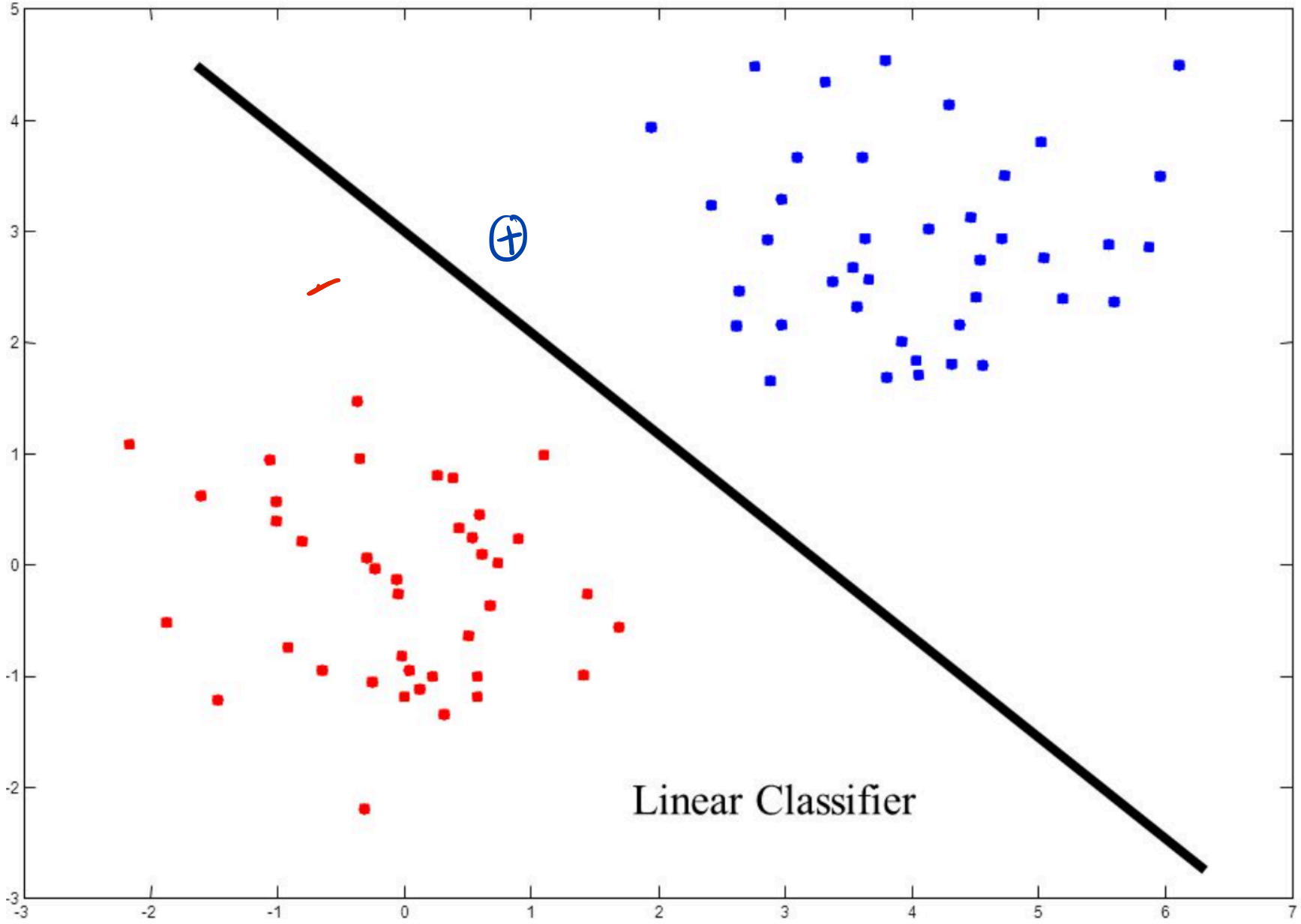
