם כל עדכון

2) הנורמה הריבועית $\| \theta^{(*)} \|^2$ גדלה לכל היותר ליניארית עם כל עדכון.

על-ידי צירוף של שתי התוצאות נראה שקוסינוס הזוית בין θ^* ל- $\theta^{(k)}$ גדלה עם תוספות סופיות עבור כל עדכון.

מאחר ו- $|\cos(\theta)| \leq 1$ נקבל שאפשר לבצע רק מספר סופי של עדכונים.

הפרספטרון – התכנסות והכללה

על-ידי צירוף של שתי התוצאות נראה שקוסינוס הזוית[1] בין θ^* ל- $\theta^{(k)}$ גדל עם תוספות סופיות עבור כל עדכון.

מאחר ו- $|\cos(\theta)| \le 1$ נקבל שאפשר לבצע רק מספר סופי של עדכונים.

[1] הגדרה:
לכל שני וקטורים x,y נסמן:

$$\cos(x,y) = \frac{x^T y}{\|x\| \cdot \|y\|}$$

הפרספטרון – התכנסות והכללה

טענה ראשונה:

נתבונן במכפלה הפנימית לפני ואחרי כל עדכון:

כאשר העדכון הוא עבור הצעד ה- k, נניח בגלל שגיאה על התבנית x_t

$$(\theta^*)^T \theta^{(k)} =$$

$$(\theta^*)^T (\theta^{(k-1)} + y^t x^t) =$$

$$(\theta^*)^T \theta^{(k-1)} + y^t (\theta^*)^T x^t \geq (\theta^*)^T \theta^{(k-1)} + \gamma$$

הפרספטרון – התכנסות והכללה

נרשום שוב:

$$(\theta^*)^T \theta^{(k)} =$$

$$(\theta^*)^T (\theta^{(k-1)} + y^t x^t) =$$

$$(\theta^*)^T \theta^{(k-1)} + y^t (\theta^*)^T x^t \geq (\theta^*)^T \theta^{(k-1)} + \gamma$$

תזכורת: $\gamma > 0$ משמש להבטיח שכל דוגמא מסווגת באופן נכון עם שוליים סופיים.

הסבר: לפי ההנחה θ^* הוא הוקטור עבורו אין שגיאות כלל, ולכן:

$$y^t (\theta^{(*)})^T \underline{x}^t \geq \gamma$$

הפרספטרון – התכנסות והכללה

לפיכך, לאחר k עדכונים:

$$(\theta^*)^T \theta^{(k)} \geq (\theta^*)^T \theta^{(k-1)} + \gamma$$

$$\geq (\theta^*)^T \theta^{(k-2)} + 2\gamma$$

$$\geq ...(\theta^*)^T \theta^{(k-(k-1))} + (k-1)\gamma$$

$$\geq (\theta^*)^T \theta^{(0)} + k\gamma \geq k\gamma$$

כלומר:

$$(\theta^*)^T \theta^{(k)} \geq k\gamma$$

הפרספטרון – התכנסות והכללה

טענה שניה:

$$\|\theta^{(k)}\|^2 = \|\theta^{(k-1)} + y^t x^t\|^2$$

$$= \|\theta^{(k-1)}\|^2 + 2y^t(\theta^{(k-1)})^T x^t + |y^t| \cdot \|x^t\|^2$$

$$= \|\theta^{(k-1)}\|^2 + 2y^t(\theta^{(k-1)})^T x^t + \|x^t\|^2$$

$$\leq \|\theta^{(k-1)}\|^2 + \|x^t\|^2$$

$$\leq \|\theta^{(k-1)}\|^2 + R^2$$

כלומר:

$$\|\theta^{(k)}\|^2 \leq \|\theta^{(k-1)}\|^2 + R^2$$

הפרספטרון – התכנסות והכללה

הסבר:

מאחר ו- $y^t(\theta^{(k-1)})^T x^t < 0$

(כי אחרת לא היה מתבצע עדכון בשלב ה- k).

