CS725 : LOGISTIC REGRESSION

Goal is binary classification i.e., assign a label $y = \{0, 1\}$ to an input $x \in \mathbb{R}^d$.

Say we want to repurpose a linear regression estimator to do binary classification

Assign $\hat{y} = 1$ if $w^T \hat{x} \geq \tau$, where τ is a predefined threshold
($\hat{y}$ is the predicted label) ($\hat{x}$ is a test sample)

Limitations: (A) How do we pick τ?
(B) Hard to calibrate the strength of the prediction

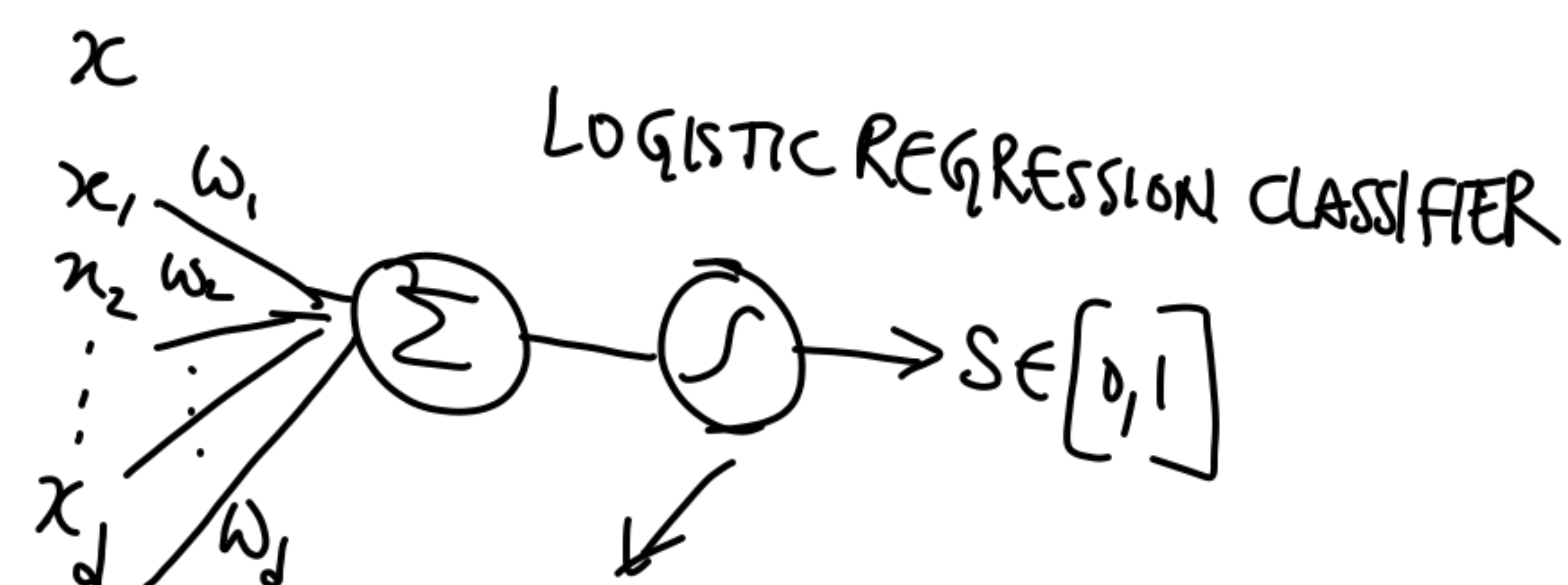
Range of $w^T x$: $(-\infty, \infty)$

$[0,1]$

Squash the $w^T x$ scores down to the range $[0,1]$

LOGISTIC REGRESSION

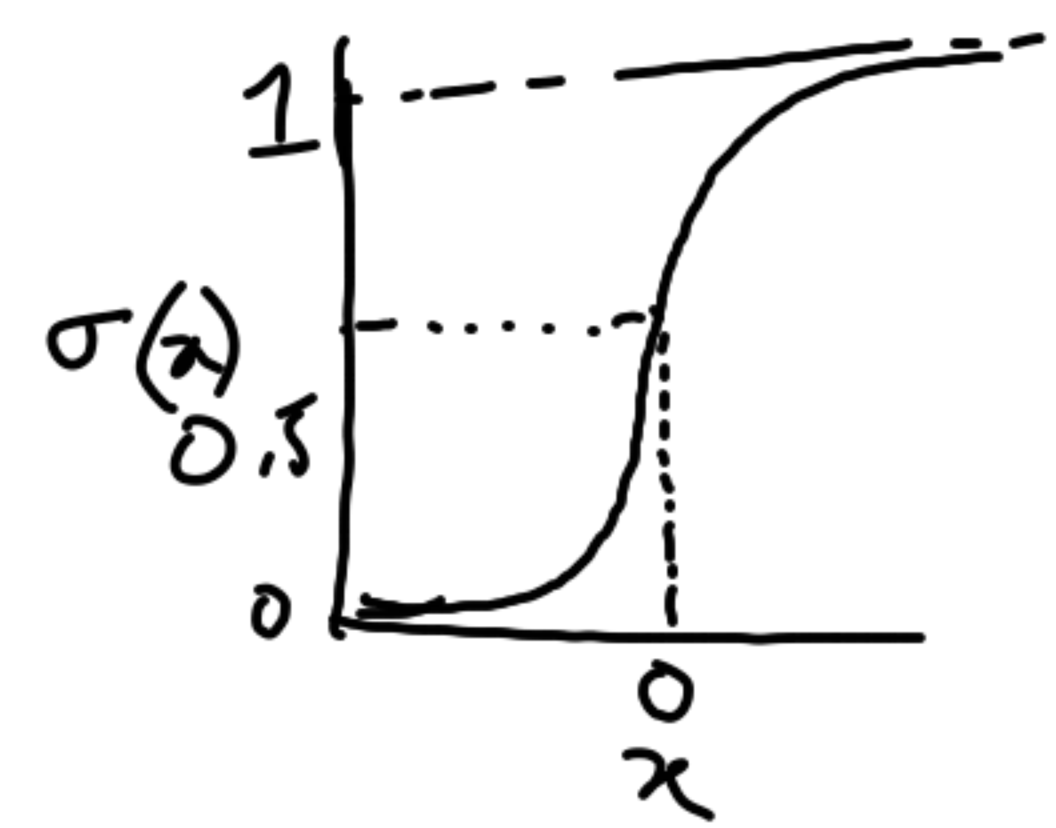
In a logistic regression model,



Choice of function to compress the real range?

SIGMOID FUNCTION [choice of squashing fn for logistic regression]