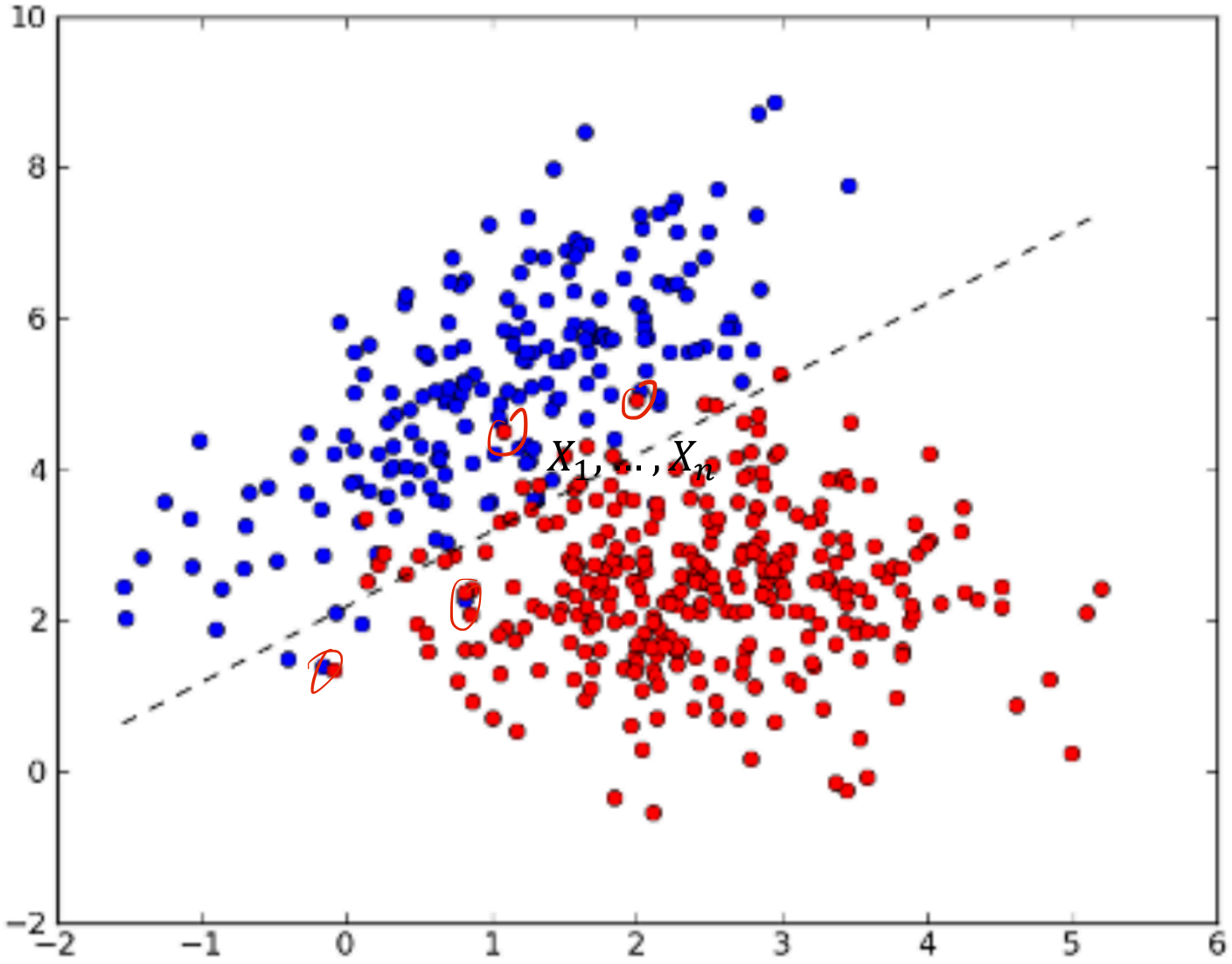
Linear Classifiers