וכן לפי ההנחה: $\| x^t \| \leq R$

מאחר ואפשר להמשיך ולפרק את הנורמה:

$$\| \theta^{(k)} \|^2 \leq \| \theta^{(k-1)} \|^2 + R^2 \leq \| \theta^{(k-2)} \|^2 + 2R^2$$

$$\ldots \leq \| \theta^{(k-k)} \|^2 + k \cdot R^2$$

מקבלים: $\| \theta^{(k)} \|^2 \leq k \cdot R^2$

הפרספטרון – התכנסות והכללה

עתה אפשר לצרף את טענות 1) ו- 2) ולקבל:

$$\cos(\theta^*, \theta^{(k)}) = \frac{(\theta^*)^T \theta^{(k)}}{\|\theta^*\| \cdot \|\theta^{(k)}\|}$$

$$\geq^{(1)} \frac{k \cdot \gamma}{\|\theta^*\| \cdot \|\theta^{(k)}\|}$$

$$\geq^{(2)} \frac{k \cdot \gamma}{\sqrt{kR^2}\, \|\theta^*\|}$$

(הסבר: לפי ההגדרה: $\cos(x, y) = \dfrac{x^T y}{\|x\| \cdot \|y\|}$)

הפרספטרון – התכנסות והכללה

מאחר וקוסינוס חסום על-ידי 1, מקבלים:

$$1 \geq \cos(\theta^*, \theta^{(k)}) \geq \frac{k \cdot \gamma}{\sqrt{kR^2 \| \theta^* \|}}$$

או:

$$\sqrt{k} \leq \frac{R \cdot \| \theta^* \|}{\gamma}$$

$$k \leq \frac{R^2 \cdot \| \theta^* \|^2}{\gamma^2}$$

שוליים וגאומטריה

מצאנו כי:

$$k \le \frac{R^2 \cdot \| \theta^* \|^2}{\gamma^2}$$

האם הגודל $\dfrac{\| \theta^* \|^2}{\gamma^2}$

קשור לכמה בעיית הסווג קשה?

נראה בהמשך כי קיים קשר בין $\dfrac{\| \theta^* \|^2}{\gamma^2}$ לבין קושי בעיית הסווג.

שוליים וגאומטריה

טענה: ההפכי של ערך זה, כלומר: $\dfrac{\gamma}{\|\theta^*\|}$

הוא המרחק הקטן ביותר במרחב התבניות

מכל תבנית למשטח ההפרדה (המסומן על-ידי θ^*)

במלים אחרות הגודל $\dfrac{\gamma}{\|\theta^*\|}$ משמש כמדד לכמה

מופרדות התבניות של שתי המחלקות (על-ידי משטח
הפרדה ליניארי).

נסמן שוליים אלה ב- γ_{geom}

שוליים וגאומטריה

לפיכך γ_{geom}^{-1} הוא מדד סביר לכמה הבעייה קשה.

ככל שהשוליים הגאומטריים המפרידים בין תבניות האימון הם יותר קטנים, כך הבעייה יותר קשה.

שוליים וגאומטריה

נחשב את γ_{geom}

נמדוד את המרחק בין משטח ההפרדה:

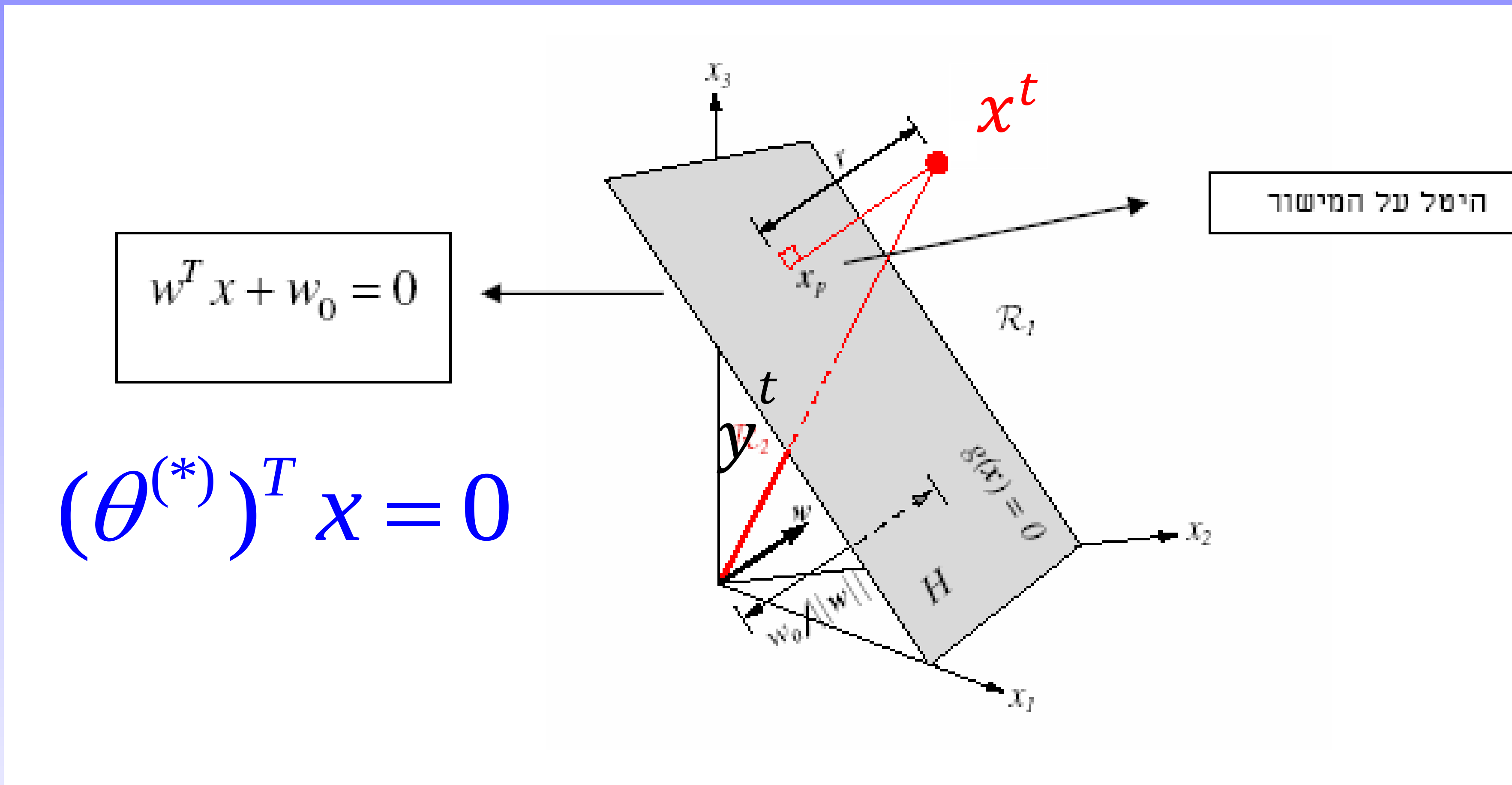
$$(\theta^{(*)})^T x = 0$$

לאחת התמונות (הדוגמאות) עבורה

$$y^t (\theta^{(*)})^T x^t = \gamma$$

מאחר ו- $\theta^{(*)}$ ניצב למשטח ההפרדה (נורמל), אזי המרחק הקצר ביותר בין המשטח ל- x^t הוא מקביל לוקטור הניצב $\theta^{(*)}$

שוליים וגאומטריה



כלומר הדוגמא x^t היא אחת הדוגמאות הקרובות ביותר למשטח ההפרדה. $(y^t(\theta^{(*)})^T x^t = \gamma)$

שוליים וגאומטריה

נסמן:

$$x_p = x^t - \frac{r \cdot y^t \cdot \theta^*}{\|\theta^*\|}$$

כאשר x_p הוא ההיטל של x^t על משטח ההפרדה $\theta^{(*)}$

וכאשר r הוא המרחק של הישר בין התבנית x^t לנקודה x_p על המשטח.

