Week03 - Part 01

Before we start, Some announcements:

1 - TA OH are added to Course page
There's at least one OH per day.

2 - WS1 posted. Solve the questions before the tutorial session. TAs will take your questions and review WS1.

3 - A1 posted: Pocket Alg. & linear regression
I created a Colab notebook so that you can easily do your work there.
Please download the latest version of notebook,
Visit "Assignments/A1" to get the link.

4. A1 is to be done in groups of two. you can join a different group for each assignments

5. Are there any hidden test cases for A1?
   - No. the marking scheme is provided in the handout.

Week03 - Part01

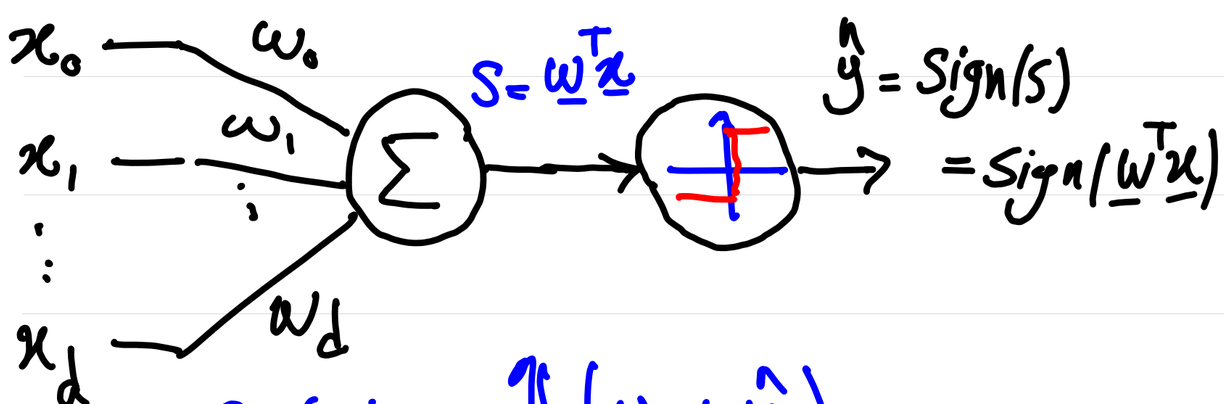
So far: Linear classification & linear Regression

- Given : $\mathcal{D} = \{(\underline{x}_n, y_n)\}_{n=1}^{N}$
- Unknown Target Function : $y = f(\underline{x})$
- Hypothesis Set : $\hat{y} = h(\underline{x})$ where $h \in \mathcal{H}$

Linear classification

$\underline{x} \in \mathbb{R}^{d+1}$, $y \in \{-1, +1\}$, $\hat{y} = \text{Sign}(\underline{w}^T \underline{x})$