ECE 4200

What is a linear classifier?



What is a linear classifier?



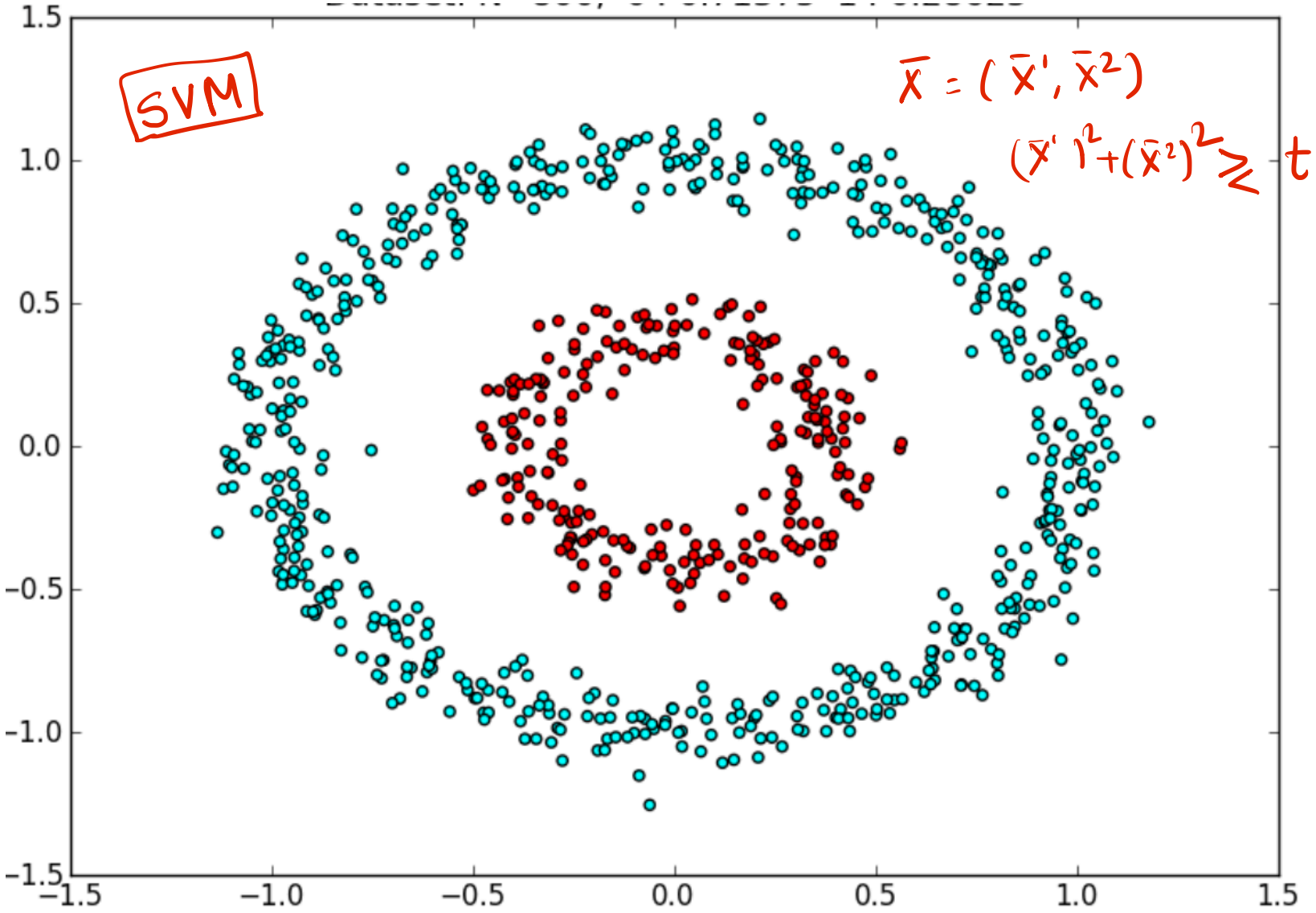
Advantages

easy to predict.