שוליים וגאומטריה

מהו הערך r עבורו

$$(\theta^{(*)})^T \cdot x_p = 0 \quad ?$$

או באופן שווה ערך :

$$y^t \cdot (\theta^{(*)})^T \cdot x_p = 0 \quad ?$$

שוליים וגאומטריה

$$y^t \cdot (\theta^{(*)})^T \cdot x_p = y^t \cdot (\theta^{(*)})^T [x^t - \frac{r \cdot y^t \cdot \theta^*}{\|\theta^*\|}]$$

$$= y^t \cdot (\theta^{(*)})^T x^t - y^t \cdot (\theta^{(*)})^T \frac{r \cdot y^t \cdot \theta^*}{\|\theta^*\|}$$

$$= y^t \cdot (\theta^{(*)})^T x^t - r \cdot \frac{\|\theta^*\|^2}{\|\theta^*\|} =$$

$$= \gamma - r \cdot \|\theta^*\|$$

שוליים וגאומטריה

לפיכך המרחק הוא

$$r = \frac{\gamma}{\| \theta^* \|}$$

ולכן אפשר לכתוב את החסם למספר העדכונים באופן יותר ממצה, במונחים של השוליים הגאומטריים:

$$k = \left(\frac{R}{\gamma_{geom}} \right)^2$$

בהבנה ש- γ_{geom} הוא השוליים הגאומטריים הגדולים ביותר שאפשר לקבל עם מסווג ליניארי עבור בעייה זו.

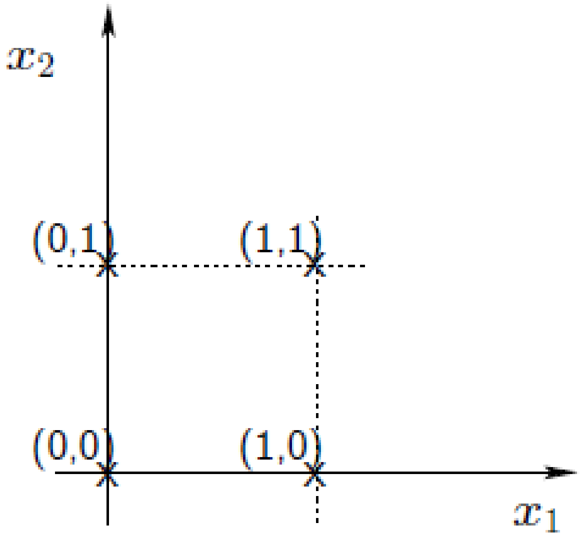
שוליים וגאומטריה

הסבר:

$$k \le \frac{R^2 \|\theta^*\|^2}{\gamma^2} = \frac{R^2}{\left(\frac{\gamma^2}{\|\theta^*\|^2}\right)} = \frac{R^2}{\gamma_{geom}^2} = \left(\frac{R}{\gamma_{geom}}\right)^2$$


perceptron
abc

The simple linear classifier cannot solve all the problems (e.g., XOR)



Non-Separable Data

מה עושים כאשר הנתונים הם לא פרידים ליניארית?

האם אלגוריתם לימוד הפרספטרון (PLA) יעצור ?

Non-Separable Data

- גם כאשר הנתונים הם לא פרידים ליניארית לעיתים נראה שעל-מישור (או ישר במקרה של 2 מימדים) עשוי לפתור את הבעייה.
- צריך לפתור את הבעייה הקומבינטורית הבאה:

$$\min_{\theta \in R^{d+1}} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} [sign(\theta^T x^{(i)}) \neq y^{(i)}]$$

זוהי בעייה המוכרת כבעיית NP קשה, ולכן נרחיב את אלגוריתם לימוד הפרספטרון.

