SUPERVISED LEARNING

CLASSIFICATION

$\{x_1, \ldots, x_n\}$ $\quad x_i \in \mathbb{R}^d$

$y_1, \ldots, y_n$ $\quad y_i \in \{0, 1\}$ / $y_i \in \{+1, -1\}$

Goal: $\quad h : \mathbb{R}^d \rightarrow \{0, 1\}$

$\mathbb{1}(z) = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{o/w} \end{cases}$

ERROR: $\quad \sum_{i=1}^{n} \mathbb{1}\left(h(x_i) \neq y_i\right)$

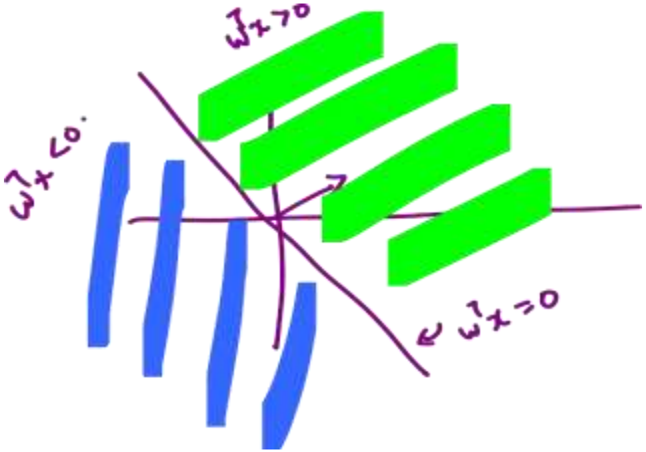
$$\mathcal{H}_{linear} = \left\{ h_w \; : \; h_w(x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ 0 & \text{o/w} \end{cases} \right.$$

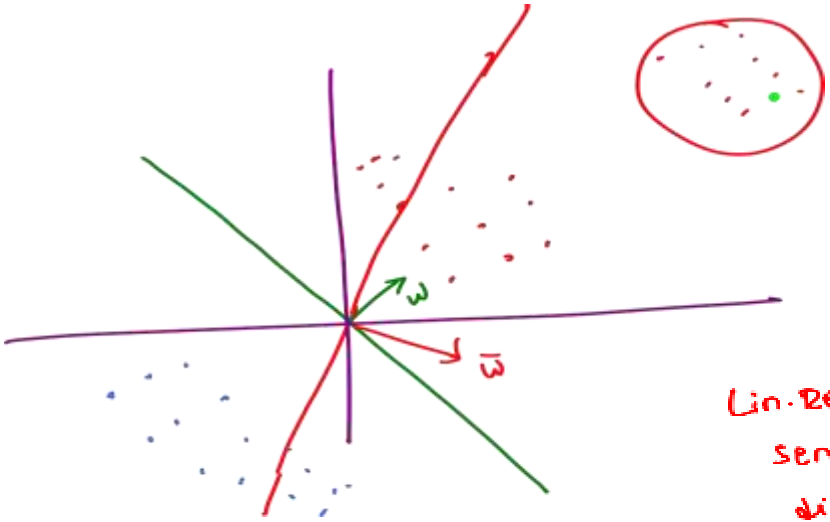
$$\min_{h \in \mathcal{H}_{linear}} \sum_{i=1}^{n} \mathbb{1}(h(x_i) \neq y_i)$$

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

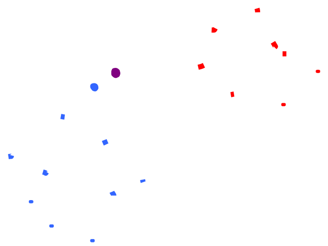
$$= \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \mathbb{1}\left(\text{sign}(w^T x_i) \neq y_i \right)$$

$\hookrightarrow$ NP HARD Problem





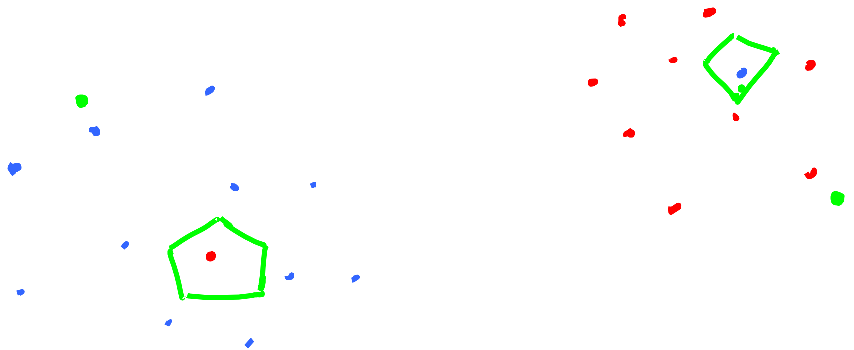
Lin. Reg is sensitive to distances.



SIMPLEST POSSIBLE ALGORITHM

- Given x_{test}, find x^* - the closest point to x_{test} in the training set

Predict $\hat{y}_{test} = y^*$.



Issue: outliers.

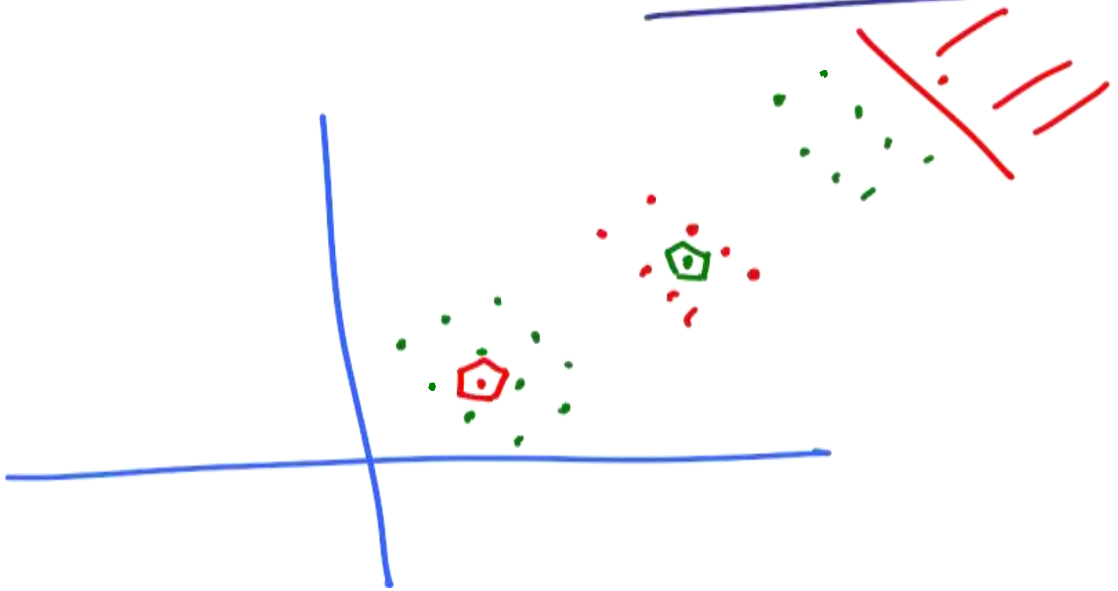
Fix: check more neighbours

K - Nearest Neighbour

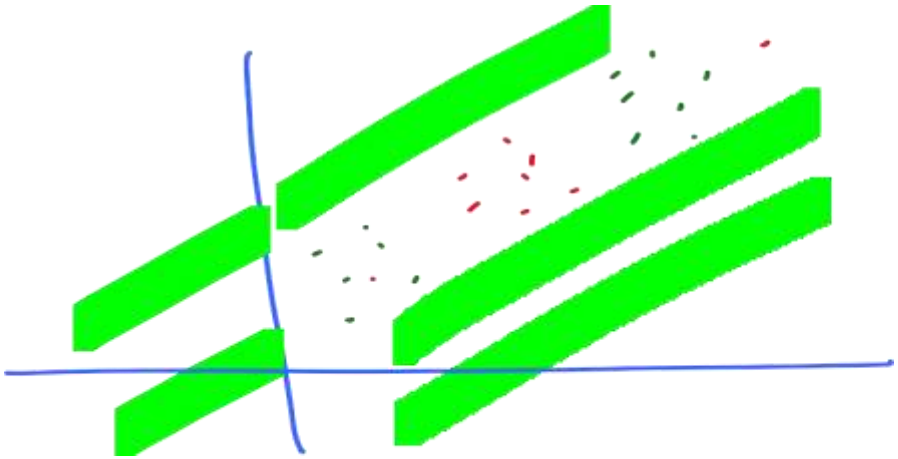
→ Given x_{test}, look at K-nearest neighbours $x_1^*, x_2^*, \ldots, x_k^*$

→ $y_{test} = \text{majority}(y_1^*, \ldots, y_k^*)$

DECISION BOUNDARY



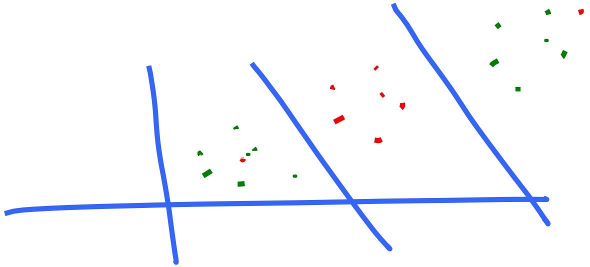
K=1



$k = n$

Choosing k

Cross validate.



k^*

ISSUES with K-NN

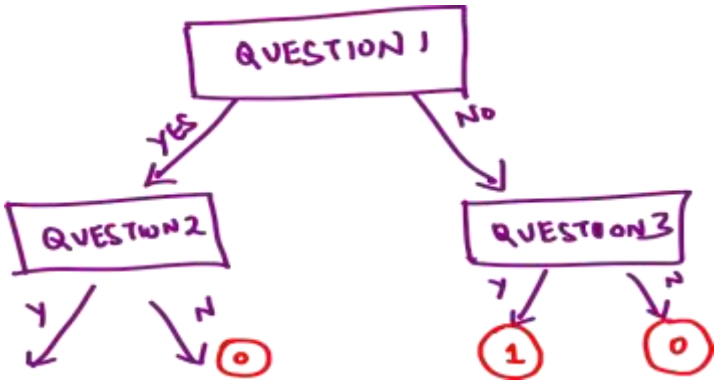
- PREDICTION IS COMPUTATIONALLY EXPENSIVE.
- NO MODEL is learnt. Cannot throw away data after "learning"

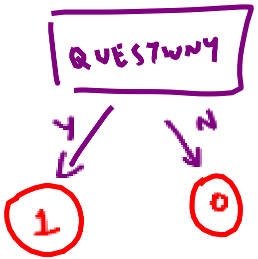
DECISION TREES

INPUT : Dataset $\{ (x_1, y_1) \ . \ . \ . \ , (x_n, y_n) \}$

OUTPUT : DECISION TREE

Decision tree





PREDICTION: Given x_{test}, traverse through the tree to reach a leaf node. Predict y_{test} = answer in leaf node

QUESTION:

A question is a (feature, value) pair.

Eg: height ≤ 180 cm ?

- How to measure "goodness" of a Question?