- Simple to describe
- Can use linear algebra
- Robust against over-fitting

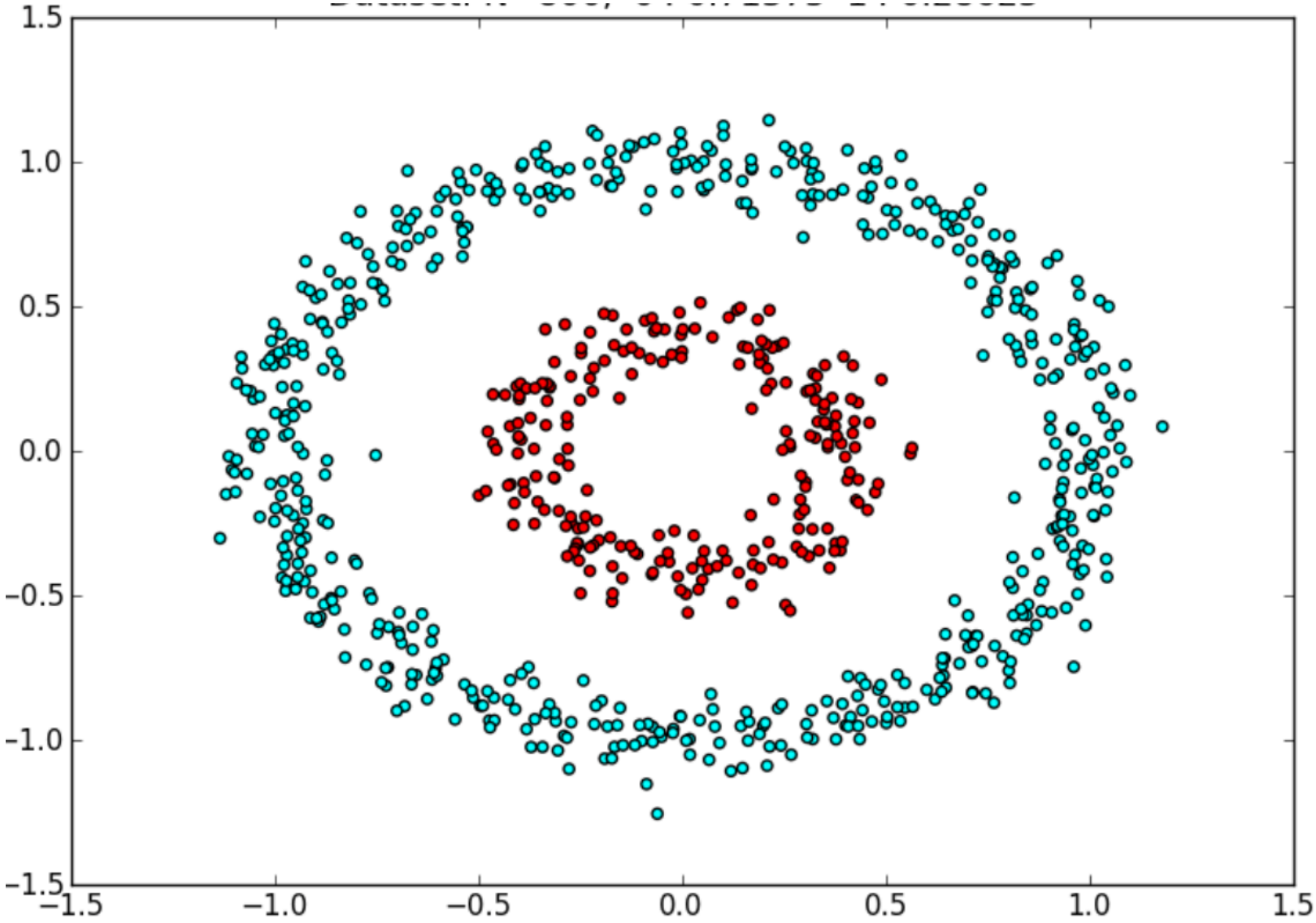
Disadvantages

- Too simple to capture complicated structures



Disadvantages

- Will see later how to use linear classifiers for this.



Examples

Perceptron

Today

Logistic Regression

Support Vector Machines

---- Linear Regression (not really classification, still linear)

Mathematically

Recall that features are points in $\mathbb{R}^d$

$(\bar{x}, y)$

A linear classifier is given by a hyper-plane in $\mathbb{R}^d$.

Classify new test examples based on which side of the hyper-plane they lie on.

Binary $\rightarrow \{+1, -1\}$

Hyperplane

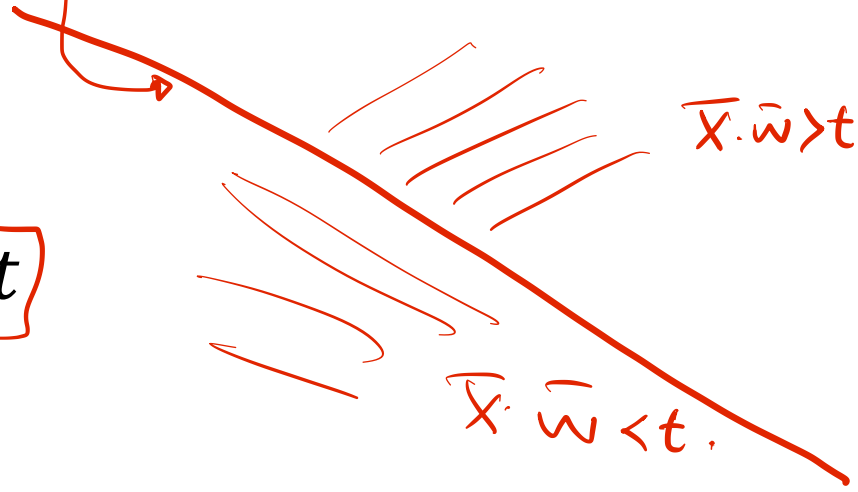
What is a hyperplane? Given $\bar{w}, t$, the set of all points $\bar{x}$ s.t.,

$$\bar{X} \cdot \bar{w} = t$$

** Note that $\bar{X}, \bar{w}$ are both vectors in $\mathbb{R}^d$, and $\bar{X} \cdot \bar{w}$ is their inner product (dot product).**

Decision rule:

$$\vec{X} \cdot \vec{w} \lessgtr t$$



WHAT IS THE $\bar{w}$?

** It is the vector perpendicular to the hyperplane.

Classification rule

What is a hyperplane? Given $\bar{w}, t$, the set of all points $\vec{x}$ s.t.,

$$\bar{x} \cdot \bar{w} = t$$

Decision rule:

$$\bar{x} \cdot \bar{w} \leqslant t$$

When labels are $\{+1, -1\}$:

$$\hat{y} = sign(\bar{x} \cdot \bar{w} - t).$$

What is the problem

How to find the **best** hyperplane?