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



S-shaped function

Good properties of the sigmoid fn: ① Continuous, differentiable

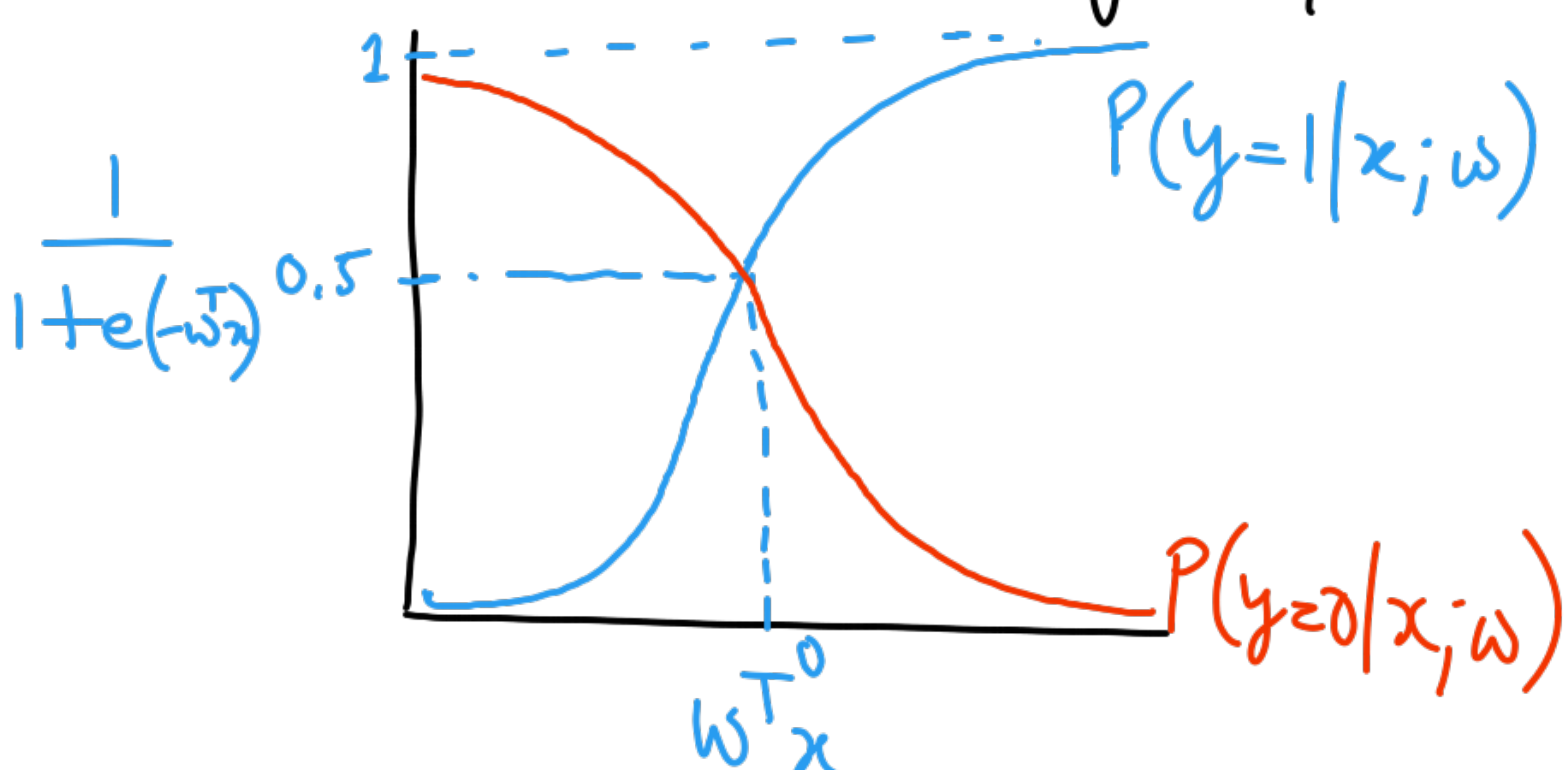
② Control over window around $x=0$ to control the slope of the sigmoid & enable better separation b/w classes

③ Nice derivative in that the derivative can be written using sigmoid fn invocations