- Illustrating as Neuron

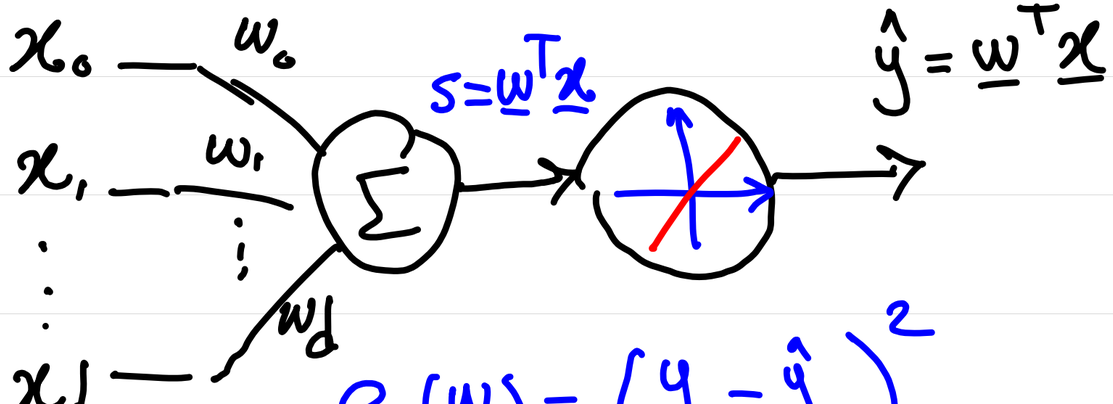


$e_n(\underline{w}) = \mathbb{I}(y_n \neq \hat{y}_n)$

Linear Regression

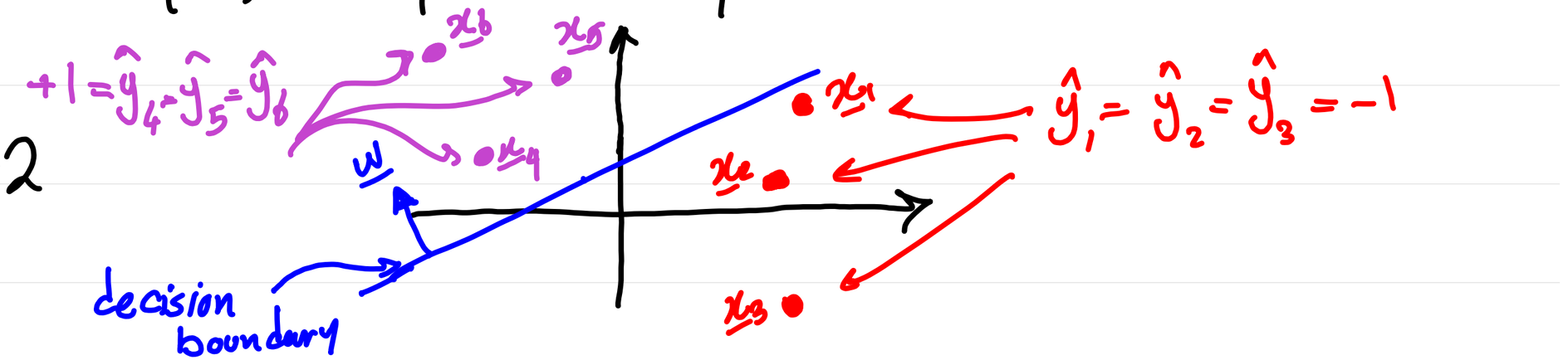
$\underline{x} \in \mathbb{R}^{d+1}$, $y \in \mathbb{R}$, $\hat{y} = \underline{w}^T \underline{x}$

- Illustrating as Neuron



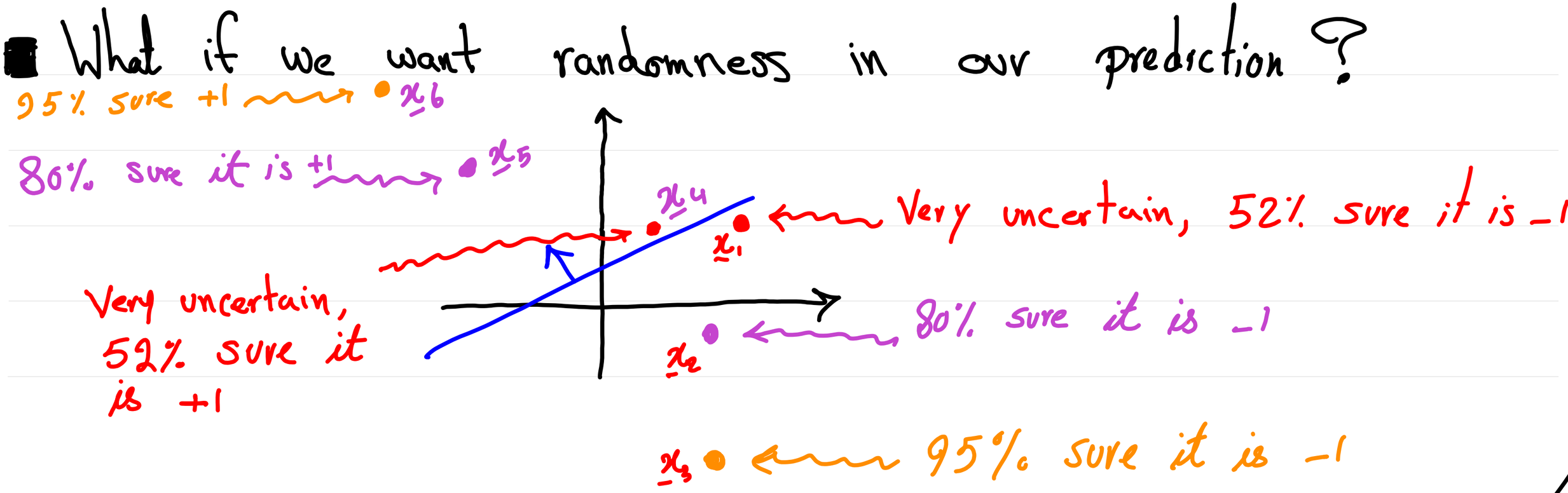
$e_n(\underline{w}) = (y_n - \hat{y}_n)^2$

- So far, We have studied deterministic hypothesis.
  - Given an input, they tell you what is its label



- E.g.: d=2

- What if we want randomness in our prediction?



- Why would Someone need a predictor with randomness?
  - E.g.: Suppose we want to predict the occurrence of heart attacks based on diet.
    - $\underline{x}$ = (average sugar in the diet, average fat in the diet)
    - $\underline{y}$ = heart attack
    - Given $\mathcal{D} = \{(\underline{x}_n, y_n)\}_{n=1}^{N}$,
    - For a new input $\underline{x}$, predict ...
      - How likely it is for the new person to have a heart attack, i.e., $\hat{P}_{\underline{N}}(+1 \mid \underline{x})$

$\hat{P}_{\underline{N}}(+1 \mid \underline{x})$:
- the likelihood of having a heart attack given that the person has $\underline{x}$ as its feature vector.
- the event of having a heart attack
- conditioned on the observed input feature vector