Best :- the hyperplane that minimizes the number of misclassified training examples.

What is a good hyper-plane?

How to find a good hyper-plane?

- Find one that minimizes the # of training errors?
- Attempt to draw on board
- HARD to compute (NP HARD)

Canonical Examples

- The Perceptron

- Logistic Regression

- Support Vector Machines

Perceptron Algorithm

Frank Rosenblatt – at CORNELL in the 50's.

Is a linear classifier that guarantees to have zero training error hyper-plane when the data is **linearly separable.**

A set S is linearly separable if there is a hyper-plane correctly classifying all examples in S

How to find such a hyper-plane ?

Perceptron Algorithm

$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{o/w} \end{cases}$

- Input: $(\bar{X}_1, y_1), (\bar{X}_2, y_2), \ldots, (\bar{X}_n, y_n)$, $y_i \in \{+1, -1\}$.
- Output: $\bar{w}, t$
- Initialize: $\bar{w} = \bar{0}, t = 0$

$\bar{w} \cdot X - t \geq 0$

- REPEAT:
  - For $i = 1, \ldots, n$:
    - If $\text{sign}(\bar{w} \cdot \bar{X}_i - t) \neq y_i$:
      - $\bar{w} \leftarrow \bar{w} + y_i \cdot \bar{X}_i$
      - $t = t - y_i$

STOP when no errors.

Perceptron Algorithm

- Decision Rule:

$$\hat{y}_i = \text{sign}(\bar{w} \cdot \bar{X}_i - t)$$

What do we want?

$$\forall i \in [n], \quad \hat{y}_i = y_i$$

Equivalently:

$$\forall i \in [n], \quad y_i \cdot (\bar{w} \cdot \bar{X}_i - t) \geq 0$$

$$\Updownarrow$$

$$y_i = \text{sign}(\bar{w} \cdot \bar{x}_i - t)$$

Why does it work?

What happens for the example after one mistake?

i^{th} example. $(\bar{w}, t)$ ← current hyperplane.

$y_i \neq \text{sign}(\bar{w} \cdot \bar{x}_i - t) \Leftrightarrow y_i \cdot (\bar{w} \cdot \bar{x}_i - t) < 0$ — ①

$\bar{w}_{new} \leftarrow \bar{w} + y_i \cdot \bar{x}_i$, $t_{new} \leftarrow t - y_i$.

$$\bar{w}_{new} \cdot \bar{x}_i - t_{new} = (\bar{w} \cdot \bar{x}_i - t) + y_i (1 + \underbrace{\bar{x}_i \cdot \bar{x}_i}_{\|x_i\|^2})$$

$$= \underbrace{(\bar{w} \cdot \bar{x}_i - t)}_{-ve} + y_i (1 + \|x_i\|^2)$$

Ex:- $y_i = +1$,

$$\vec{w} \leftarrow \vec{w} + \boxed{y_i \cdot \vec{x}_i}$$

$$\vec{w} \cdot \vec{x}_i - t \longrightarrow y_i$$

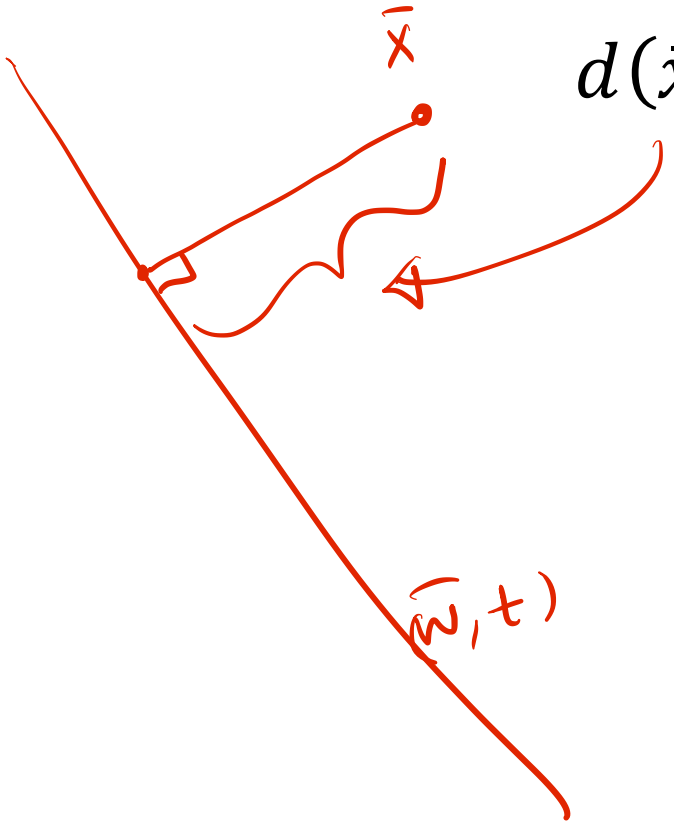
Distance from a hyperplane?