$$\sigma'(x) = \sigma(x)\left(1-\sigma(x)\right)$$

In a logistic regression model, $P(y=1|x;w) = \sigma(w^T x)$

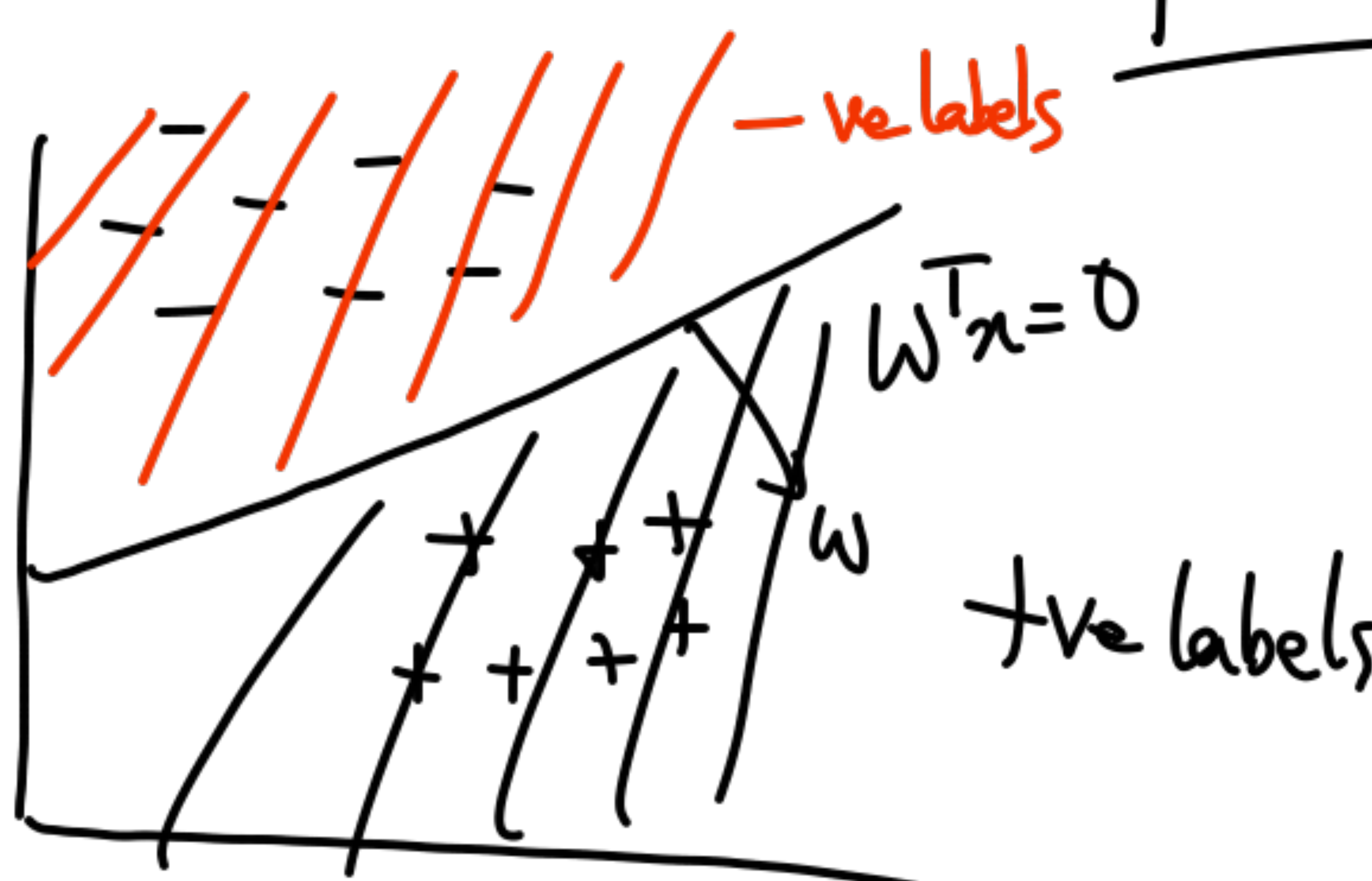
$P(y=0|x;w) = 1 - \sigma(w^T x)$



When do we assign an x to be of label $y=1$?

$$\frac{P(y=1 \mid x; w)}{P(y=0 \mid x; w)} \geq 1 \Rightarrow \frac{\sigma(w^T x)}{1-\sigma(w^T x)} \geq 1$$