Lecture Outline

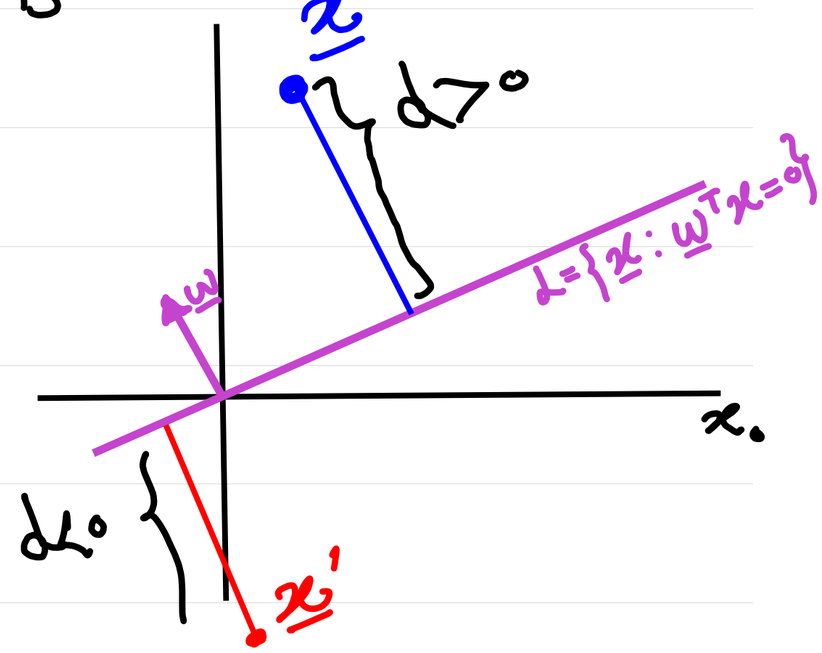
we want to predict with randomness $\hat{P}_{\underline{w}}(y=+1 \mid \underline{x})$

what hypothesis Set $\mathcal{H}$ should we use

Bingo! $\mathcal{H}$ is

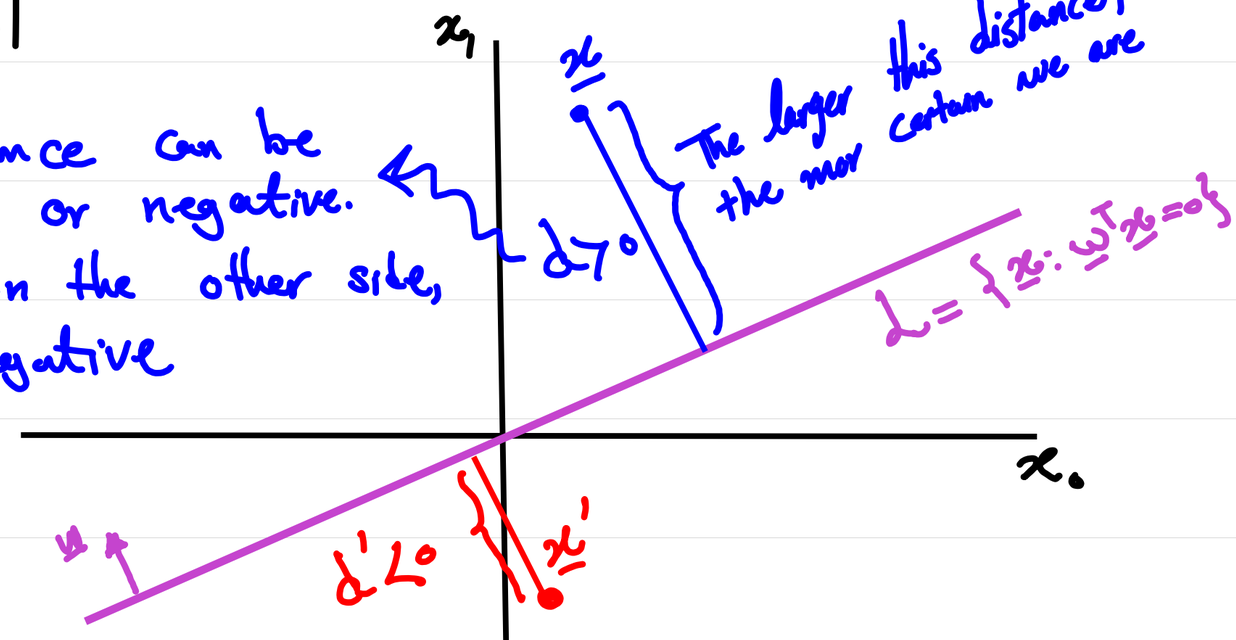
Today: Logistic Regression

- Predicts the probability of label: $\hat{P}_{\underline{w}}(+1|\underline{x})$
- Let's see what could be a reasonable formulation for Logistic regression model to output probability of label, $\hat{P}_{\underline{w}}(+1|\underline{x})$
  - We are looking for a decision rule that its prediction depends on
    - which side of the hyperplan the datapoint is x_1
    - How far away the datapoint is from the hyperplane.
      - The larger the distance b/w the datapoint and the hyperplane the more certain we should be



E.g.: $d=1$

This distance can be positive or negative. If it's on the other side, it is negative