What is the distance of a point $\bar{x}$ from a hyperplane specified by $(\bar{w}, t)$?

$(10\bar{w}, 10 \cdot t)$

$-\bar{w} \cdot \bar{x} - t = 0$

$-(10\bar{w}) \cdot \bar{x} - (10t) = 0$



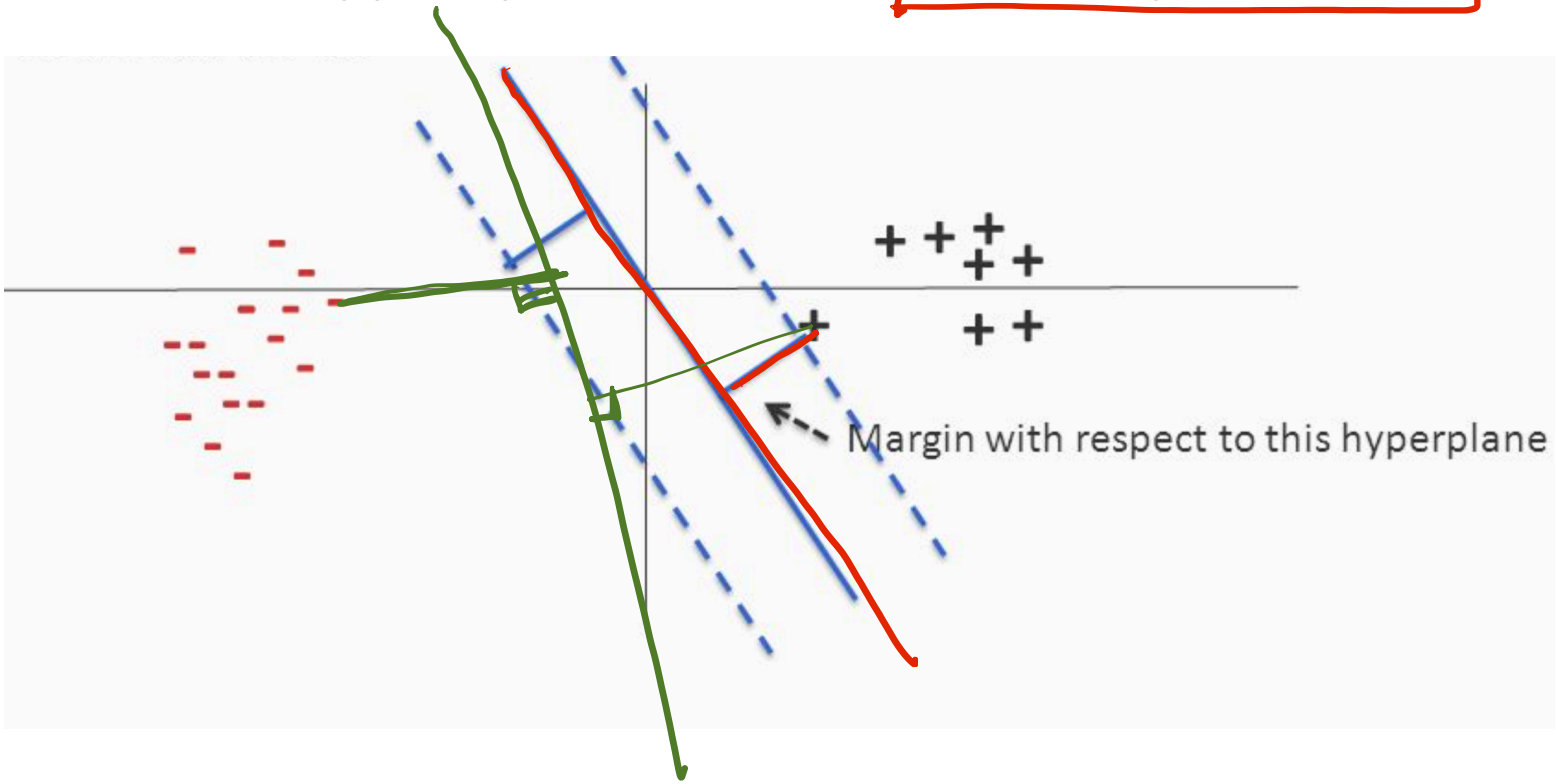
$$d(\bar{x}; \bar{w}, t) = \frac{|\bar{w} \cdot \bar{x} - t|}{\| \bar{w} \|}.$$

$\| \bar{w} \| = \| \bar{w} \|_2 = \sqrt{\bar{w}^{1^2} + \ldots + \bar{w}^{d^2}}$

Margin of a hyperplane

Suppose $S = (\bar{X}_1, y_1), (\bar{X}_2, y_2), \ldots, (\bar{X}_n, y_n)$ are linearly separable.