$$\Rightarrow e^{w^T x} \geq 1 \Rightarrow \boxed{w^T x \geq 0}$$

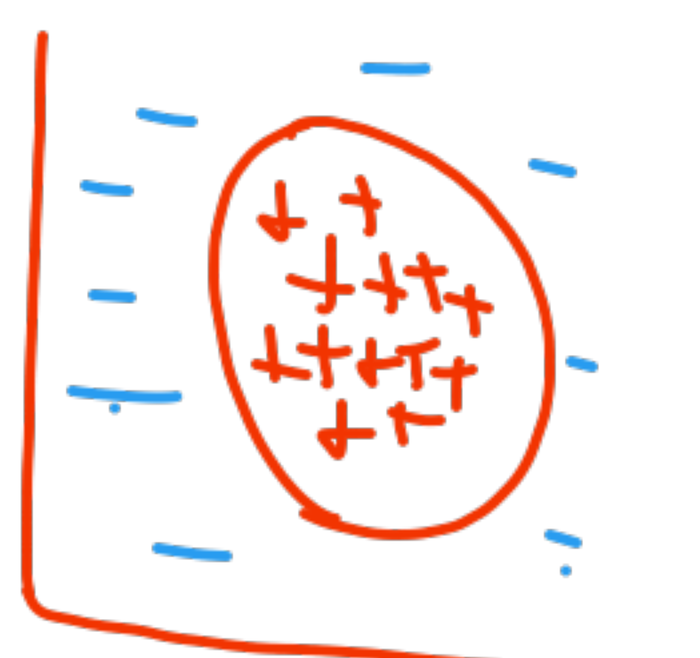


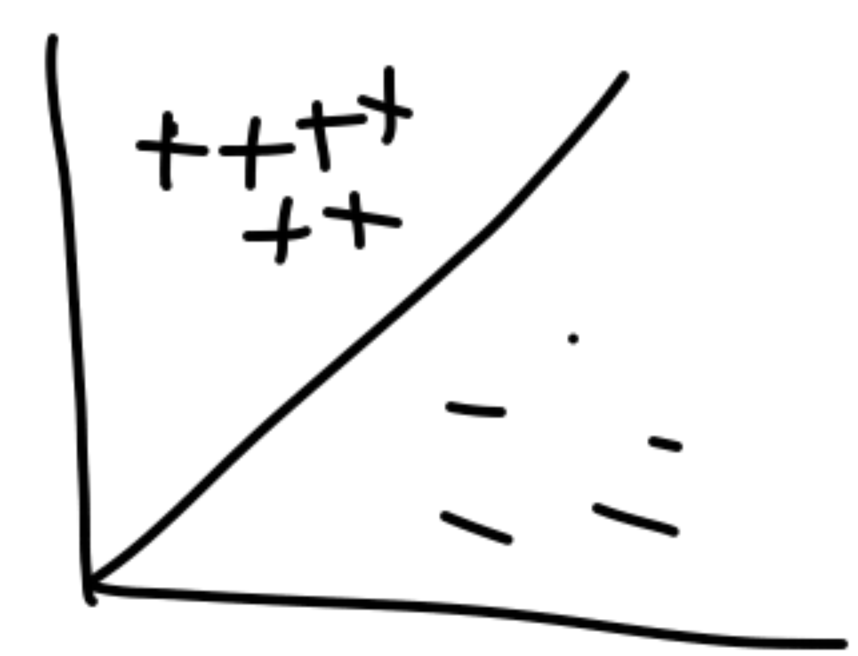
Here $w^T x = 0$ is referred to as a <u>DECISION BOUNDARY</u>

for logistic regression (LR), the decision boundary is a hyperplane i.e. LR is a linear model