- The distance b/w the new point $\underline{x}$ and the decision boundary indicates the probability of our prediction.
  - As $d \to +\infty$, $\hat{P}_{\underline{w}}(+1|\underline{x}) \to 1$
  - As $d \to -\infty$, $\hat{P}_{\underline{w}}(+1|\underline{x}) \to 0$
  - When $d \approx 0$, $\hat{P}_{\underline{w}}(+1|\underline{x}) \approx 0.5$

Lecture Outline

we want to predict with randomness $\hat{P}_{\underline{w}}(y=+1 \mid \underline{x})$

what hypothesis Set $\mathcal{H}$ should we use

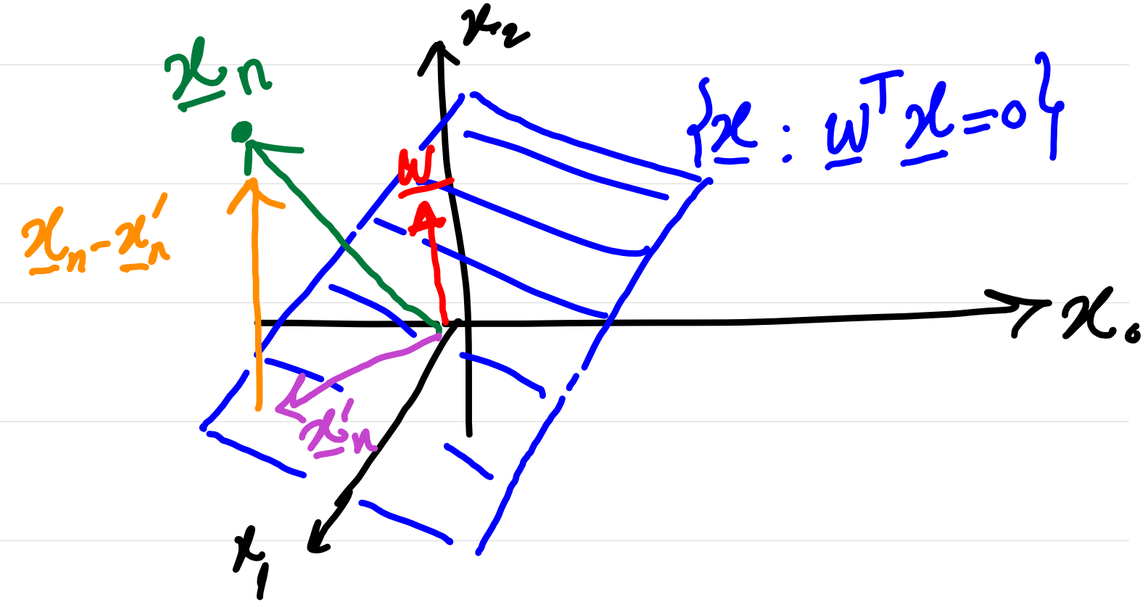
Let's use hyperplane $\underline{w}^T \underline{x} = 0$, and make prediction based on the distance b/w the point and the hyperplane

what is the distance?

Bingo! $\mathcal{H}$ is

Let's find the distance b/w a point and a hyperplane

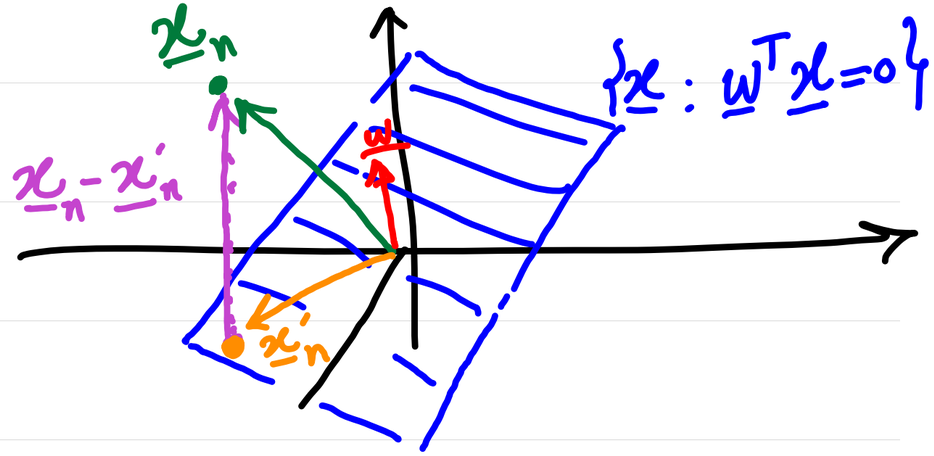
- Assume $\{\underline{x} : \underline{w}^T \underline{x} = 0\}$ is the decision boundary (hypeplane.)
  - WLG, assume that $\|\underline{w}\|_2 = 1$.
  - Recall: $\underline{w}$ is orthogonal to the hyperplane $\{\underline{x} : \underline{w}^T \underline{x} = 0\}$



- To find the distance b/w $\underline{x}_n$ and the hyperplane, ...
  - Find the closest point on the hyperplane to $\underline{x}_n$, i.e. project $\underline{x}_n$ onto the hypeplane
  - length $(\underline{x}_n - \underline{x}'_n)$ is what we are looking for.