$D \quad \{(x_1, y_1) \; \cdots \; (x_n, y_n)\}$

Y

N

$D_{yes} \quad \{(x_5, y_5), (x_8, y_8) \cdots\}$

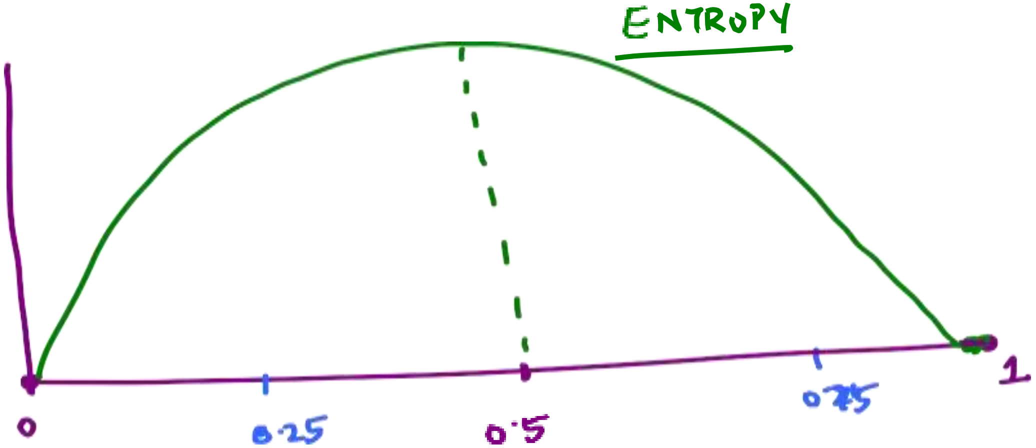
$\{(x_1, y_1), (x_2, y_2), \; \cdots\} \quad D_{No}$

$\{\underbrace{y_1, y_2, \ldots, y_n}\}$ $y_i \in \{0, 1\}$.

Measure of Impurity.

$\text{entropy}(p) = -\left(p \log p + (1-p) \log(1-p)\right)$

$[\log(0) = 0]$



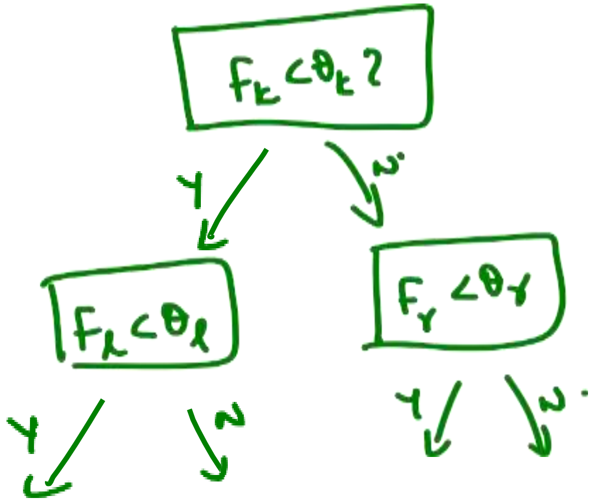
$p \to$ Fraction of 1's.

$$\text{Information gain}(\text{feature}, \text{value}) = \text{ENTRODY}(D) - \left(\frac{\gamma\ \text{ENTROPY}(D_{yes}) + (1-\gamma)\ \text{ENTROPY}(D_{no})}{2} \right)$$

$$\gamma = \frac{|D_{yes}|}{|D|} \qquad (1-\gamma) = \frac{|D_{no}|}{|D|}$$

ALGORITHM

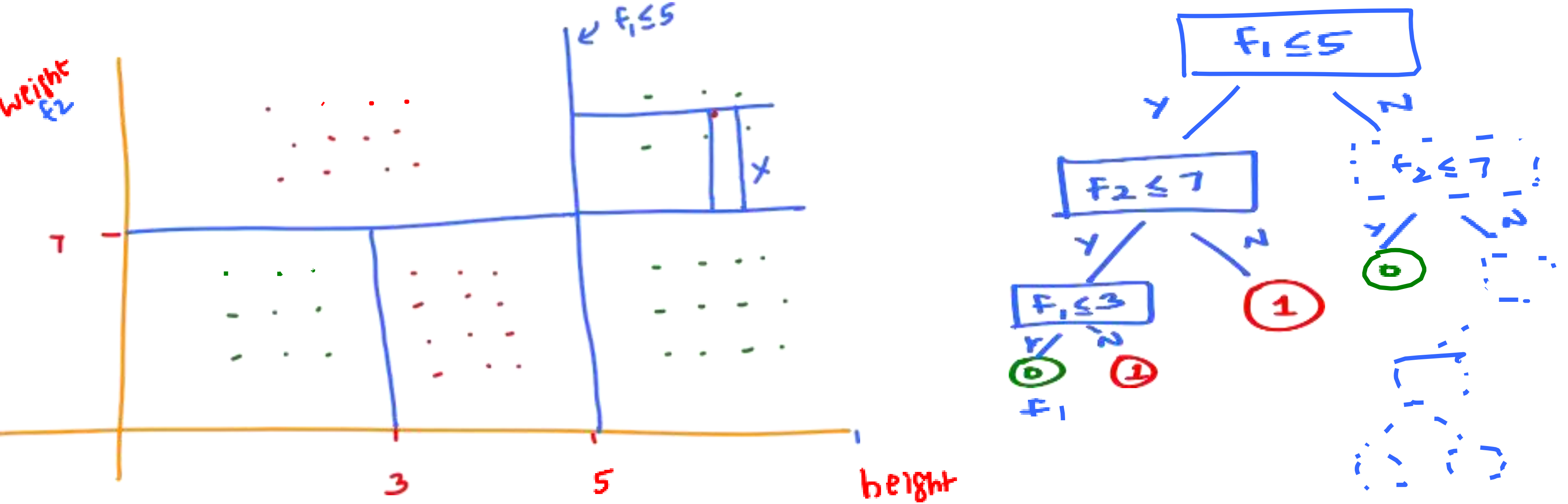
- → Discretize each feature in [min, max] range.
- → Pick question ($f_k \le \theta$) that has highest information gain
- → Repeat for D_{yes} & D_{no}.



Points

- Depth is a hyper-parameter.
- Can also stop growing if node is "sufficiently" smooth.

Decision Boundary



Note Title 28-Sep-22

Formal Treatment of the Learning Problem

Instance space

Label space.

Assume $D \sim \mathcal{X} \times \mathcal{Y}$

$\mathcal{X} \to \mathbb{R}^d$

$\mathcal{Y} \to \underline{\{0,1\}}$

Both training and test data are drawn from the same distribution!

If one is given access to D, what is the best classifier?

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim D}[\mathbb{I}(h(x) \neq y))]$$

In words, h* is the classifier that minimizes the average **test** error

Formal Treatment of the Learning Problem

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim D}[\mathbb{I}(h(x) \neq y))]$$

Expectation of a indicator function is probability. So,

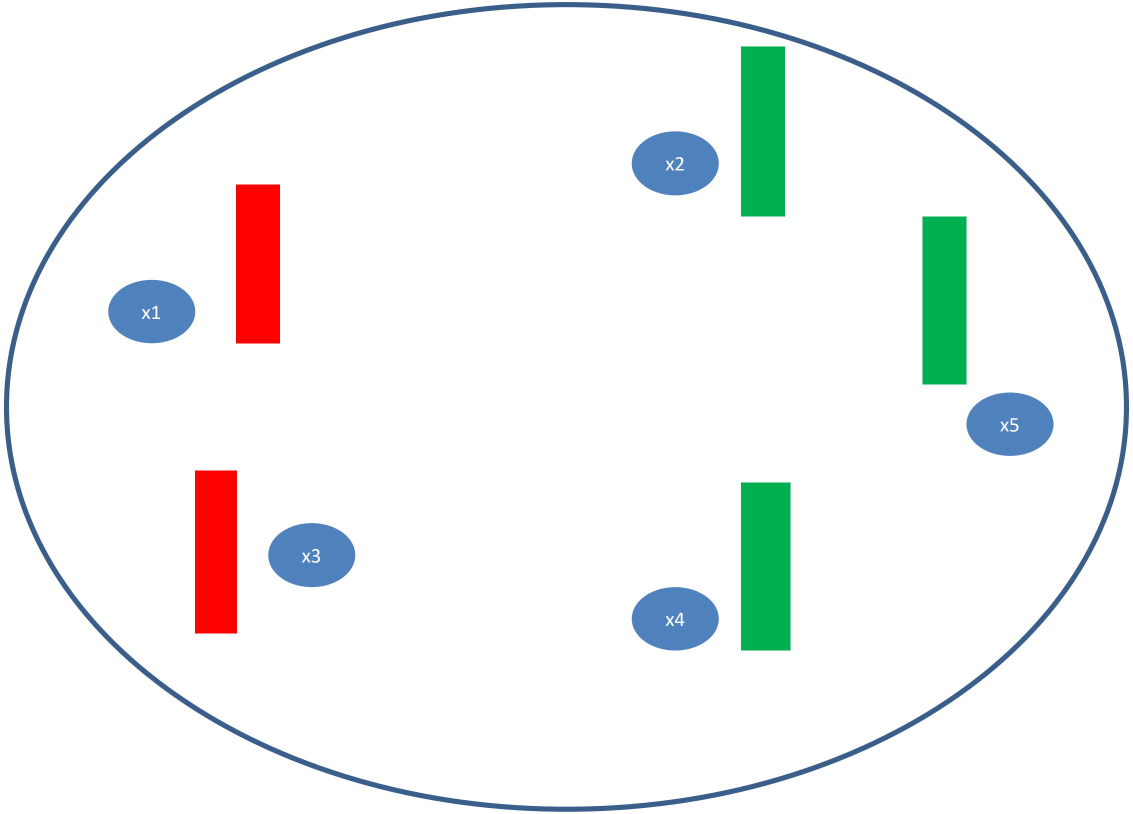
$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{P}_{(x,y)\sim D}[h(x) \neq y)]$$