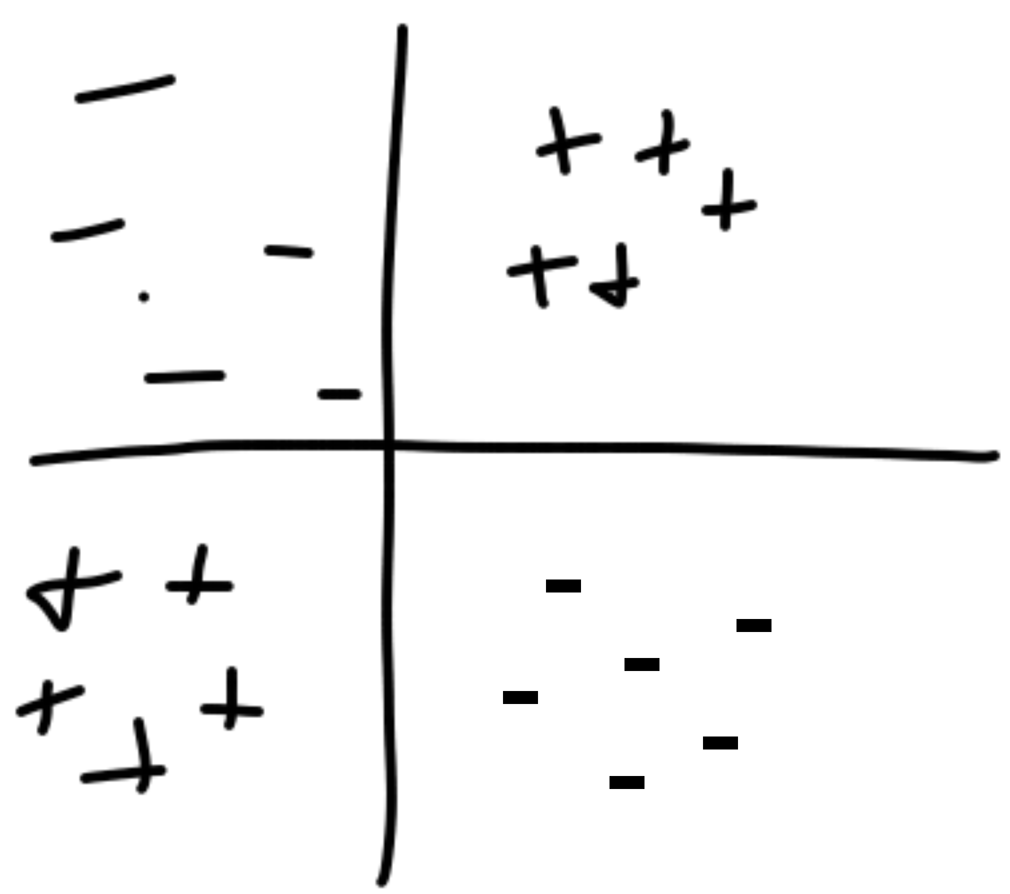
Log. regression is a linear model $\Rightarrow$ it can perfectly classify a <u>linearly separable</u> dataset and the decision boundary is linear

Linearly separable data: A dataset $\mathcal{D}$ is said to be linearly separable if there exists a w s.t. for all +ve examples $w^T x \geq 0$ and for all −ve examples $w^T x < 0$





Linear decision boundary is restrictive



Consider $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where

$$y = \begin{cases} 1 & \text{if } \operatorname{sign}(x_1) = \operatorname{sign}(x_2) \\ 0 & \text{otherwise} \end{cases}$$

There is no linear separator for this particular dataset

If we expand the feature space to include $x_1 x_2$, then the data becomes linearly separable

How do we learn the parameters w of a logistic regression model?

Use <u>maximum likelihood estimation</u>!

$$w_{MLE} = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i \mid x_i; w)$$

// find w that maximizes the conditional likelihood

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} -\log P(y_i \mid x_i; w)$$

Loss function in Logistic Regression

Loss of a single example $L(x_i, y_i; w) = \begin{cases} -\log P(y_i = 1 \mid x_i; w) & \text{if } y_i = 1 \\ -\log\left(1 - P(y_i = 1 \mid x_i; w)\right) & \text{if } y_i = 0 \end{cases}$

In a more compact form, $L(x_i, y_i; w) = -y_i \log P(y_i = 1 \mid x_i; w) - (1 - y_i) \log\left(1 - P(y_i = 1 \mid x_i; w)\right)$

Loss over n training instances $= \sum_{i=1}^{n} L(x_i, y_i; w)$] CROSS-ENTROPY LOSS

CROSS ENTROPY LOSS;

(A) Differentiable

(B) Convex function

(C) No closed form solution. Use gradient descent to find w that minimizes the cross-entropy loss