- Note that $\underline{x}_n - \underline{x}_n'$ is orthogonal to the hyperplane.
- Thus, $\underline{x}_n - \underline{x}_n' = d\,\underline{w}$
- We must find d.
  - $\underline{x}_n - \underline{x}_n' = d\,\underline{w} \Rightarrow \underline{x}_n' = \underline{x}_n - d\,\underline{w}$ ①
  - Since, $\underline{x}_n'$ is on the hyperplane $\underline{w}^T \underline{x}_n' = 0$

$$\Rightarrow \underline{w}^T \underline{x}_n - d\,\underline{w}^T \underline{w} = 0 \Rightarrow \underline{w}^T \underline{x}_n = d\,\|w\|_2^2 = d.$$



Lecture Outline

We want to predict with randomness $\hat{P}_{\underline{w}}(y=+1 \mid \underline{x})$

What hypothesis Set $\mathcal{H}$ should we use

Let's use hyperplane $\underline{w}^T\underline{x}=0$, and make prediction based on the distance b/w the point and the hyperplane

What is the distance?

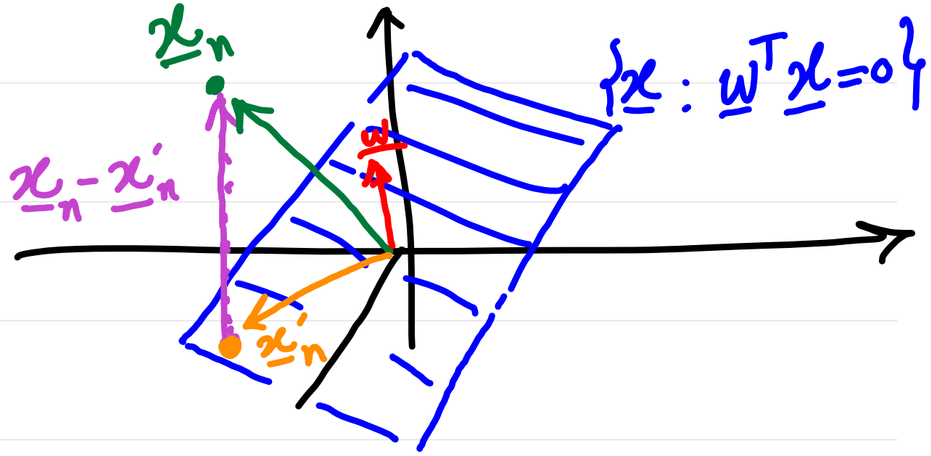
Assuming $\|w\|=1$,

$d = \underline{w}^T\underline{x}$

How do you want to model your predicted likelihood based on d?

Bingo! $\mathcal{H}$ is

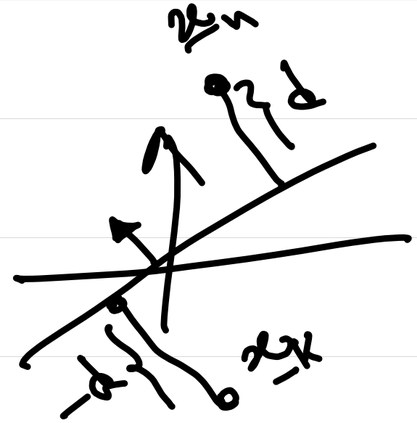
■ So, assuming $\|\underline{w}\| = 1$, the distance b/w $\underline{x}_n$ and the hyperplane is $d = \underline{w}^T \underline{x}_n$



■ Nice! we now know how to formulate the distance.

■ Now, let's see how to formulate our decision rule $\hat{P}_{\underline{w}}(+1|\underline{x}_n)$ based on this distance.

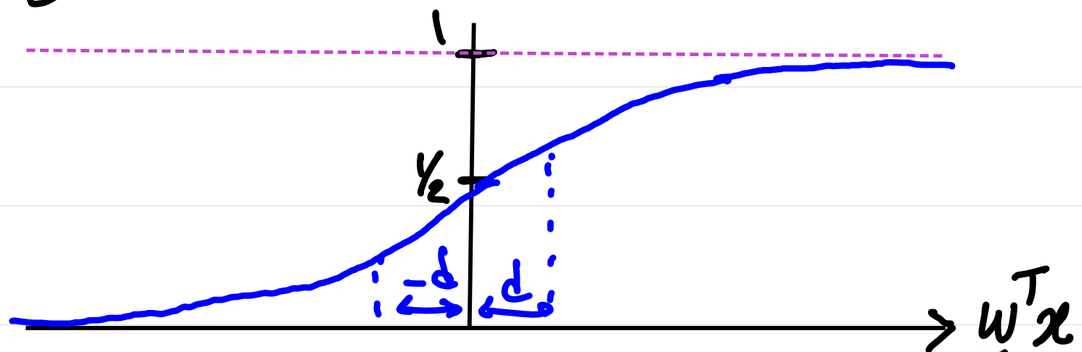
■ We should formulate $\hat{P}_{\underline{w}}(+1|\underline{x}_n)$ in a way that