For a hyper plane $\vec{w}, t$ separating S, the margin is the distance of the hyper plane to the closest point in S.



Distance from a hyperplane?

What is the distance of a point $\bar{x}$ from a hyperplane specified by $\bar{w}, t$?

$$d(\bar{x}; \bar{w}, t) = \frac{|\bar{w} \cdot \bar{x} - t|}{\| \bar{w} \|}.$$

Margin of a separating hyperplane: $(\bar{w}, t)$

$$\text{Margin}(S, \bar{w}, t) := \min_{i \in [n]} d(\bar{X}_i; \bar{w}, t)$$

Margin of a data set

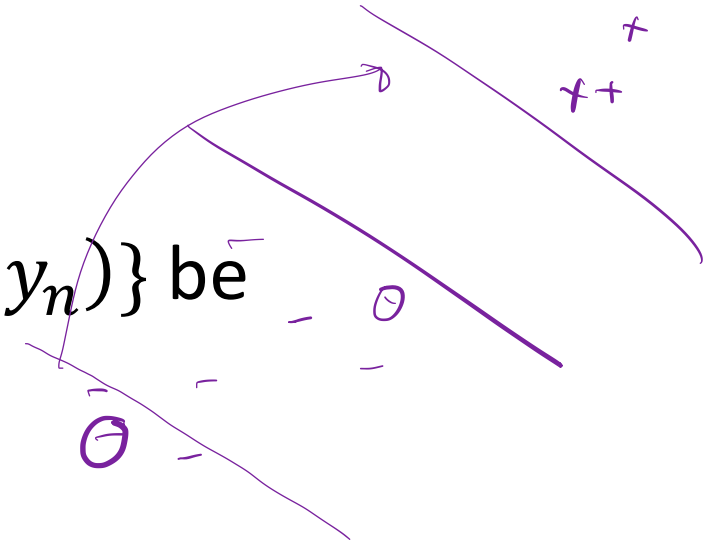
Suppose $S = (\bar{X}_1, y_1), (\bar{X}_2, y_2), \ldots, (\bar{X}_n, y_n)$ are linearly separable.

Margin of S is the **largest** margin of all separating hyper planes.

$$\max_{(\bar{w}, t)} \text{Margin}(S, (\bar{w}, t)) = \text{Margin}(S)$$

$(\bar{w}, t)$ ↓ separating hyperplane

Theorem for Perceptron



Suppose $S = \{(\bar{X}_1, y_1), (\bar{X}_2, y_2), \ldots, (\bar{X}_n, y_n)\}$ be s.t. $\| \bar{X}_i \|_2 \leq 1$, and $\text{margin}(S) = \gamma$.

Then the perceptron converges after at most

$$\frac{4}{\gamma^2}$$

updates.

How to prove the algorithm terminates?

Online Demo:

http://mlweb.loria.fr/book/en/perceptron.html

https://github.com/Xelio/Machine-Learning

Proof for Perceptron

$\text{Margin}(S, \bar{w}_{opt}, t_{opt}) = \gamma$.

Let $\bar{w}_{opt}, t_{opt}$ denote the max-margin hyperplane.

Assume $\| \bar{w}_{opt} \|_2 = 1$

Normalizing the weights

Let $(\bar{w}, t)$ denote **any** hyperplane, and $D \neq 0$. Then hyperplane $\left(\frac{\bar{w}}{D}, \frac{t}{D}\right)$ provides the same decision rule as $(\bar{w}, t)$.

$$\text{sign}\left(\frac{\bar{w}}{D} \cdot \bar{X}_i - \frac{t}{D}\right) = \text{sign}\left(\frac{\bar{w} \cdot \bar{X}_i - t}{D}\right) = \text{sign}(\bar{w} \cdot \bar{X}_i - t)$$