Finding the gradient of the CE loss :

$$\nabla_w \sigma(w^T x_i) = \sigma(w^T x_i) \cdot \left(1 - \sigma(w^T x_i)\right) \nabla_w w^T x_i$$

$$= \sigma(w^T x_i)\left(1 - \sigma(w^T x_i)\right) x_i$$

$$\nabla_w \log(\sigma(w^T x_i)) = \left(1 - \sigma(w^T x_i)\right) x_i \longrightarrow A$$

$$\nabla_w \log\left(1 - \sigma(w^T x_i)\right) = -\sigma(w^T x_i) x_i \longrightarrow B$$

$\Longrightarrow$ Gradient of CE loss : $\nabla_w \sum_i L(x_i, y_i; w) = \nabla_w \sum_i -y_i \log(\sigma(w^T x_i)) - (1 - y_i) \log(1 - \sigma(w^T x_i))$

$$\nabla_w \sum_i L(x_i, y_i; w) = \nabla_w \sum_i -y_i \log(\sigma(w^T x_i)) - (1-y_i) \log(1 - \sigma(w^T x_i))$$

Using (A) and (B) from earlier, we have:

$$\nabla_w \sum_i L(x_i, y_i; w) = \sum_i \big(\underbrace{\sigma(w^T x_i)}_{\hat{y}_i} - y_i\big) x_i$$

$$= \sum_i (\hat{y}_i - y_i) x_i$$

Very similar to the gradient obtained for linear regression except $\hat{y}_i = \sigma(w^T x_i)$

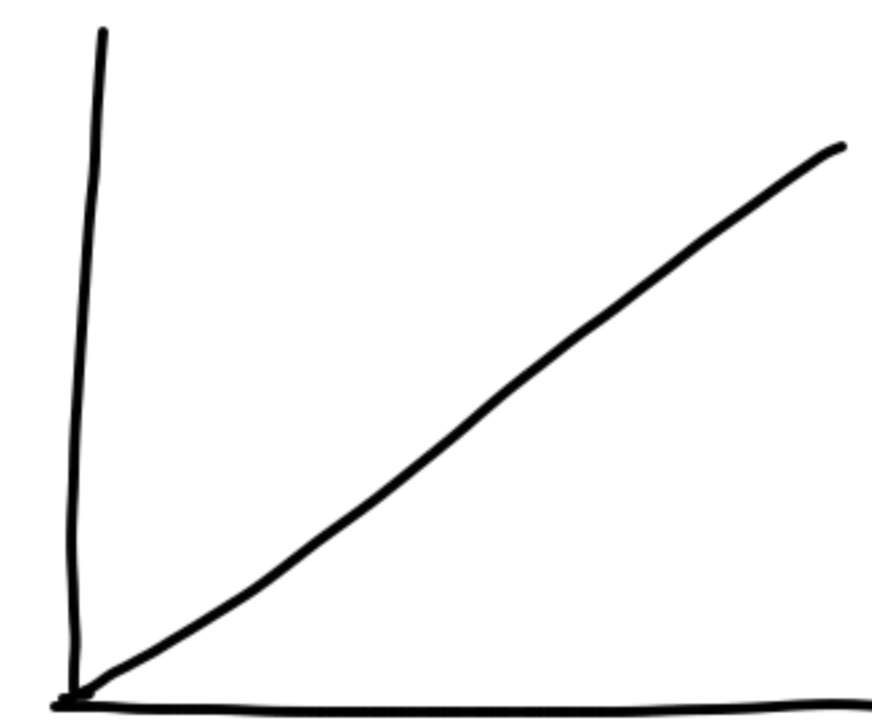
(opposed to $\hat{y}_i = w^T x_i$ for linear regression)

Given the gradient, do (S)GD with the following update rule:

$$w \leftarrow w - \eta \sum_i (\hat{y}_i - y_i) x_i, \quad \hat{y}_i = \sigma(w^T x_i)$$

Consider the following linear separator $x_1 - 2x_2 = 0$

with $w_1 = 1$, $w_2 = -2$



Let's scale up w with a constant factor. E.g.

$w_1 = 10$, $w_2 = -20$

OR $w_1 = 10^4$, $w_2 = -2 \times 10^4$

Which of these w's is a logistic regression model likely to choose?