- When $d = \underline{w}^T \underline{x}_n \to +\infty$, $\hat{P}_{\underline{w}}(+1|\underline{x}_n) \to 1$
- When $d = \underline{w}^T \underline{x}_n \to -\infty$, $\hat{P}_{\underline{w}}(+1|\underline{x}_n) \to 0$
- When $d = \underline{w}^T \underline{x}_n \to 0$, $\hat{P}_{\underline{w}}(+1|\underline{x}_n) \to 0.5$
- We want symmetry, i.e., if $d = \underline{w}^T \underline{x}_n$ and $-d = \underline{w}^T \underline{x}_K$, then

$$\hat{P}_{\underline{w}}(+1|\underline{x}_n) = \hat{P}_{\underline{w}}(-1|\underline{x}_K) = 1 - \hat{P}_{\underline{w}}(+1|\underline{x}_K), \text{ if } \underline{w}^T \underline{x}_n = -\underline{w}^T \underline{x}_K$$

- So, we are looking for a function like this

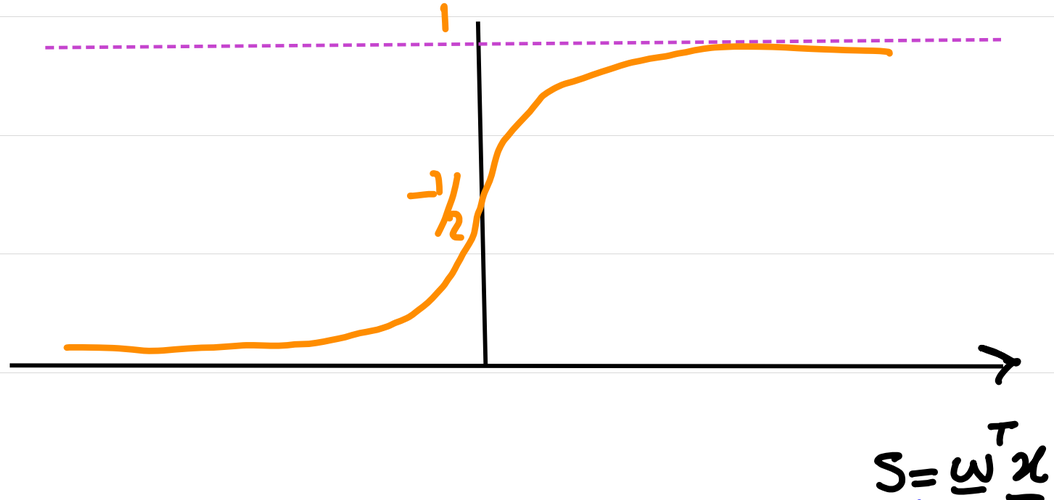


- There is a nice function that we can use to model such behaviour.
  - The "logistic" function

$$\theta(s) = \frac{1}{1+e^{-s}}$$

$$\hat{P}_{\underline{w}}(1|\underline{x}) = \theta(s) = \frac{1}{1+e^{-s}}$$

$$= \frac{1}{1+e^{-\underline{w}^T\underline{x}}}$$



Lecture Outline

We want to predict with randomness $\hat{P}_{\underline{w}}(y=+1|\underline{x})$

What hypothesis Set $\mathcal{H}$ should we use

Let's use hyperplane $\underline{w}^T\underline{x}=0$, and make prediction based on the distance b/w the point and the hyperplane

How do you want to model your predicted likelihood based on d?

- $\underline{w}^T\underline{x}\to+\infty$, $\hat{P}_{\underline{w}}(1|\underline{x})\to 1$
- $\underline{w}^T\underline{x}\to 0$, $\hat{P}_{\underline{w}}(1|\underline{x})\to 1/2$
- $\underline{w}^T\underline{x}\to-\infty$, $\hat{P}_{\underline{w}}(1|\underline{x})\to 0$
- Symmetry

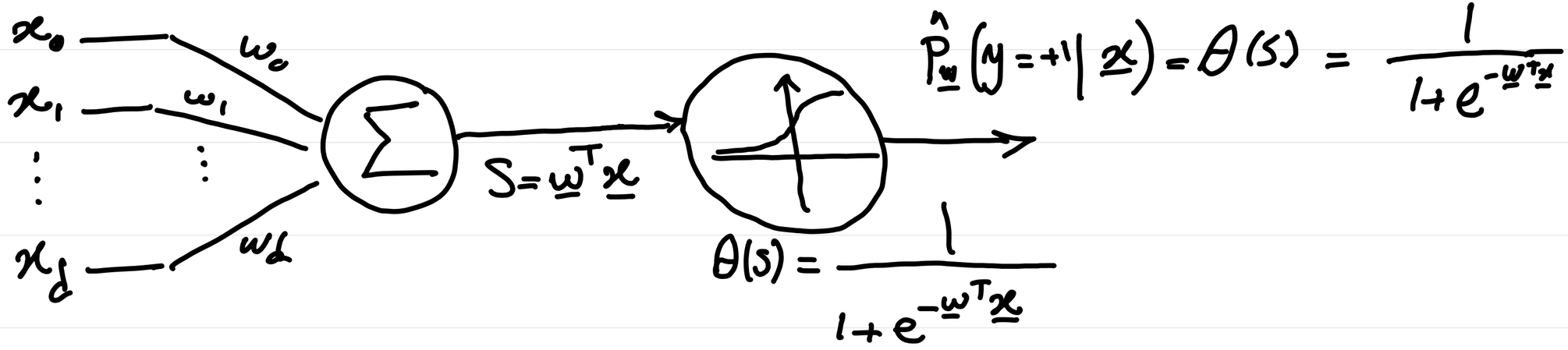
What is the distance? Assuming $\|w\|=1$, $d=\underline{w}^T\underline{x}$