אלגוריתם הכיס (pocket algorithm)

- האלגוריתם שומר את הפתרון הטוב ביותר – וקטור הפרמטרים עם השגיאה הנמוכה ביותר בה הוא נתקל עד לאיטרציה ה- t ב- PLA.
- התוצאה הסופית היא וקטור הפרמטרים המביא לשגיאה הנמוכה ביותר עד האיטרציה הנוכחית.

The Pocket Algorithm

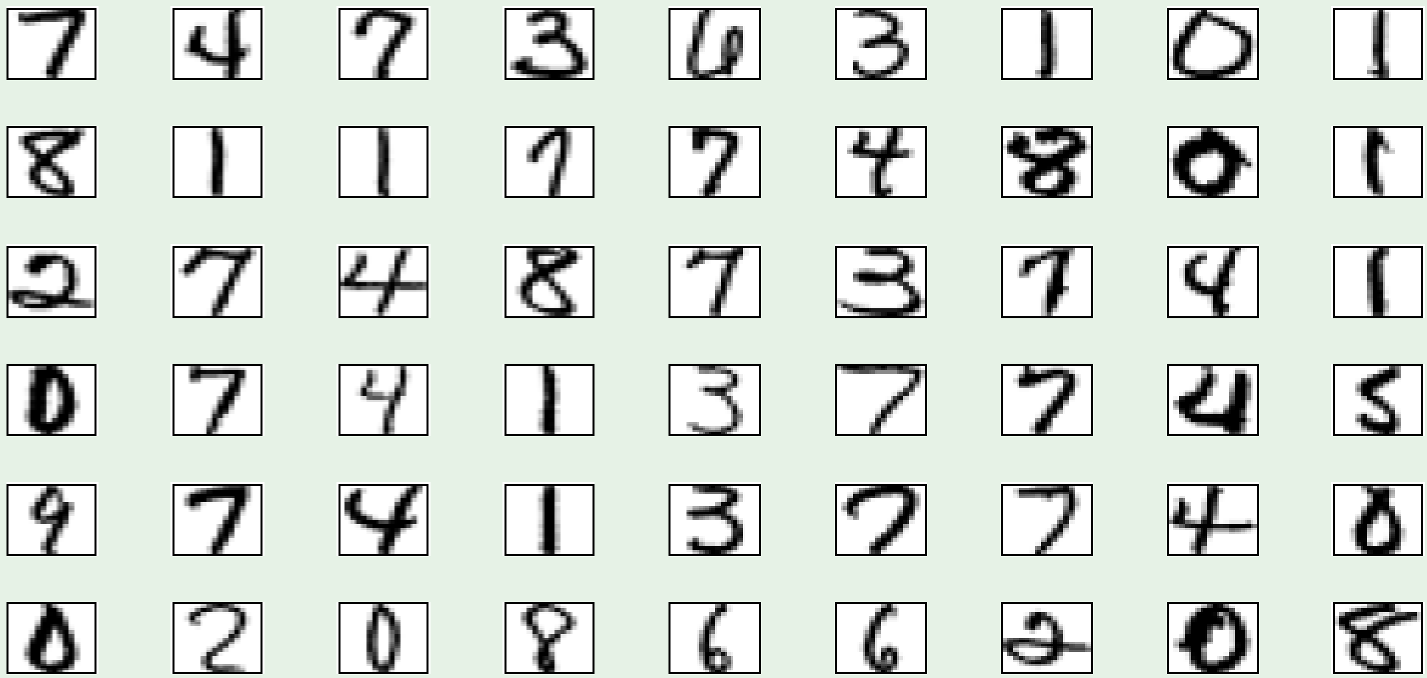
1. Set the pocket weight vector $\hat{\theta}$ to $\theta(0)$ of PLA
2. For t=0,1,…,T
3. Run PLA for one update to obtain $\theta(t+1)$
4. Evaluate $J(\theta(t+1))$
5. If $\theta(t+1)$ is better than $\hat{\theta}$ in terms of J

   set $\hat{\theta}$ to $\theta(t+1)$
6. Return $\hat{\theta}$

ההבדל בין האלגוריתמים:

- ה- PLA עובר על הדוגמאות ומשנה את הפרמטרים θ בכל איטרציה, בעוד שאלגוריתם ה- pocket דורש שלב נוסף בו עבור כל ערך של θ בכל איטרציה משערכים את J(θ), כלומר את פונקציית השגיאה, ולכן ה- pocket הרבה יותר איטי מה- PLA.
- לא מובטח באיזו מהירות ה- pocket יתכנס לפתרון טוב.
- אלגוריתם ה- pocket מאפשר הפרדה גם כאשר הנתונים הם לא פרידים ליניארית.