$\|\bar{w}_{opt}\| = 1$

Proof for Perceptron

Let $\bar{w}_{opt}, t_{opt}$: max-margin hyperplane. $\| \bar{w}_{opt} \| = 1$

Then, for all i

$$y_i\left(\bar{w}_{opt} \cdot \bar{X}_i - t_{opt}\right) \geq \gamma.$$

How many updates can happen? …

$$d\left(\bar{X}_i, (\bar{w}_{opt}, t_{opt})\right) = \frac{|\bar{w}_{opt} \cdot \bar{X}_i - t_{opt}|}{\|w_{opt}\|} = |\bar{w}_{opt} \cdot \bar{X}_i - t_{opt}| \geq \gamma$$

$$y_i = \text{sign}\left(\bar{w}_{opt} \cdot \bar{X}_i - t_{opt}\right)$$

Proof for Perceptron

Let $\overline{w}_{opt}, t_{opt})$ be the $d + 1$ dimensional vector.

Claim: $|\mathbf{t_{opt}}| \leq \mathbf{1}$.

PROOF: If not, then the distance of the origin from the hyperplane $\overline{w}_{opt}, t_{opt}$ is

$$\left|t_{opt}\right| \geq 1.$$

The hyperplane is at a distance at least 1 from the origin.

SHOW that this means all training examples are on the same side of $\overline{w}_{opt}, t_{opt})$ (hint: use $\| \bar{X}_i \| \leq 1$)

Proof for Perceptron

Therefore, $\| (\overrightarrow{\boldsymbol{w}_{opt}}, \boldsymbol{t}_{opt}) \|^{2} \leq \mathbf{2}$.

Proof for Perceptron

Let $(\overline{w}_j, t_j)$ be the vectors after the jth **update**. Let

Note that $(\overline{w}_0, t_0) = (\vec{0}, 0)$.

Claim: $(\overline{\boldsymbol{w}}_{\boldsymbol{j}}, \boldsymbol{t}_{\boldsymbol{j}}) \cdot (\overline{\boldsymbol{w}}_{\boldsymbol{opt}}, \boldsymbol{t}_{\boldsymbol{opt}}) \geq \boldsymbol{j\gamma}$.

PROOF:

Suppose you made a mistake on $(\bar{X}_i, y_i)$ with the weights $(\overline{w}_{j-1}, t_{j-1})$, before the jth update:

$$\overline{w}_j \leftarrow \overline{w}_{j-1} + y_i \cdot \bar{X}_i$$
$$t_j = t_{j-1} - y_i$$

Proof for Perceptron

$$\bar{w}_j \leftarrow \bar{w}_{j-1} + y_i \cdot \bar{X}_i$$
$$t_j = t_{j-1} - y_i$$

$(\bar{w}_j, t_j) \cdot (\bar{w}_{opt}, t_{opt})$
$= (\bar{w}_{j-1} + y_i \cdot \bar{X}_i, t_{j-1} - y_i) \cdot (\bar{w}_{opt}, t_{opt})$
$= (\bar{w}_{j-1}, t_{j-1}) \cdot (\bar{w}_{opt}, t_{opt}) + y_i(\bar{w}_{opt} \cdot \overrightarrow{x_i} - t_{opt})$ *(WHY?)*

$\geq (\bar{w}_{j-1}, t_{j-1}) \cdot (\bar{w}_{opt}, t_{opt}) + \gamma$ (why?)

… (induction goes here)

$\geq j \cdot \gamma$.

Proof for Perceptron

Claim: $\| (\bar{w}_j, t_j) \|^2 \leq 2j$.

PROOF:

$\| (\bar{w}_j, t_j) \|^2 = (\bar{w}_j, t_j) \cdot (\bar{w}_j, t_j)$(plugging in the update equations):
$= (\bar{w}_{j-1}, t_{j-1}) \cdot (\bar{w}_{j-1}, t_{j-1}) + \| \vec{x_i} \|^2 + 1 + 2y_i(\bar{w}_{j-1} \cdot \bar{X}_i - t_{j-1})$

Term in red is at most 0, (WHY?). Using $\| \bar{X}_i \|^2 \leq 1$,

$\| (\bar{w}_j, t_j) \|^2 \leq \| (\vec{w}_{j-1}, t_{j-1}) \|^2 + 2 \leq \cdots \leq 2j$.

Proof for Perceptron

Cosine of an angle is always at most 1:

$$\frac{(\bar{w}_j, t_j) \cdot (\bar{w}_{opt}, t_{opt})^2}{\| (\bar{w}_j, t_j) \|^2 \| (\bar{w}_{opt}, t_{opt}) \|^2} \leq \mathbf{1}$$

By the three claims, this implies

$$\frac{j^2 \gamma^2}{2j \cdot 2} \leq 1,$$

Which tells us that $j \leq 4/\gamma^2$.