Logistic function is a suitable choice $\theta(s)=\frac{1}{1+e^{-s}}$

Bingo! $\mathcal{H}$ is $\hat{P}_{\underline{w}}(1|\underline{x})=\frac{1}{1+e^{-\underline{w}^T\underline{x}}}$

Illustrating Logistic Regression Model as Neuron

- Given : $D = \{(\underline{x}_n, y_n)\}_{n=1}^{N}$, $\underline{x}_n \in \mathbb{R}^{d+1}$, $y_n \in \{+1, -1\}$
- Unknown Target Function : $y = f(\underline{x})$
- Hypothesis Set : $\hat{y} = \theta(\underline{w}^T \underline{x}) = \dfrac{1}{1+e^{-\underline{w}^T \underline{x}}}$



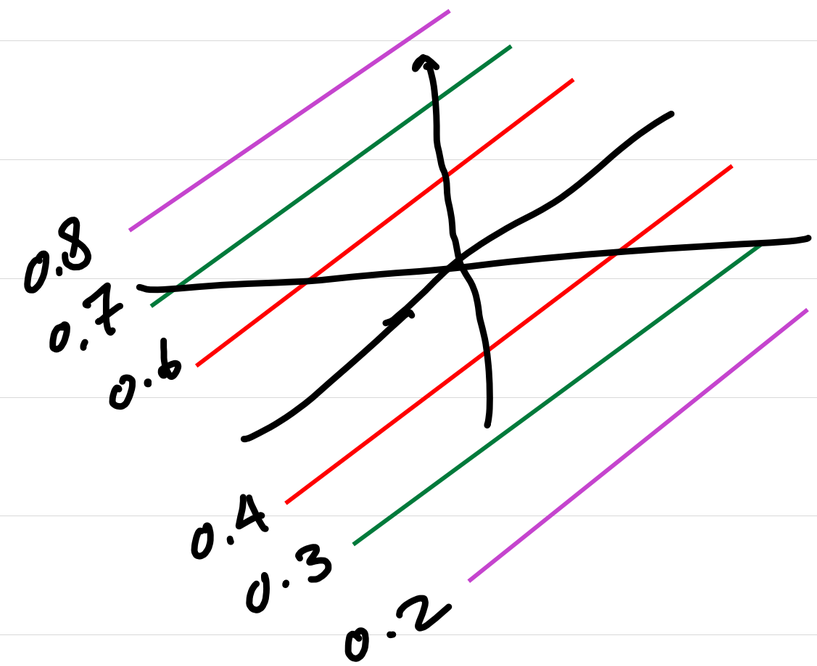
- $\theta(S)$ is called the "logistic function."
- Logistic function is an example of "Sigmoid functions"
  - "Sigmoid" functions is the class of functions that are "S" shaped.

Remarks:

① This is a soft threshold, in contrast to the hard threshold used by linear classification

② $\hat{P}_{\underline{w}}(-1|\underline{x}) = 1 - \hat{P}_{\underline{w}}(+1|\underline{x}) = 1 - \dfrac{1}{1+e^{-\underline{w}^T\underline{x}}}$

$= \dfrac{1}{1+e^{\underline{w}^T\underline{x}}} = \theta(-\underline{w}^T\underline{x})$



③ Notation:

$$\hat{P}_{\underline{w}}(1 \mid \underline{x}) = \theta(\underline{w}^T \underline{x}) \qquad \text{estimate} \qquad \mathbb{P}[y=+1 \mid \underline{x}]$$

$$\hat{P}_{\underline{w}}(-1 \mid \underline{x}) = \theta(-\underline{w}^T \underline{x})$$

- We can combine them:

$$\hat{P}_{\underline{w}}(y \mid \underline{x}) = \theta(y\, \underline{w}^T \underline{x}) = \frac{1}{1 + e^{-y\, \underline{w}^T \underline{x}}}, \text{ where } y \in \{+1, -1\}$$

Lecture Outline

We want to predict with randomness

→

Let's use $\mathcal{H}$ as $\hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