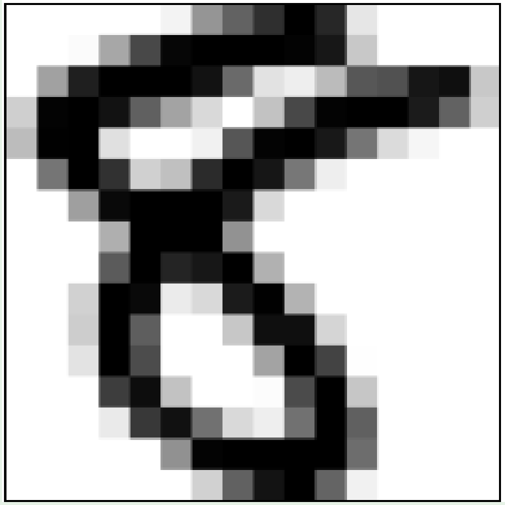
A real data set



Input representation

'raw' input $\mathbf{x} = (x_0, x_1, x_2, \cdots, x_{256})$

linear model: $(w_0, w_1, w_2, \cdots, w_{256})$



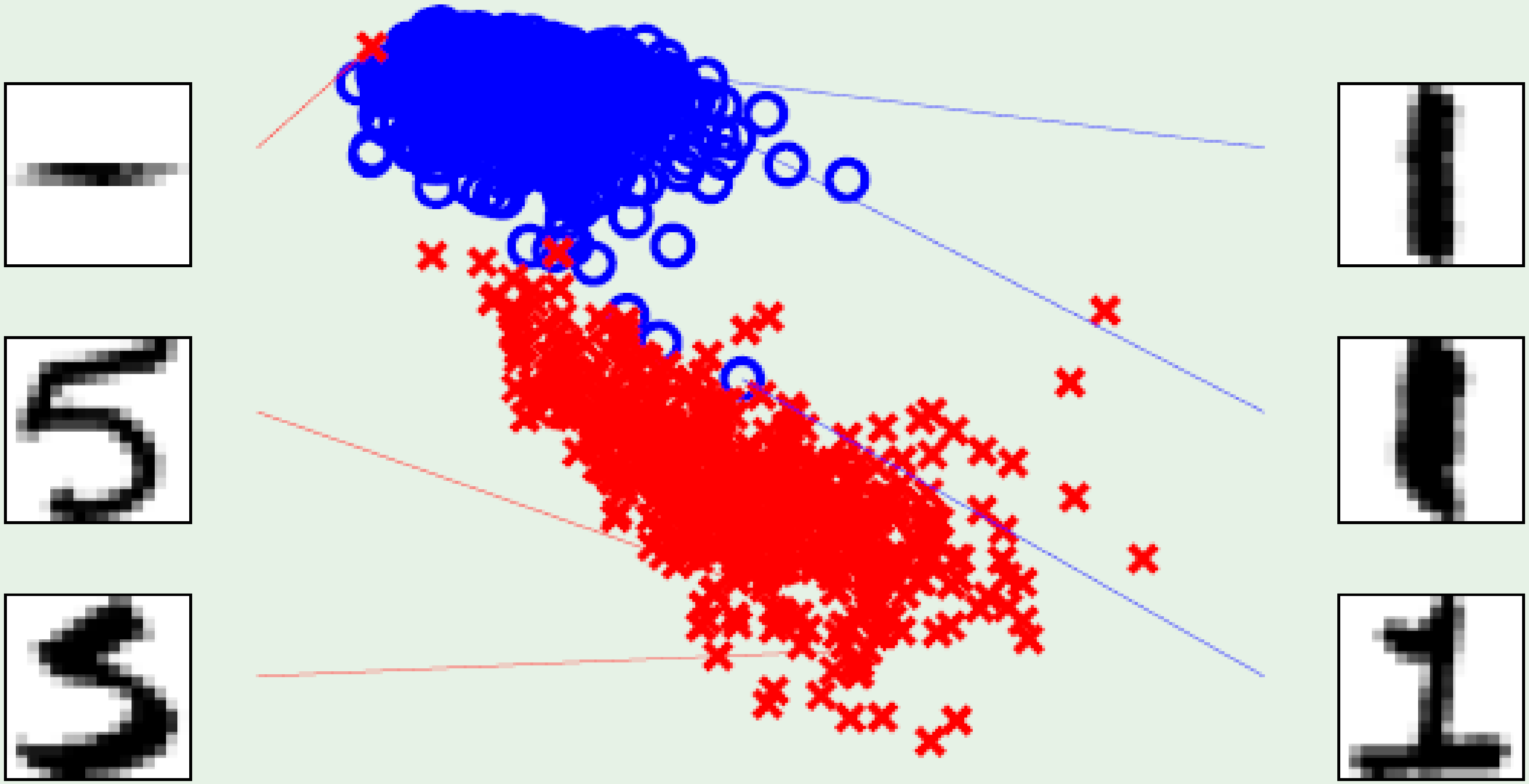
Features: Extract useful information, e.g.,