P(x) is Uniform over {x1,x2,x3,x4,x5}

Data points

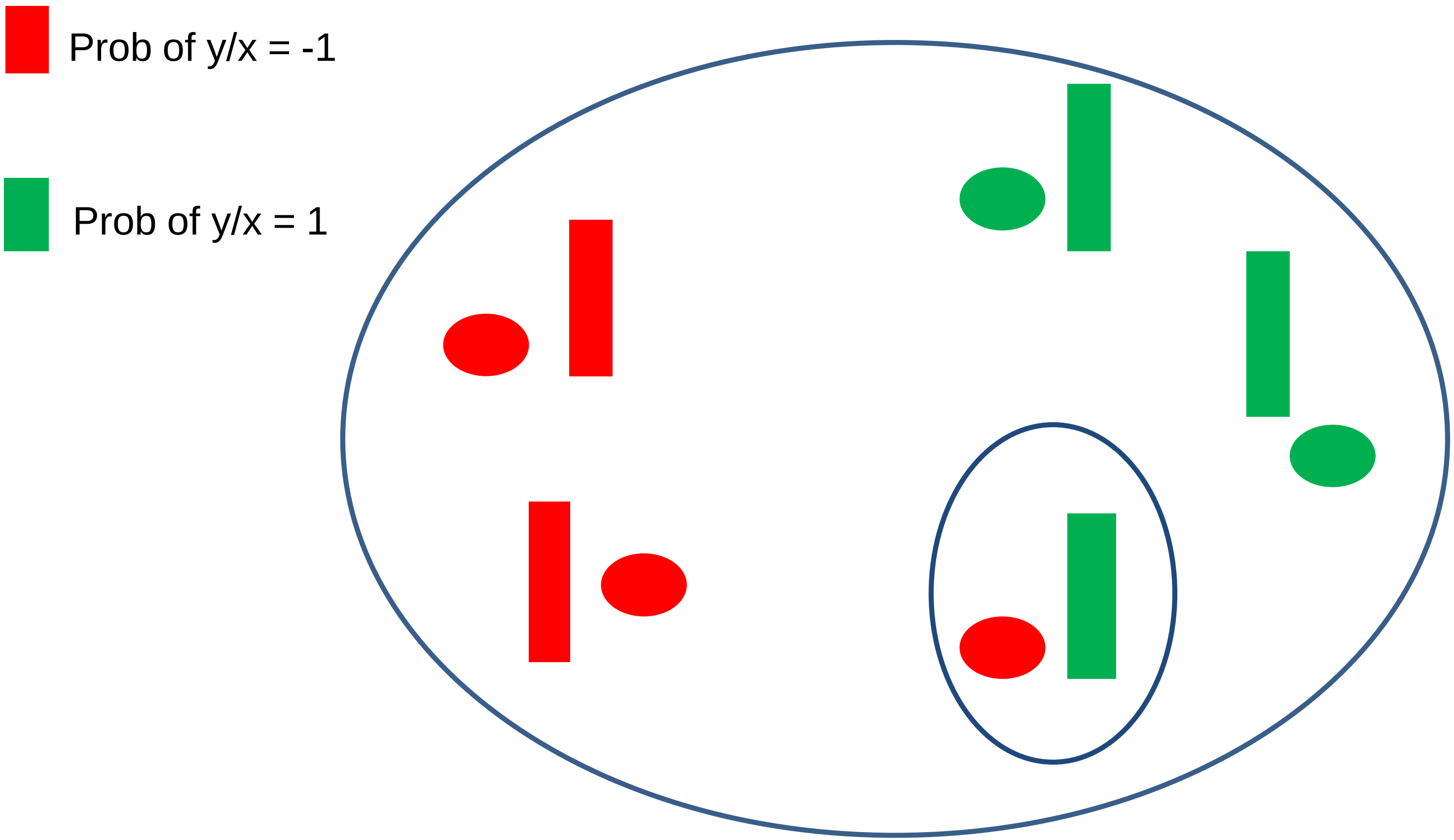
Prob of y/x = -1

Prob of y/x = 1

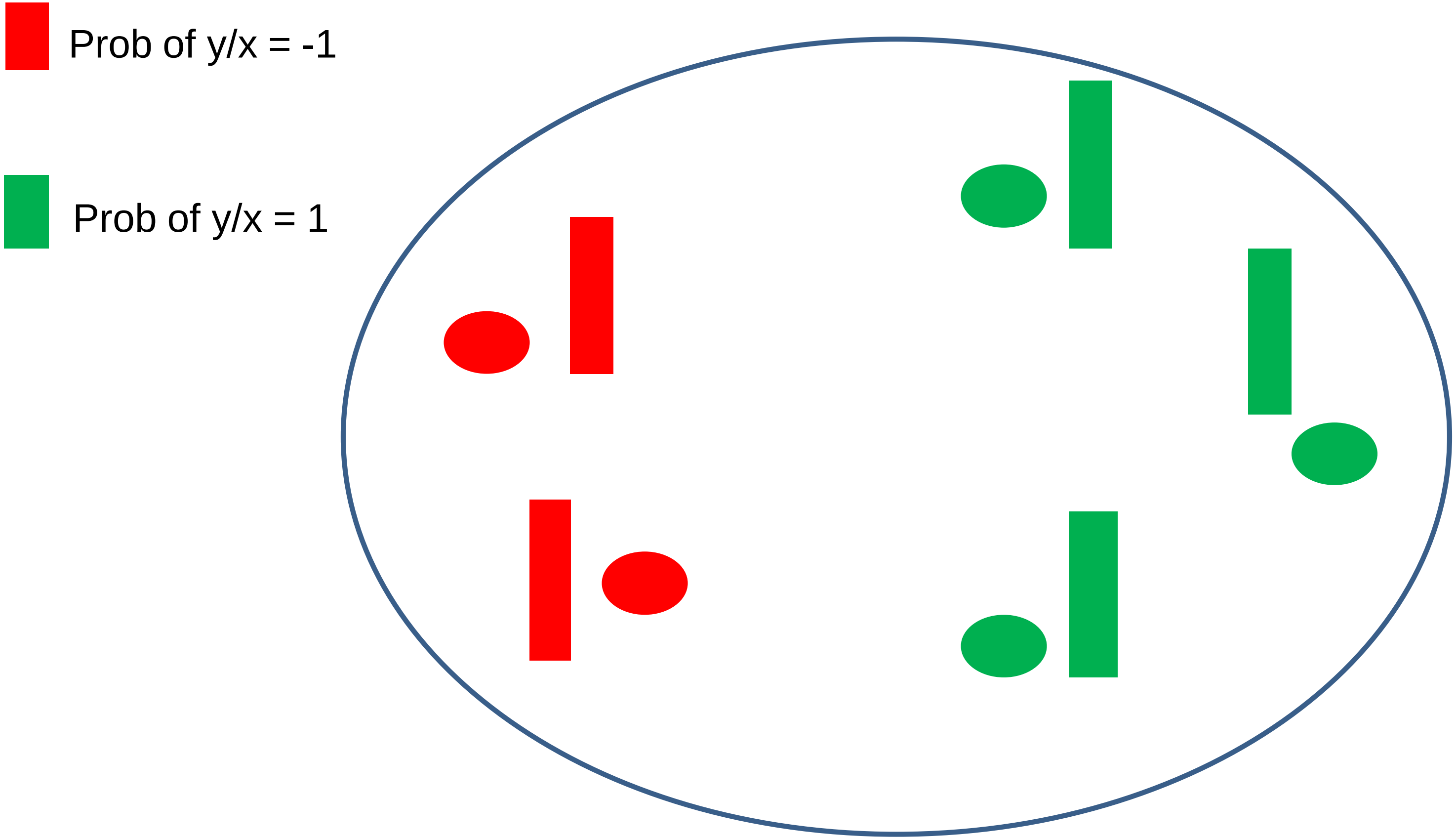


A typical dataset from this distribution

x5	1
x3	-1
x2	1
x4	1
x3	-1
x1	1



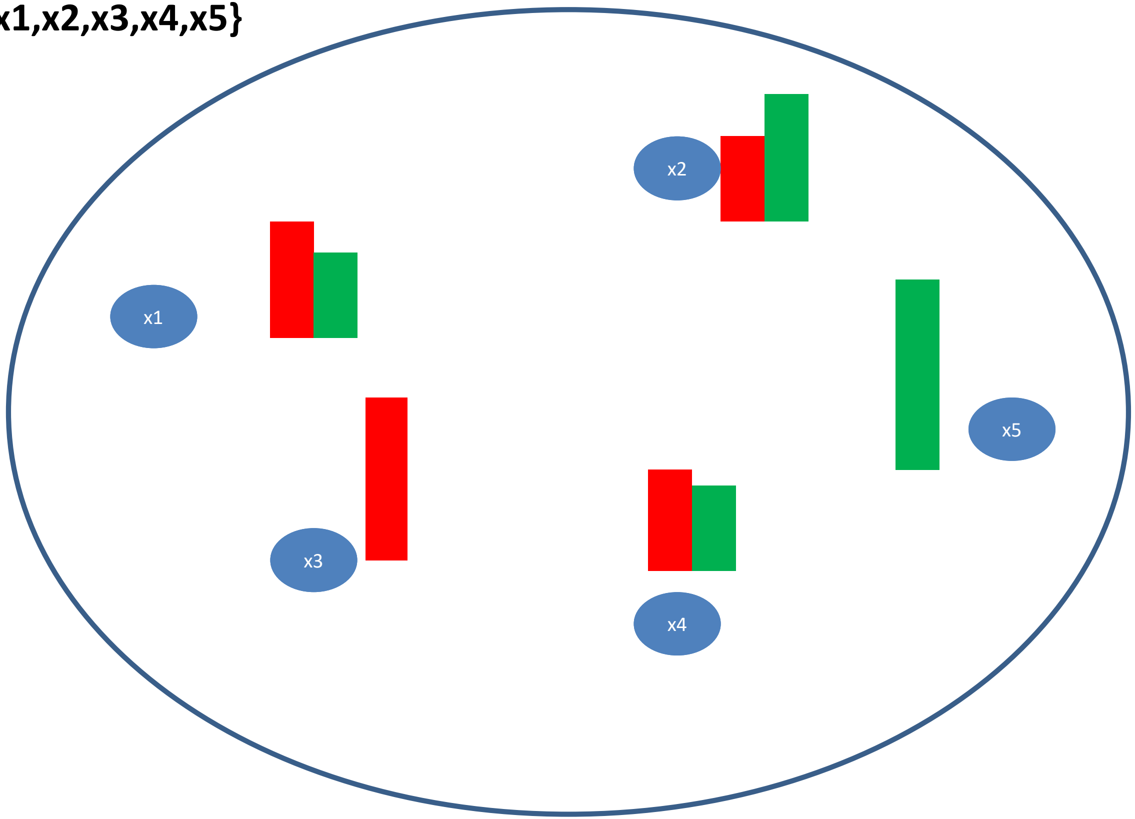
A sub-optimal classifier



Optimal classifier – Gives 0 test error!!

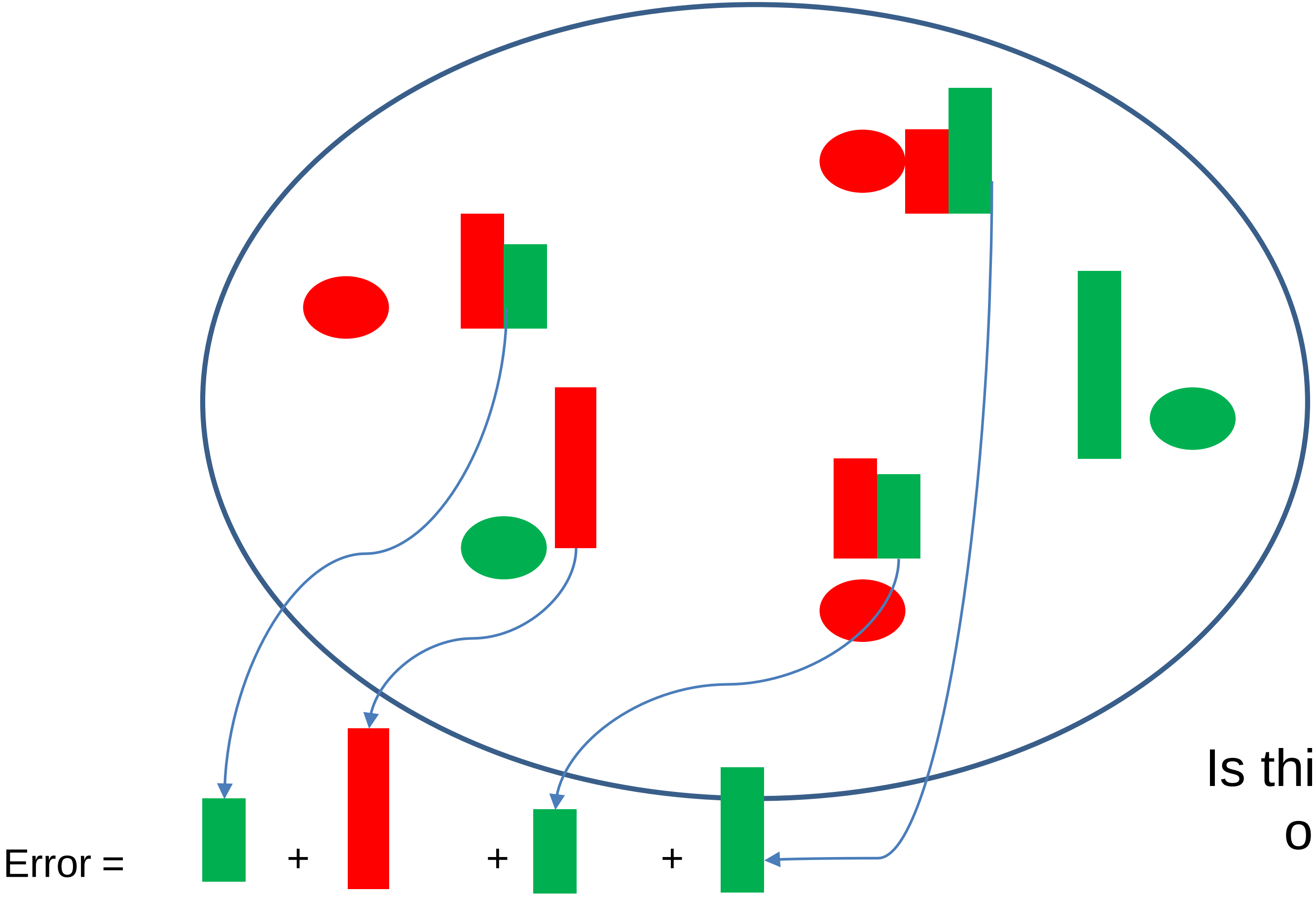
What if there is no classifier can make zero error?

P(x) is Uniform over {x1,x2,x3,x4,x5}

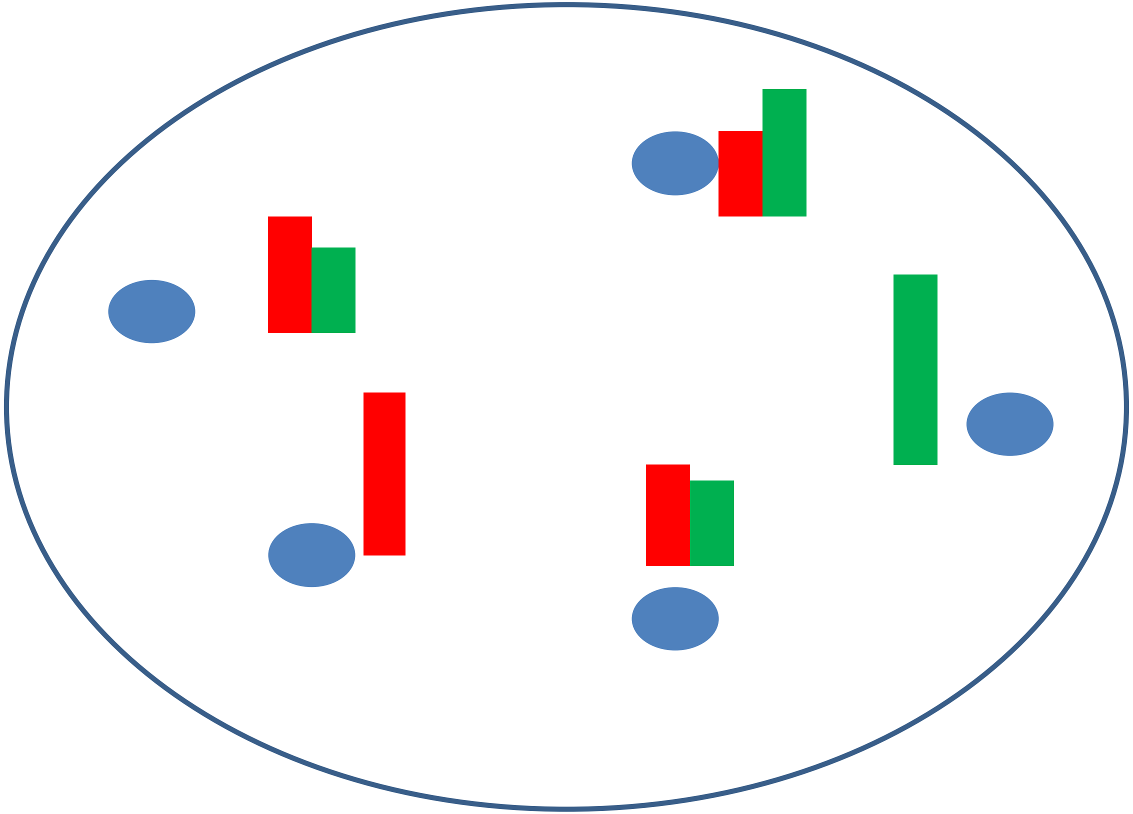


A typical dataset from this distribution

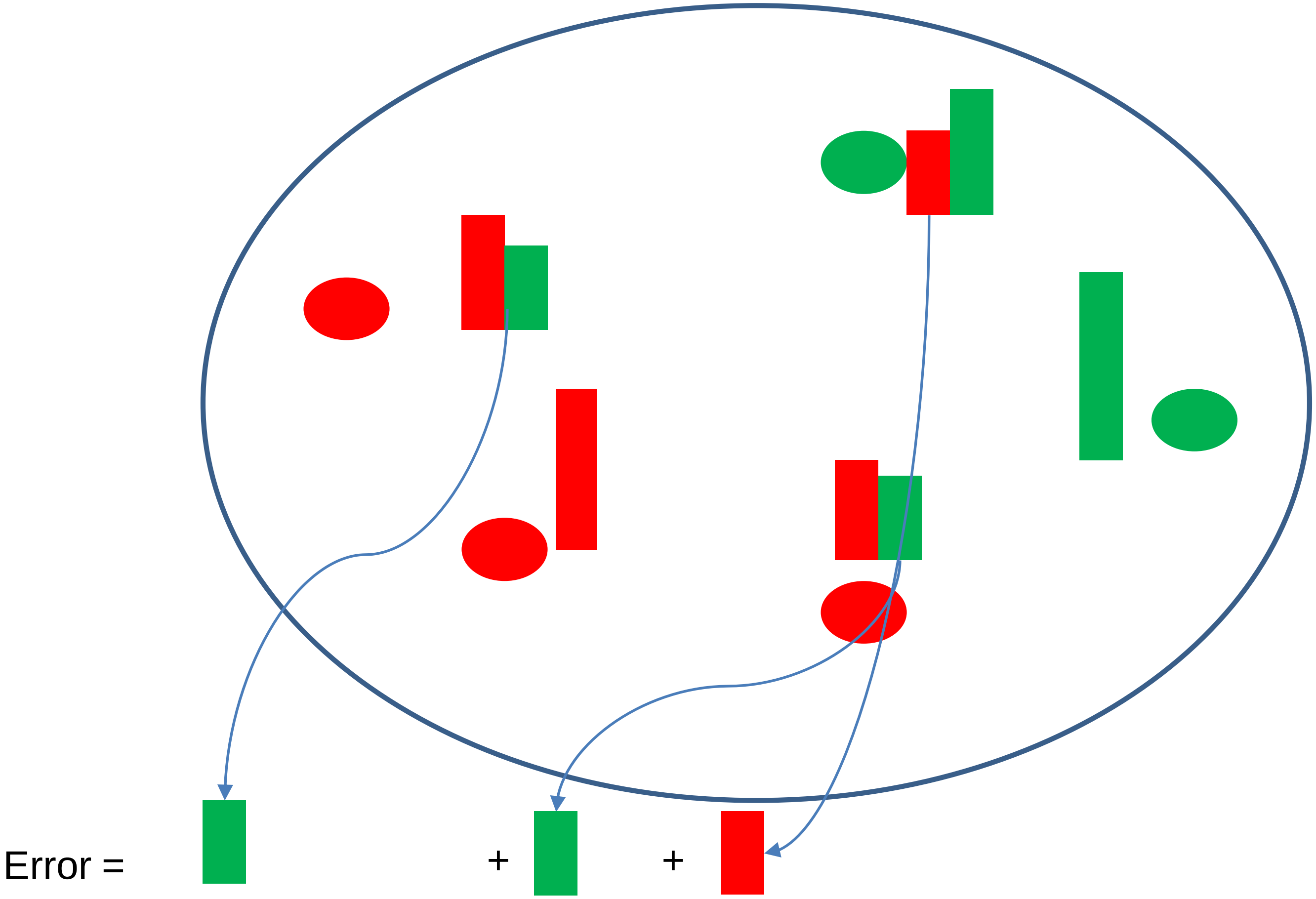
x5	1
x3	-1
x2	1
x4	1
x2	-1
x4	-1



Is this classifier optimal?



How should the “best” classifier predict?


Error =
+
+

Formal Treatment of the Learning Problem

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim D}[\mathbb{I}(h(x) \neq y))]$$

Expectation of a indicator function is probability. So,

$$h^* = \arg\min_{h \in \mathcal{H}} \mathbb{P}_{(x,y)\sim D}[h(x) \neq y)]$$

BAYES OPTIMAL CLASSIFIER

$$h^*(x) = \{1 \;\; if \;\; P(y/x) \geq 0.5 \;\; and \;\; -1 \;\; otherwise\}$$

GOOD NEWS

We know the form of the best classifier

BAD NEWS

We don't know the distribution D over X,Y

We will make assumptions about distribution generating data

TYPES OF MODELING

- GENERATIVE MODEL
- DISCRIMINATIVE MODEL

Gen. Model

$P(x,y)$

Discriminative model

$P(y|x)$

Eg: K-NN

: Decision-tree

$y|_x \sim N(w^T x, \sigma^2)$

Note that in both models, we only need P(y/x) to make predictions

Generative Models

$$P(x,y) = \underline{\boxed{P(x)} \cdot \boxed{P(y/x)}} = \boxed{P(y)} \boxed{P(x/y)}$$

Data: $\{ (x_1, y_1) \ \ldots \ (x_n, y_n) \}$

$$x_i \in \{0,1\}^d \qquad y_i \in \{0,1\}$$

words in dictionary

Eg: Spam-Classification

"Hello, how are you?"

$$\begin{array}{cccccccccccccc} & \text{ARE} & & & & \text{HELLO} & & \text{HOW} & & & \text{YOU} & \\ [0 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 1 & \cdots & 1 &] \end{array}$$

GENERATIVE STORY

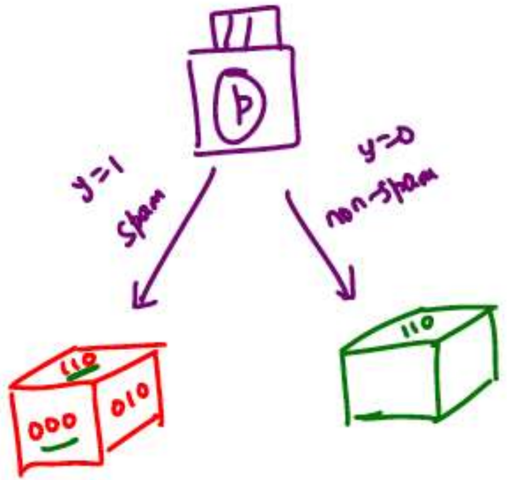
- **STEP 1:** Decide the label by tossing a coin

$$P(y_i = 1) = p$$

- **STEP 2:** Decide features given label using $P(x_i / y_i)$

$d=3$

Lottery	Money	Hello
1	1	0

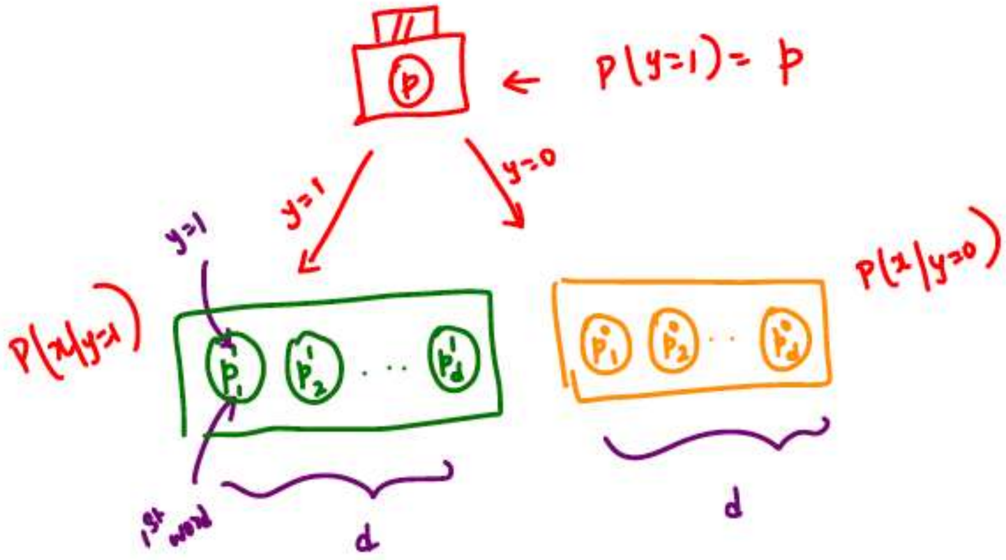


How many parameters? $\quad 1 + (2^d - 1) + (2^d - 1)$

$$= 2^{d+1} - 1$$

Issue: . Too many parameters

. Need Alternate Story.



parameters?

$1 + d + d$

$= \underline{2d+1}.$

Step 1: $P(y=1) = p$

$f_i \in \{0,1\}$

$$p_1^y$$

Step 2: $P\left(x = [f_1 \; f_2 \; \cdots \; f_d] / y\right)$

$$= \prod_{k=1}^{d} \left(p_k^y\right)^{f_k} \left(1 - p_k^y\right)^{(1-f_k)}$$

ASSUMPTION: Features are "CONDITIONALLY INDEPENDENT" given label

Parameters to estimate :

$$p, \{p_1^1, \ldots, p_d^1\}, \{p_1^0, \ldots, p_d^0\}$$

Maximum Likelihood estimators

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i \leftarrow$$

[Fraction of Spam emails in data]

label

$$\hat{p}_j^y = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j^i = 1, y_i = y)}{\sum_{i=1}^{n} \mathbb{1}(y_i = y)}$$

j^{th} word

[Fraction of y-labeled emails that contain j^{th} word]

Given $x_{test} \in \{0,1\}^d$, [test email]

What is y_{test}?

Predict 1 if $P(y_{test} = 1 / x_{test}) > P(y_{test} = 0 / x_{test})$

Predict 0 otherwise.

BAYES RULE

$$P(y^{test} = 1 / x^{test}) = \frac{P(x^{test} / y^{test} = 1) \cdot \underline{P(y^{test} = 1)}}{P(x^{test})}$$

Say $x^{test} = [f_1 \quad f_2 \quad \cdots \quad f_d] \in \{0,1\}^d.$

$$① \quad P(y^{test}=1 \mid x^{test}) \propto \left[\prod_{k=1}^{d} (\hat{p}_k^1)^{f_k} (1-\hat{p}_k^1)^{(1-f_k)}\right] \cdot \hat{p}$$

$$② \quad P(y^{test}=0 \mid x^{test}) \propto \left[\prod_{k=1}^{d} (\hat{p}_k^0)^{f_k} (1-\hat{p}_k^0)^{(1-f_k)}\right] \cdot (1-\hat{p})$$

If ① > ②, predict $y^{test}=1$

predict $y^{test}=0$ otherwise.

MODEL uses 2 key things

- CLASS CONDITIONAL INDEPENDENCE
- $\boxed{\text{BAYES}}$ THEOREM

- $\boxed{\text{Naive}}$ assumption.
- may not hold.
- works well in practice.

NAIVE-BAYES algorithm

Pitfall to watch out for

→ If a word does not appear in the train set, but appears in the test set, then both $\hat{p}_j^1 = 0$ and $\hat{p}_j^0 = 0$, this can't be predicted.

Possible fix

→ Add 2 pseudo emails to data. ↙ LAPLACE SMOOTHING.

$x = [1\ 1\ 1\ \cdots\ 1],\ y = 1$

$x = [1\ 1\ 1\ 1\ 1\ \cdots\ 1],\ y = 0$

DECISION BOUNDARY ?

Predict $y_{test} = 1$ if $\dfrac{P(y_{test}=1 / x_{test})}{P(y_{test}=0 / x_{test})} \geqslant 1.$

$$\log\left(\frac{P(y_t=1/x_t)}{P(y_t=0|x_t)}\right) \geqslant 0$$

$$\log\left(\frac{P(x_t|y_t=1)\cdot P(y_t=1)}{P(x_t|y_t=0)\cdot P(y_t=0)}\right) \geqslant 0$$

$$\Rightarrow \quad \log\left(\prod_{\underline{i=1}}^{d} \left(\frac{(\hat{p}_i^1)^{f_i} (1-\hat{p}_i^1)^{(1-f_i)}}{(\hat{p}_i^0)^{f_i} (1-\hat{p}_i^0)^{(1-f_i)}} \right) \cdot \left(\frac{\hat{p}}{1-\hat{p}} \right) \right) \geq 0$$

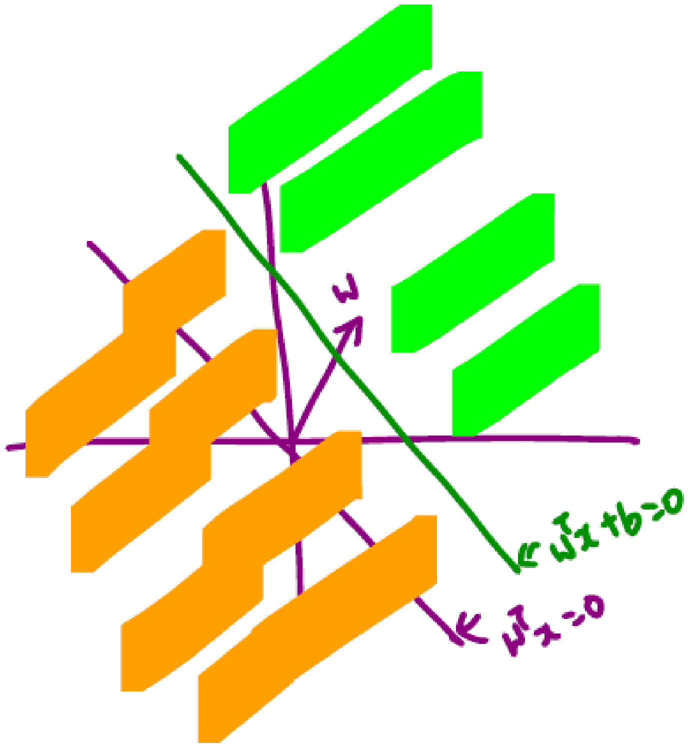
$$\sum_{i=1}^{d} \left[\underline{f_i} \log \underbrace{\left(\underline{\hat{p}_i^1} / \hat{p}_i^0 \right)} + (1-\underline{f_i}) \log \underbrace{\left(\frac{1-\hat{p}_i^1}{1-\hat{p}_i^0} \right)} \right] + \log \underbrace{\left(\frac{\hat{p}}{1-\hat{p}} \right)} \geq 0$$

$$x_{test} = [f_1 \quad f_2 \dots f_d]$$

Predict 1 if

$$\sum_{i=1}^{d} f_i \underbrace{\log\left(\frac{\hat{p}_i^1 (1-\hat{p}_i^0)}{\hat{p}_i^0 (1-\hat{p}_i^1)}\right)}_{w_i} + \underbrace{\sum_{i=1}^{d} \log\left(\frac{1-\hat{p}_i^1}{1-\hat{p}_i^0}\right) + \log\left(\frac{\hat{p}}{1-\hat{p}}\right)}_{b} \geq 0$$

$x_{test} = [f_1 \quad f_2 \dots \quad f_d]$



Data : $\{(x_1, y_1), \dots, (x_n, y_n)\}$

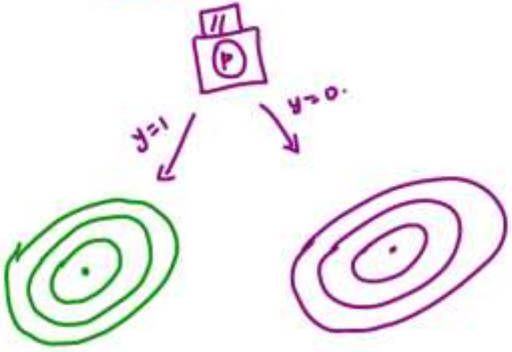
$x_i \in \boxed{\mathbb{R}^d}$

$y_i \in \{0, 1\}$

A Generative Story

Note: In this model, Co-variances are same.

$x \mid y=1 \sim N(\mu_1, \Sigma)$



$x \mid y=0 \sim N(\mu_0, \Sigma)$

Parameters:

$$\underset{p}{1} + \underset{\mu_1}{d} + \underset{\mu_0}{d} + o(d^2)$$

Maximum-Likelihood estimator

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(Fraction of points labelled 1)

$$k \in \{0,1\} \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = k) \cdot x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = k)}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^T$$ [argue this]

Prediction

As usual, by Bayes rule.

Predict 1 if $P(y_t=1 \mid x_t) > P(y_t=0 \mid x_t)$

$$P(x_t \mid y_t=1) \cdot P(y_t=1) > P(x_t \mid y_t=0) \cdot P(y_t=0)$$

$$\underline{P(x_t | y_t = 1)} \cdot P(y_t = 1) > P(x_t | y_t = 0) \cdot P(y_t = 0)$$

$$f(x_t ; \hat{\mu}_1, \hat{\Sigma}) \cdot \hat{p} > f(x_t ; \hat{\mu}_0, \hat{\Sigma}) \cdot (1 - \hat{p})$$

$$\Rightarrow \quad e^{-\frac{1}{2}(x_t - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (x_t - \hat{\mu}_1)} \cdot \hat{p} > e^{-\frac{1}{2}(x_t - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (x_t - \hat{\mu}_0)} \cdot (1 - \hat{p})$$

On Simplification

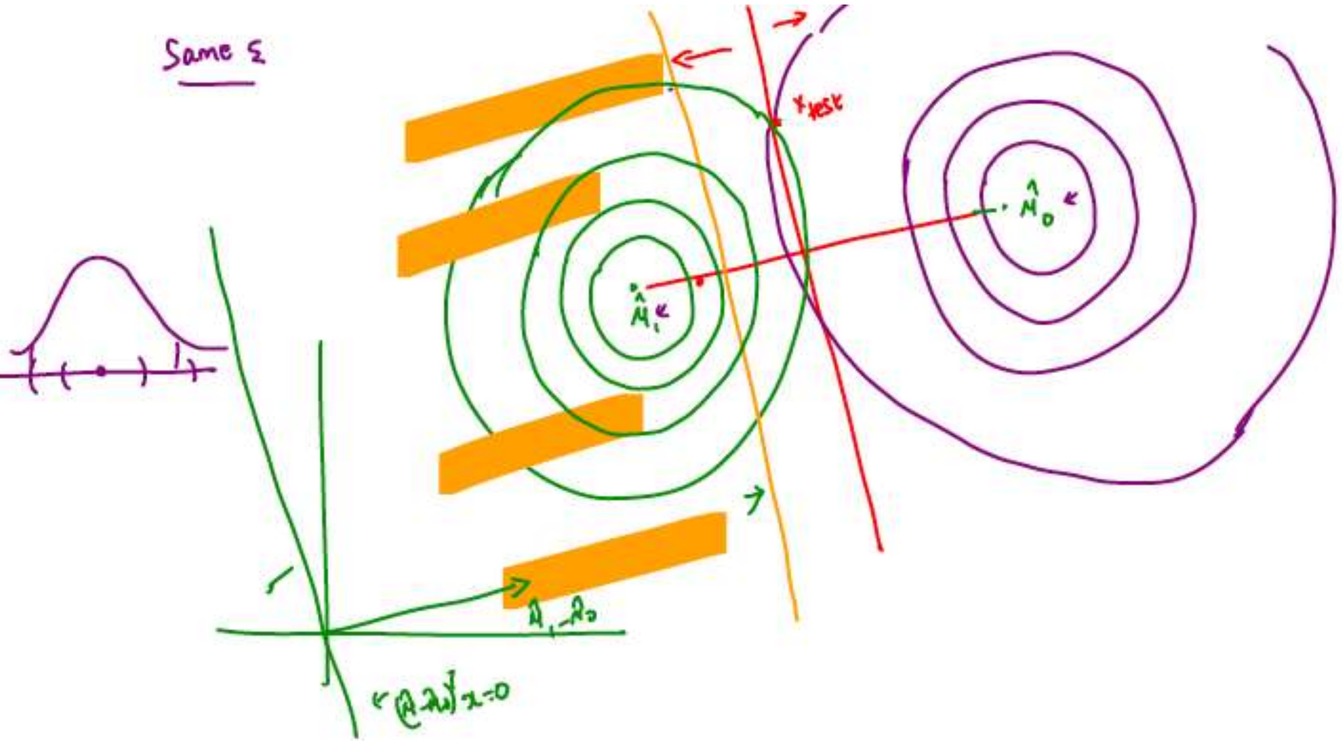
$y_{test} =$ Predict 1

$$\Rightarrow \quad \underbrace{2(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}}_{w} x_{test} + \underbrace{\hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0 - \hat{\mu}_1 \hat{\Sigma}^{-1} \hat{\mu}_1 + \log\left(\frac{1-\hat{p}}{\hat{p}}\right)}_{b} \geq 0$$

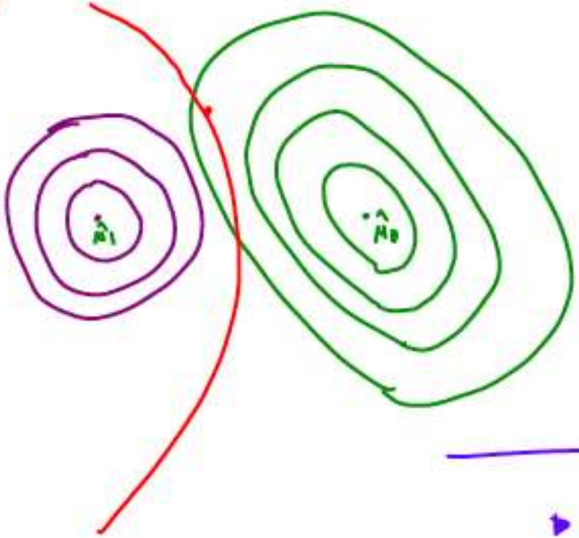
$$w^T x_{test} + b \geq 0 \quad \Rightarrow$$ DECISION FUNCTION IS LINEAR

only if Σ is same for $x|y=1$, $x|y=0$

Same Σ



Different Σ



Decision function is Quadratic.

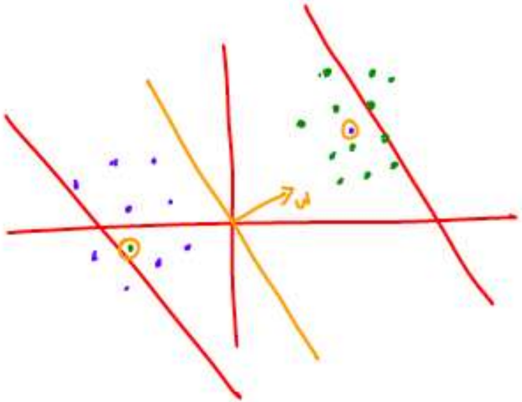
- Gaussian Naive Bayes.

Note Title 28-Sep-22

Question :

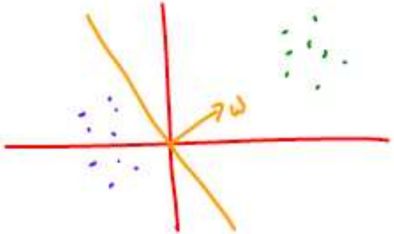
→ Can we directly make linear assumption about $P(y/x)$

→ $$\boxed{\begin{aligned} P\left(y=1/x\right) &= 1 \quad \text{if } \underline{w}^T x + b \geq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}}$$

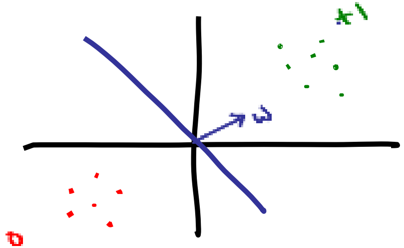


Dataset is
not allowed in
our model

LINEAR
SEPARABILITY
ASSUMPTION



Allwed.



$$P(y=1|x) = 1 \quad \text{if} \quad w^T x \geq 0$$
$$= 0 \quad \text{o/w}$$

Linear separability assumption.

Lin. sep assumption :

$$\exists w \in \mathbb{R}^d \quad \text{s.t} \quad \text{sign}(w^T x_i) = y_i \quad \forall i \in [n]$$

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

$\{1, \ldots, n\}$.

DATA $\{(x_1,y_1)\ \ldots\ (x_n,y_n)\}$

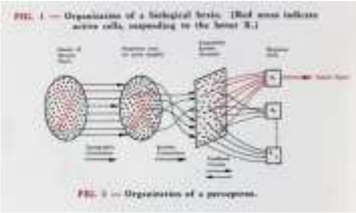
$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \mathbb{1}\left(\operatorname{sign}(w^T x_i) \neq y_i\right) \rightarrow \text{NP-HARD}$$

We know in general the problem of finding best w is NP-hard.

But is it still hard under linear separability assumption?

PERCEPTRON

Frank Rosenblatt ’50, Ph.D. ’56, works on the “perceptron” –
what he described as the first machine “capable of having an original idea.”



https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

PERCEPTRON - ALGORITHM

Input: $\{ (x_1, y_1) \cdots , \}$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$

Iteration

$w^0 = [0 \; 0 \cdots 0]$ $\quad 0 \in \mathbb{R}^d.$

until convergence

→ Pick (x_i, y_i) from dataset

IF $\left(\text{sign}(w^{t^T} x_i) = y_i \right)$

do nothing

else

$\{ w^{t+1} = w^t + x_i y_i$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$

UPDATE RULE.

end.

end

UPDATE RULE

$$w^{t+1} = w^t + x_i y_i$$

Two types of mistake

Type 1

Pred $\to$ 1

act $\to$ -1

Type 2

Pred $\to$ -1

act $\to$ 1

Type -1

Pred $\to$ 1

act $\to$ -1

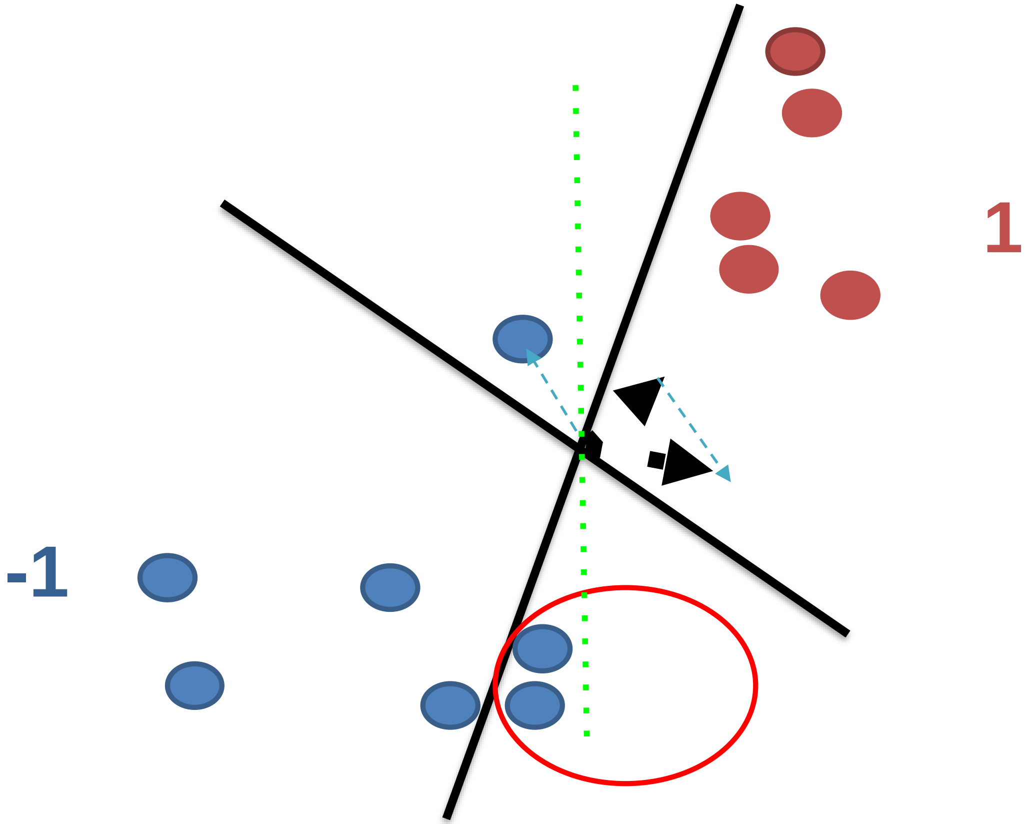
$w^{t^T} x_i \geq 0$ but $y_i = -1$

$$w^{t+1^T} x_i = (w^t + x_i y_i)^T x_i$$

$$= \underset{\geq 0}{w^{t^T} x_i} + \underset{< 0}{y_i \|x_i\|^2}$$

≤ 0 $\qquad$ > 0

Type -2



Fixing one error might lead to more errors elsewhere

In general, does perceptron work for linearly separable data?

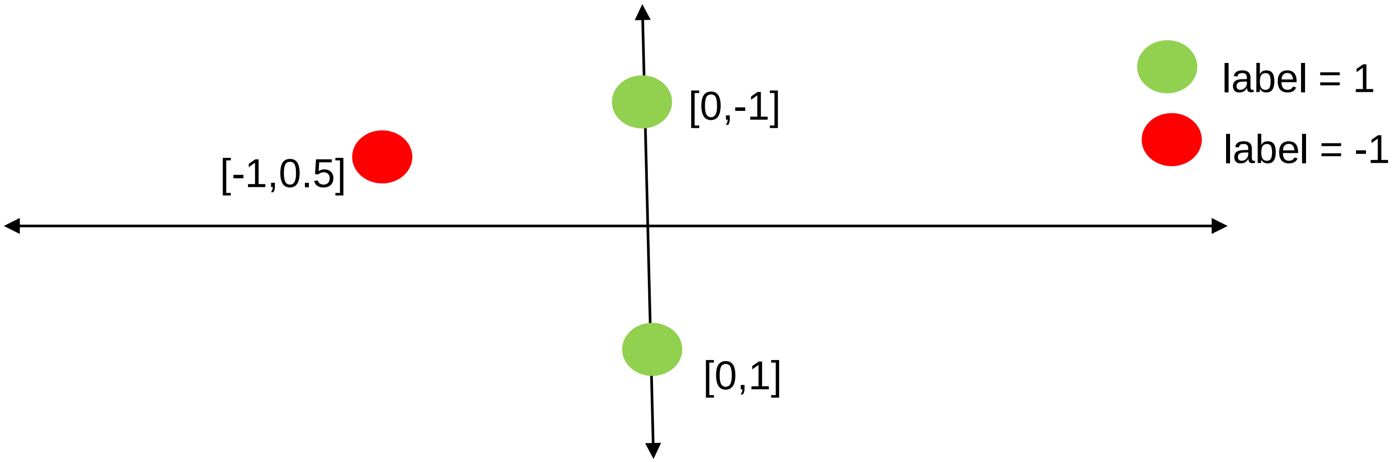
Recall

Lin. sep assumption :

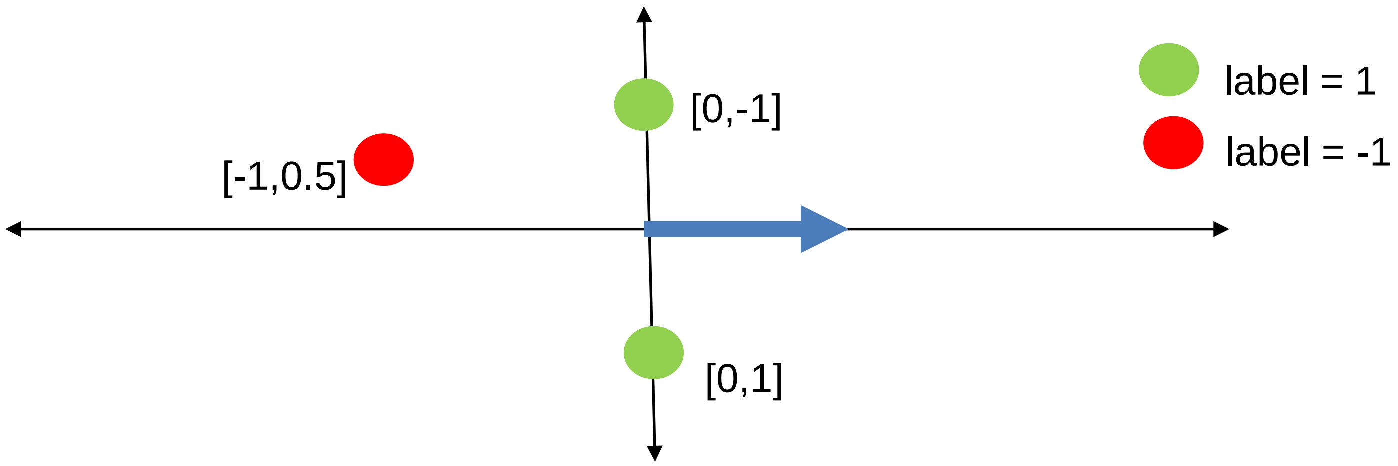
$\exists\, w \in \mathbb{R}^d$ s.t $\text{sign}(w^T x_i) = y_i \quad \forall i \in [n]$

$[n] = \{1, \dots, n\}$.

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

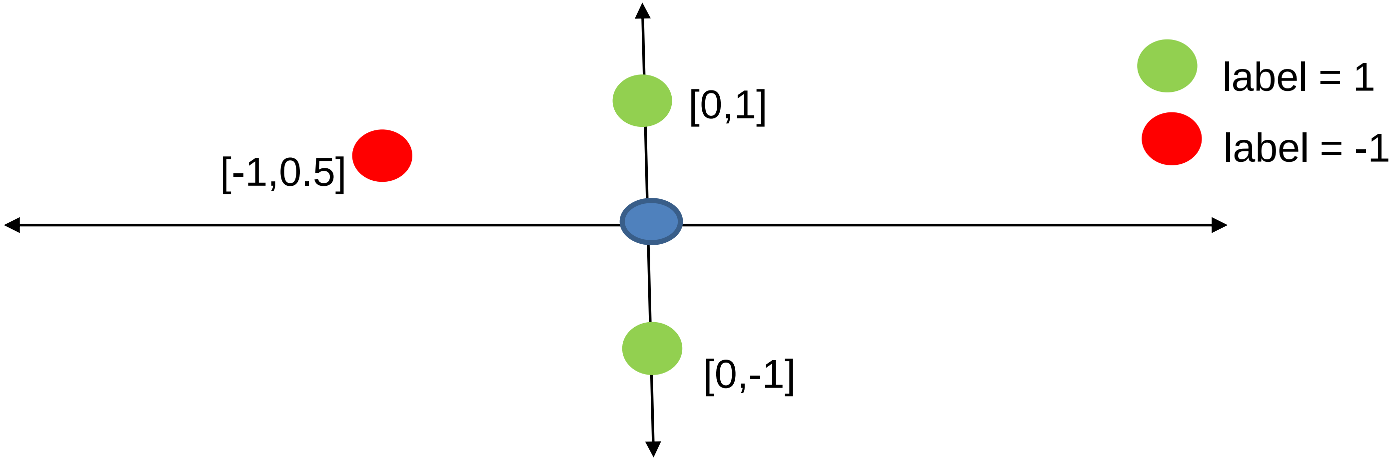


Is this a linearly separable dataset?



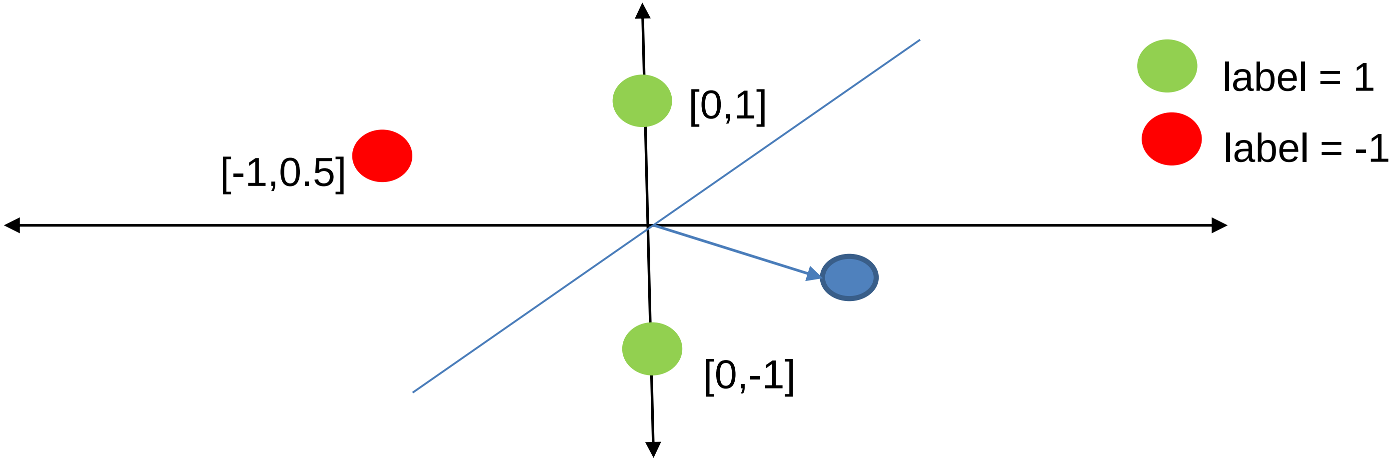
Any w in the positive x-axis linearly separates the data.
The dataset is linearly separable.

Let’s see what perceptron learns from this data!



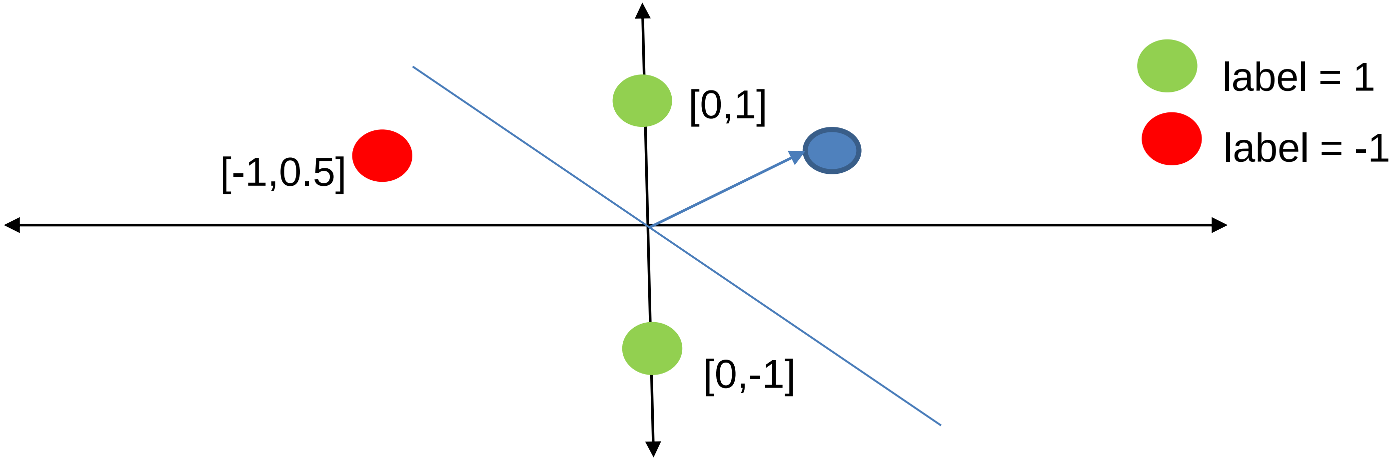
$$w^0 = [0\ 0]$$

	Predicted label	True label
[0 1]	1	1
[0 -1]	1	1
[-1 0.5]	1	-1



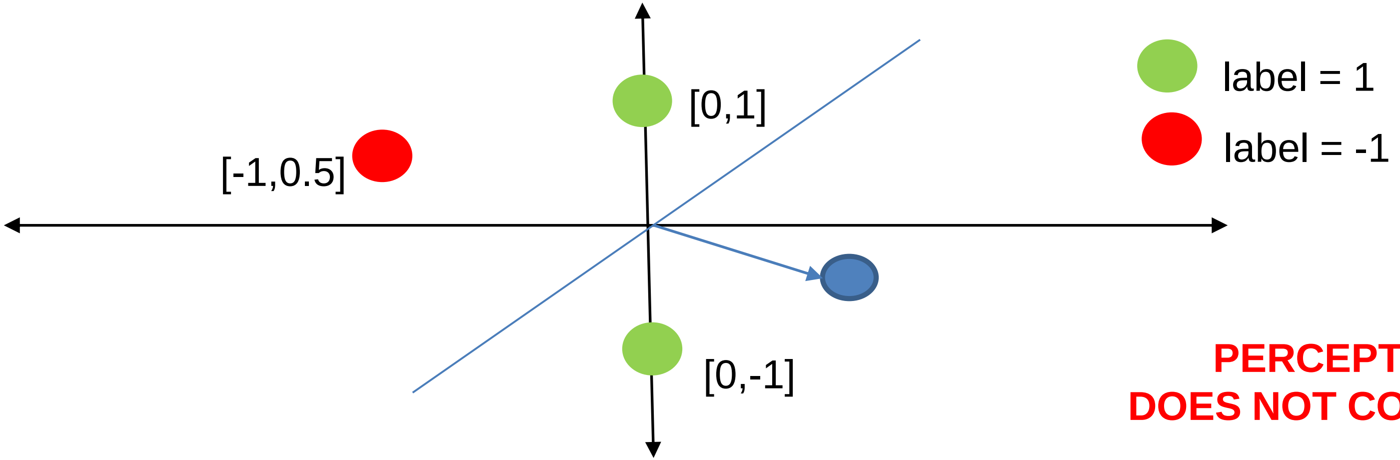
$w^1 = [1\ -0.5]$

	Predicted label	True label
[0 1]	-1	1
[0 -1]	1	1
[-1 0.5]	-1	-1



	Predicted label	True label
[0 1]	1	1
[0 -1]	-1	1
[-1 0.5]	-1	-1

$w^2 = [1\ 0.5]$



PERCEPTRON DOES NOT CONVERGE

w^3= [1 -0.5]

	Predicted label	True label
[0 1]	-1	1
[0 -1]	1	1
[-1 0.5]	-1	-1

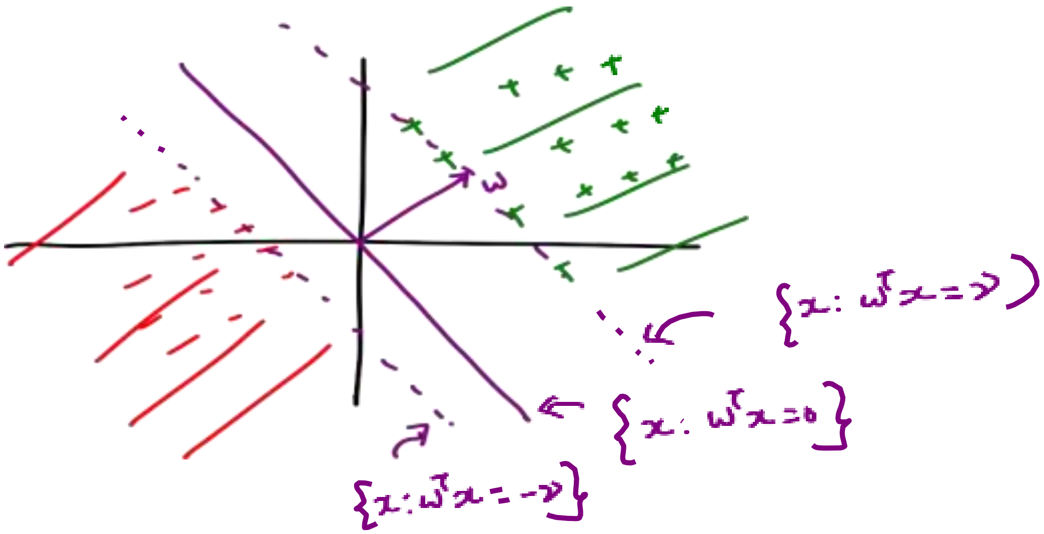
Issue $\quad\quad\quad\quad\quad\quad\quad\quad$ $c > 0$

- optimal $w^* = \begin{bmatrix} c \\ 0 \end{bmatrix}$ has datapoints that lie on the linear separator.

- If we assume this isn't the case, will perceptron converge?

ASSUMPTIONS

- LINEAR SEPERABILITY with γ- MARGIN



A Dataset $D = \{ (x_1, y_1) \ldots, (x_n, y_n)$ is $\underline{L.S}$

with γ-margin if $\exists\, w^*$ s.t

$$\underline{(w^{*T} x_i)\, y_i \geq \gamma} \quad \forall i \quad \text{for some } \underline{\gamma > 0}$$

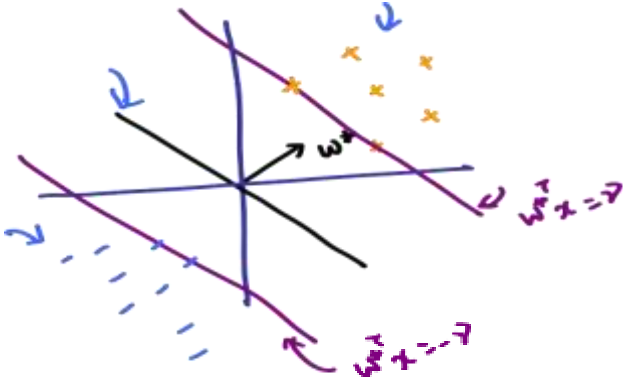
PERCEPTRON

$$\boxed{w^{t+1} = w^{t} + x_i y_i}$$

ASSUMPTIONS

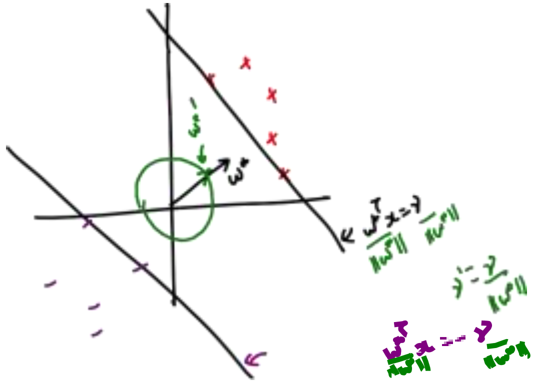
1. Linear separability with γ margin

A dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is L.S with γ margin if $\exists\, w^*$ s.t $(w^{*T} x_i) y_i \geq \gamma$ $\forall i$ for some $\gamma > 0$



(2) RADIUS ASSUMPTION

$\forall i \in D, \; \|x_i\|_2 \leq R$ for some $R > 0$



(3) Without loss of generality

$\|w^*\| = 1$

ANALYSIS of "mistakes" of Perceptron.

- observe that an update happens only when a "mistake" occurs.
- Say the current guess $= w_\ell$ and a mistake happens w.r.t (x, y)

$$w_{\ell+1} = w_\ell + x \cdot y$$

$$\|w_{\ell+1}\|^2 = \|w_\ell + x \cdot y\|^2$$

$$= (w_\ell + x \cdot y)^T (w_\ell + x \cdot y)$$

$$\|w_{\ell+1}\|^2 = \underbrace{\|w_\ell\|^2} + \underbrace{2(w_\ell^T x) \cdot y}_{\leq 0} + \underbrace{(x^T x)}_{\|x\|^2} \cdot \underbrace{y^2}_{1}$$

≤ 0 because [mistake.]

$\leq R^2$

$$\leq \|w_\ell\|^2 + R^2 \leq \left(\underbrace{\|w_{\ell-1}\|^2 + R^2}\right) + R^2$$

$$\|w_0\|^2 + \ell R^2$$

Inductively

$$\boxed{\|w_{\ell+1}\|^2 \leq \ell \cdot R^2} \quad \text{——— } (1)$$

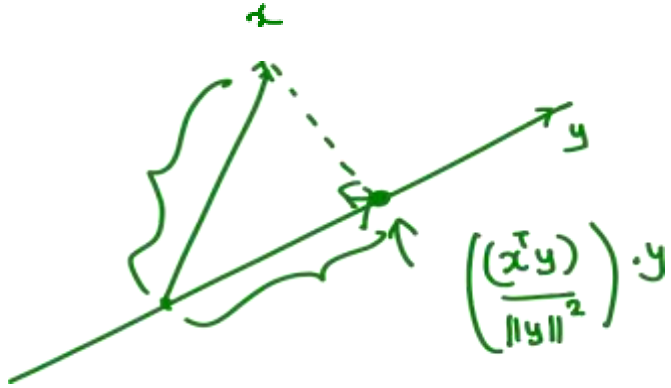
$$w_{\ell+1}^T w^* = (w_\ell + x \cdot y)^T w^*$$

$$= w_\ell^T w^* + \underbrace{(w^{*T} x) \cdot y}_{\geq \gamma}$$

$$\geq \left(w_{\ell-1}^T w^* + \underline{\gamma} \right) + \gamma$$

$\vdots$

$$\boxed{\underbrace{\underline{w_{\ell+1}^T w^*}}_{} \geq \ell \cdot \gamma}$$



$$\|x\|^2 \geq \left\| \left(\frac{x^T y}{\|y\|^2}\right) y \right\|^2$$

$$\geq \frac{(x^T y)^2}{\|y\|^{\cancel{4}\,2}} \cdot \cancel{\|y\|^2}$$

$$\Rightarrow \underline{(x^T y)^2 \leq \|x\|^2 \|y\|^2}$$

(C.S) Cauchy Schwartz inequality $\longrightarrow$ $x^T y \leq \|x\|\|y\|$

From before

$$\underline{w_{l+1}^T} w^* \geq l \cdot \gamma$$

$$\underbrace{\|w_{l+1}\|^2 \|w^*\|^2}_{} \underset{\text{C-S}}{\geq} (\underline{w_{l+1}^T} w^*)^2 \geq l^2 \gamma^2$$

$$\|w_{l+1}\|^2 \geq l^2 \gamma^2 \quad \text{— } 2$$

Combining 1 & 2.

$$\ell^2 \gamma^2 \leq \|w_{\ell+1}\|^2 \leq \ell R^2$$

$\hookrightarrow$ 2 $\qquad\qquad$ $\hookrightarrow$ 1

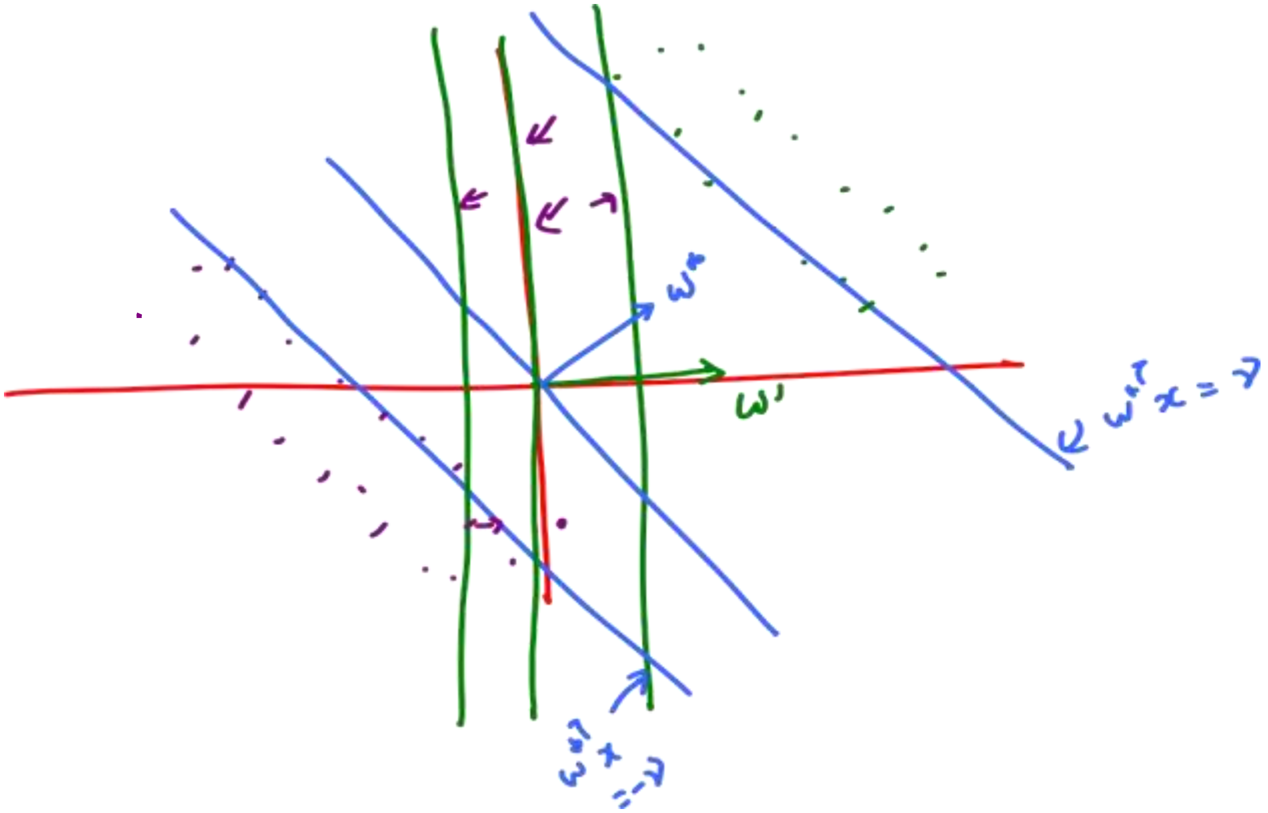
$$\Rightarrow \quad \ell^2 \gamma^2 \leq \ell R^2$$

$$\boxed{\ell \leq \frac{R^2}{\gamma^2}}$$

RADIUS-MARGIN BOUND.

$\Rightarrow$ # mistakes of Perceptron is bounded

$\Rightarrow$ Perceptron converges!



Perceptron's # mistakes depends on w^*.

But it might output w'.