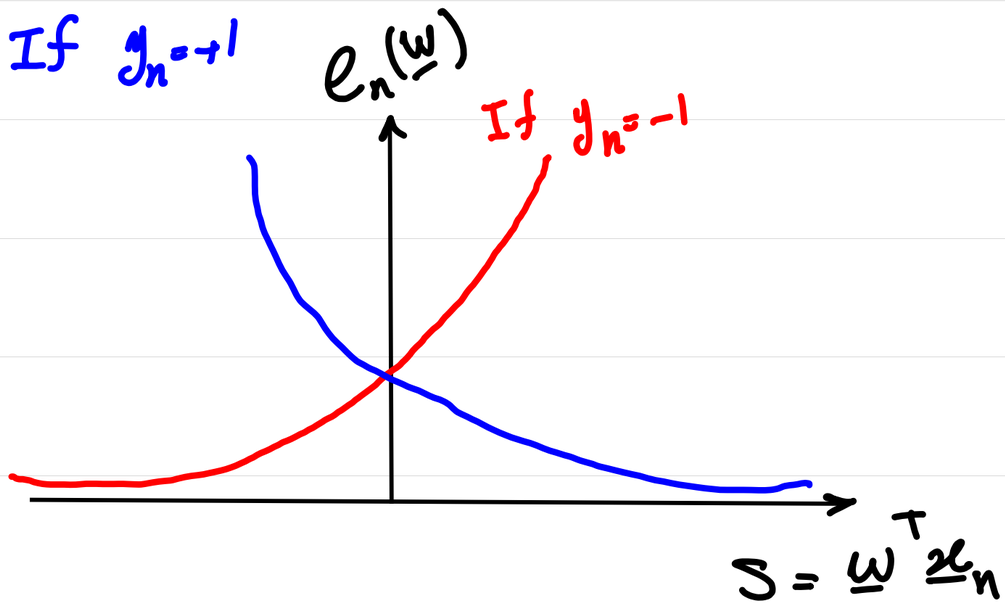
Now, we need to come up with a meaningful $E_{in}(\underline{w})$ to minimize

What is the Error Criterion ?

For the n_{th} example in train data,

natural log

$$e_n(\underline{w}) = \underbrace{-\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)}_{\text{"log-loss" function}} = \log\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$$



Note:

- If $\underline{w}^T \underline{x}_n \gg 0$ and $y_n = 1 \rightarrow$ loss = 0
- If $\underline{w}^T \underline{x}_n \ll 0$ and $y_n = -1 \rightarrow$ loss = 0

Lecture Outline

We want to predict with randomness

→ Let's use $\mathcal{H}$ as $\hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

Now, we need to come up with a meaningful $E_{in}(\underline{w})$ to minimize

← We use log-loss for our error criterion $e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Why is log-loss a reasonable choice?

■ E.g.: Suppose logistic regression outputed a $\underline{w}$ such that

$$\hat{P}_{\underline{w}}(+1 \mid \underline{x}_n) = 0.999$$

$$\left(\Rightarrow \hat{P}_{\underline{w}}(-1 \mid \underline{x}_n) = 0.001\right)$$

the "machine" is very confident that $\hat{y}_n = +1$.

☐ If $y_n = +1$, i.e., the true label was $+1$:

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) = -\log(0.999) \approx 10^{-4}$$

☐ what if $y_n = -1$. What do you expect to see for $e_n(\underline{w})$?

- The "machine" is confident in its incorrect prediction.

☑ Large $e_n(\underline{w})$ ☐ Small $e_n(\underline{w})$

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n) = -\log(0.001) \approx 3$$

■ E.g.: Suppose $\hat{P}_{\underline{w}}(+1|\underline{x}_n) = 0.2$ $\left(\hat{P}_{\underline{w}}(-1|\underline{x}_n) = 0.8\right)$

- If $y_n = +1$, $e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n) = -\log(0.2) \approx 1.61$
- If $y_n = -1$, what do we expect to see?

☐ larger $e_n(\underline{w})$ ☑ smaller $e_n(\underline{w})$

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n) = -\log(0.8) \approx 0.22$$

Lecture Outline

We want to predict with randomness

Let's use $\mathcal{H}$ as $\hat{P}_{\underline{w}}(y|\underline{x}) = \dfrac{1}{1+e^{-y\underline{w}^T\underline{x}}}$

Now, we need to come up with a meaningful $E_{in}(\underline{w})$ to minimize

We use log-loss for our error criterion $e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Why is log-loss a reasonable choice?

Numerical E.g. shows that it makes sense

Seiously? Justifying based on Numerical example? We are better than this

Summary: Logistic Regression

- Given $\mathcal{D} = \{(\underline{x}_n, y_n)\}_{n=1}^{N}$, $\underline{x}_n \in \mathbb{R}^{d+1}$ $y_n \in \{+1, -1\}$,
- find $\underline{w} \in \mathbb{R}^{d+1}$,
- to minimize $E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} e_n(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} \log\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$.
- For a new input $\underline{x}$, predict the probability of being in class y as $\hat{P}_{\underline{w}}(y \mid \underline{x}) = \frac{1}{1 + e^{-y \underline{w}^T \underline{x}}}$