intensity and symmetry $\mathbf{x} = (x_0, x_1, x_2)$

linear model: (w_0, w_1, w_2)

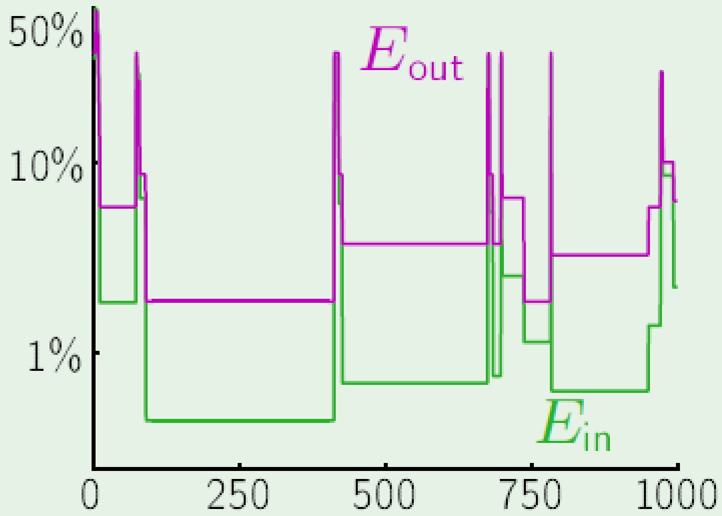
Illustration of features

$\mathbf{x} = (x_0, x_1, x_2)$ x_1: intensity x_2: symmetry

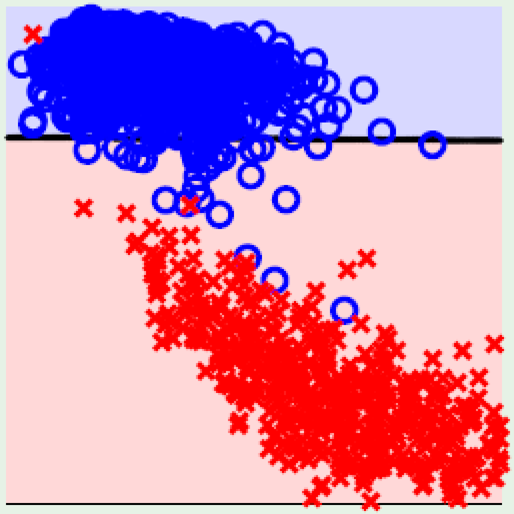


What PLA does

Evolution of E_{in} and E_{out}

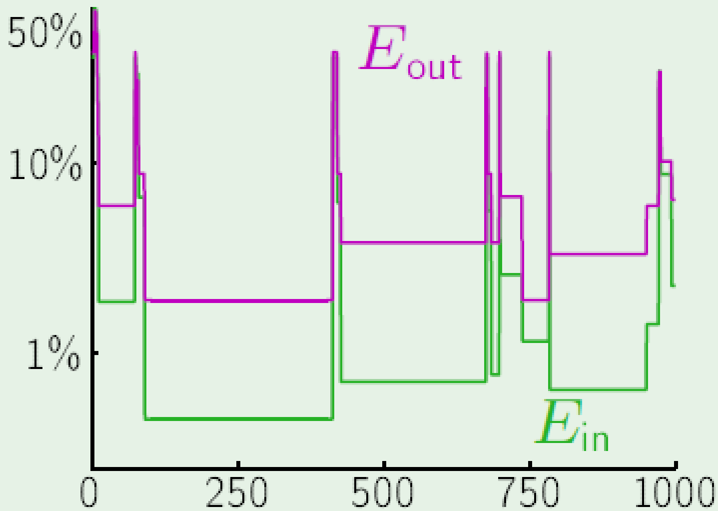


Final perceptron boundary

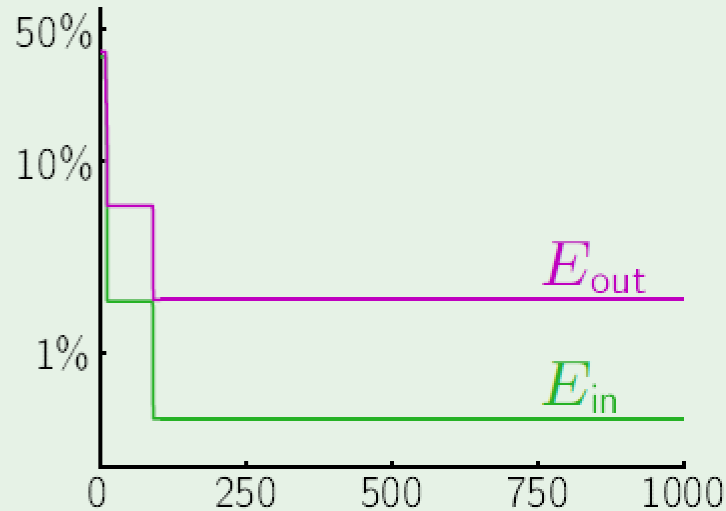


The ‘pocket’ algorithm

PLA:

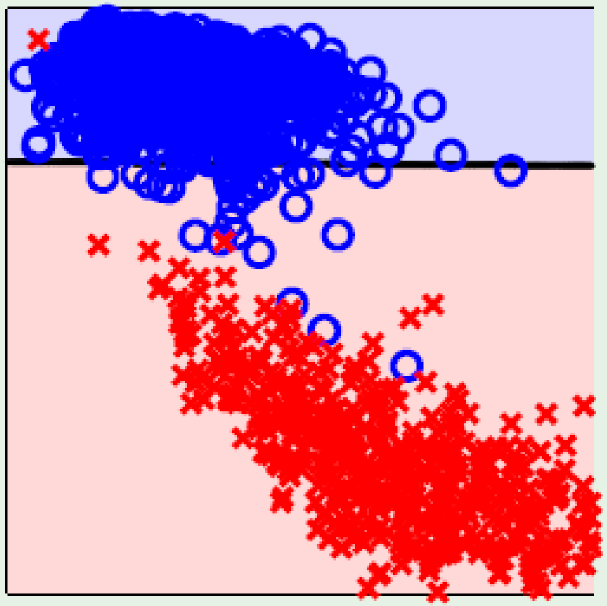


Pocket:



Classification boundary - PLA versus Pocket

PLA:



Pocket: