Plan

Review

Logistic Regression

↳ Sigmoid

↳ Softmax

Cross Entropy Loss

Logistics

- Zoom!
- check in form (10/15)
- scribed notes :)
- 2-3 work and struggle

Linear Regression

$$(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$$

$$x^{(i)} \in \mathbb{R}^d \qquad y^{(i)} \in \mathbb{R}$$

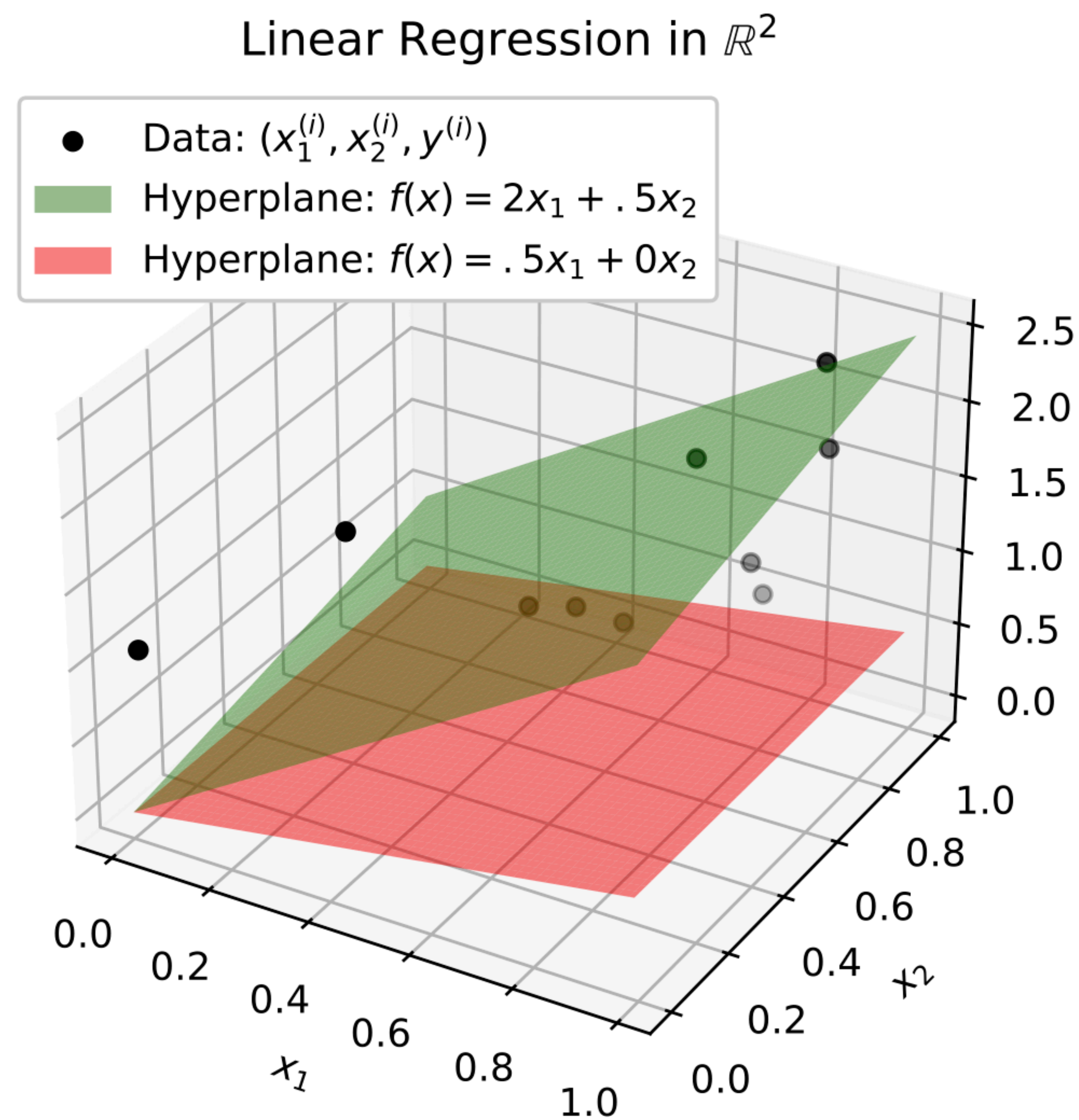
① Model: $f(x) = w \cdot x$

for $w \in \mathbb{R}^d$

② Loss: $\mathcal{L}(w) = \frac{1}{n} \|Xw - y\|_2^2$

③ optimization: $\nabla_w \mathcal{L}(w^*) = 0$

$$\iff w^* = \underbrace{(\underset{d\times n}{X^T} \underset{n\times d}{X})^{-1} \underset{d\times n}{X^T} \underset{n\times 1}{y}}_{d\times 1}$$



Motivation

- What if labels are classes (rather than values)?
- What happens when we can't find the exact optimal?

(image, cat status)

$$= \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{\text{red pixels}} \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{\text{green pixels}} \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{\text{blue pixels}} = \overset{x}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}$$

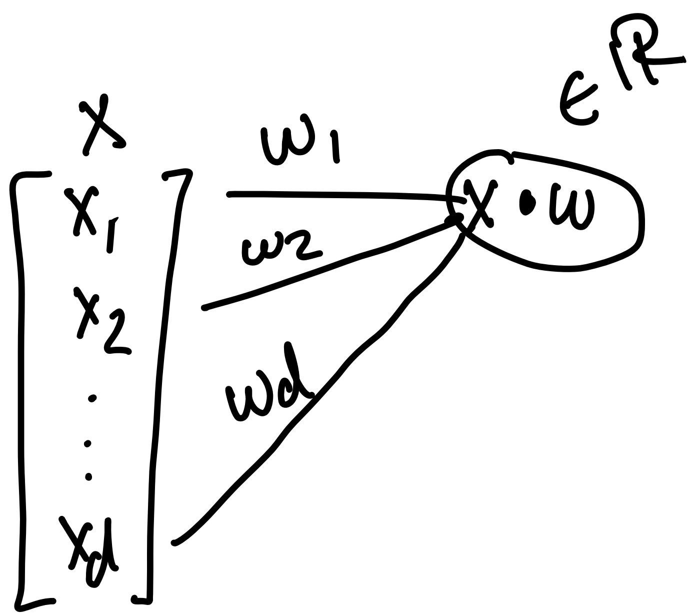
$y \in \{0, 1\}$

Supervised Binary Classification

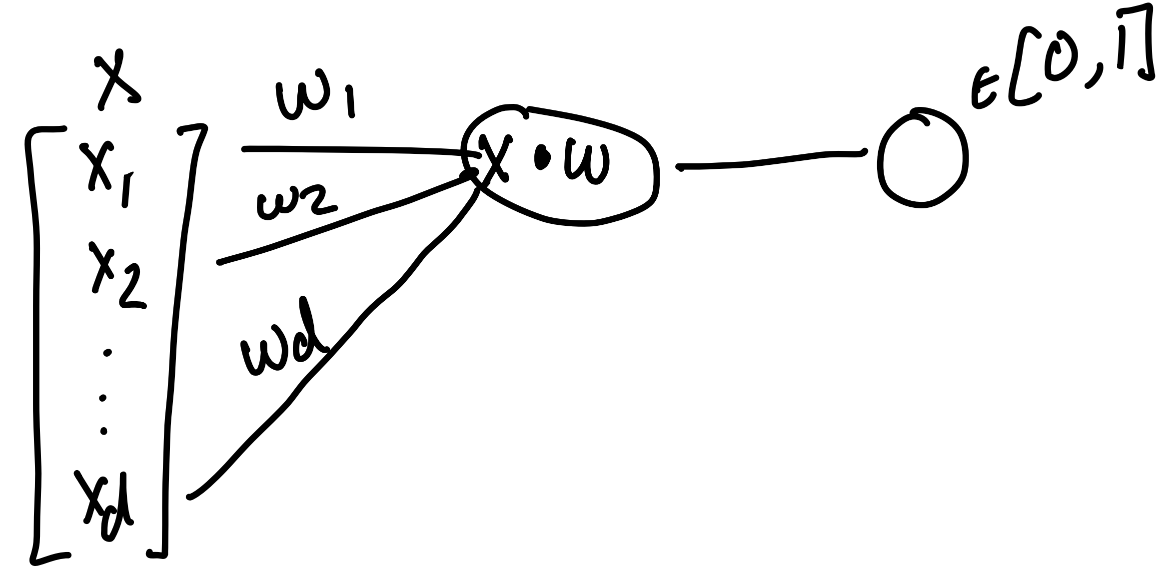
$\ldots (x^{(i)}, y^{(i)}) \ldots$

$x^{(i)} \in \mathbb{R}^d \qquad y^{(i)} \in \{0,1\}$

Goal: $f(x^{(i)}) =$ probability of positive class



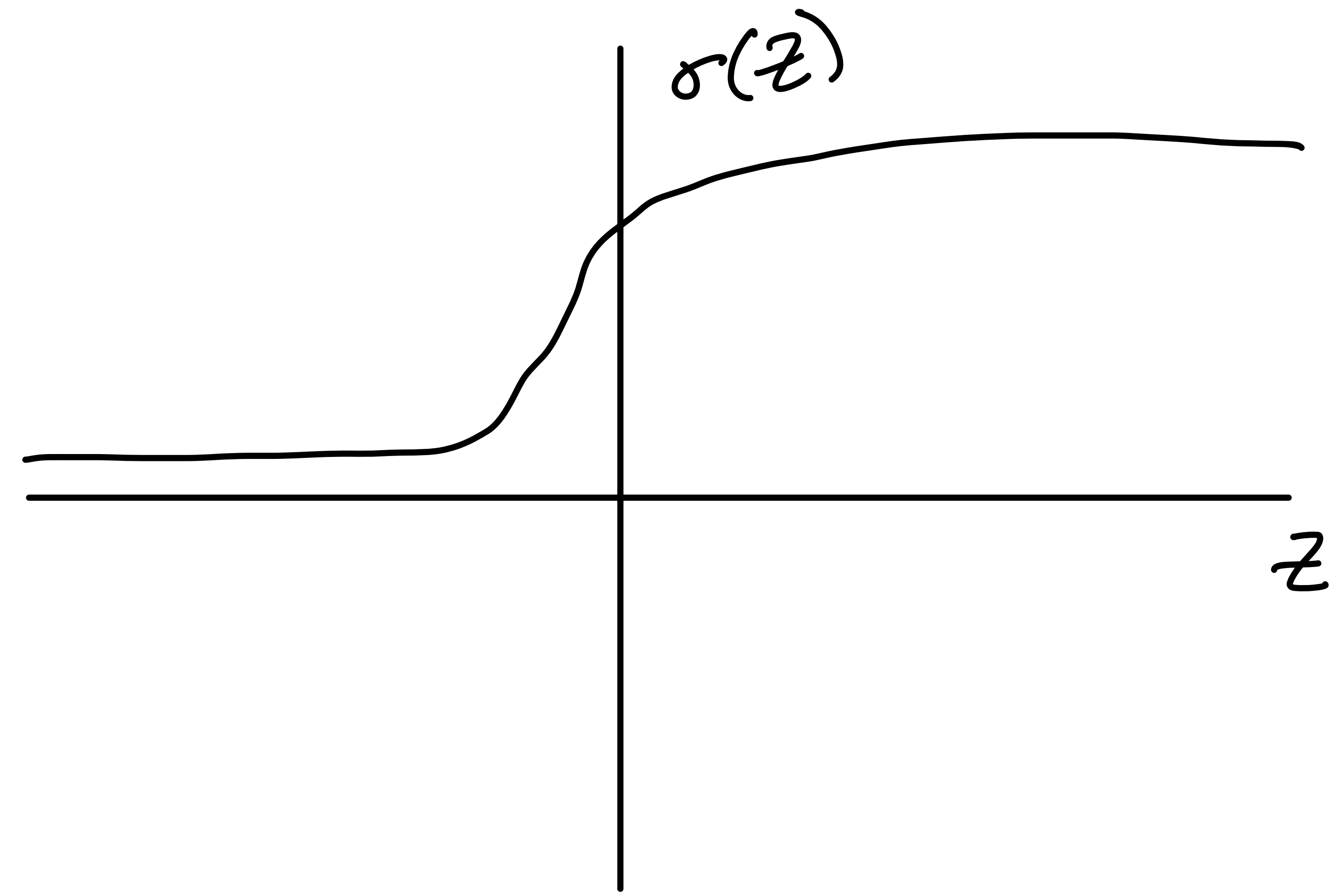
Linear Regression

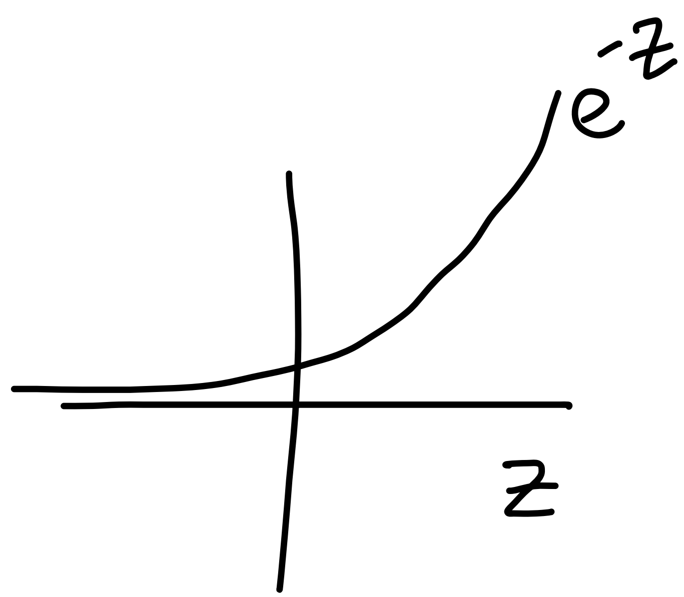


Logistic Regression

Sigmoid

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





$$\lim_{z \to \infty} \frac{1}{1+e^{-z}} = \frac{1}{1+0} = 1$$

$$\lim_{z \to -\infty} \frac{1}{1+e^{-z}} = \frac{1}{\infty} = 0$$

Multiple Classes

$\ldots (x^{(i)}, y^{(i)}) \ldots$

$x^{(i)} \in \mathbb{R}^d \qquad y^{(i)} \in \{0, 1, \ldots, k\}$

$f : \mathbb{R}^d \to [0,1]^k$

Probability distribution

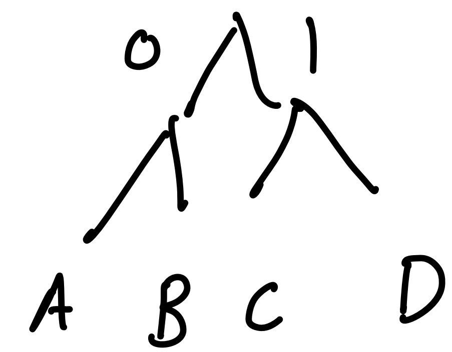
1. non-negative
2. sums to 1

Goal: $f(x^{(i)})_\ell =$ prob of class ℓ

What should the architecture be?

Cross Entropy (aside)

Communicate A, B, C, D



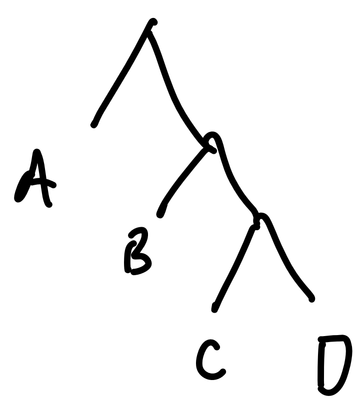
Approach #1: 00, 01, 10, 11

But what if A more likely? e.g., A - 1/2, B - 1/4, C - 1/8, D - 1/8

1, 2, 3, 3

Then

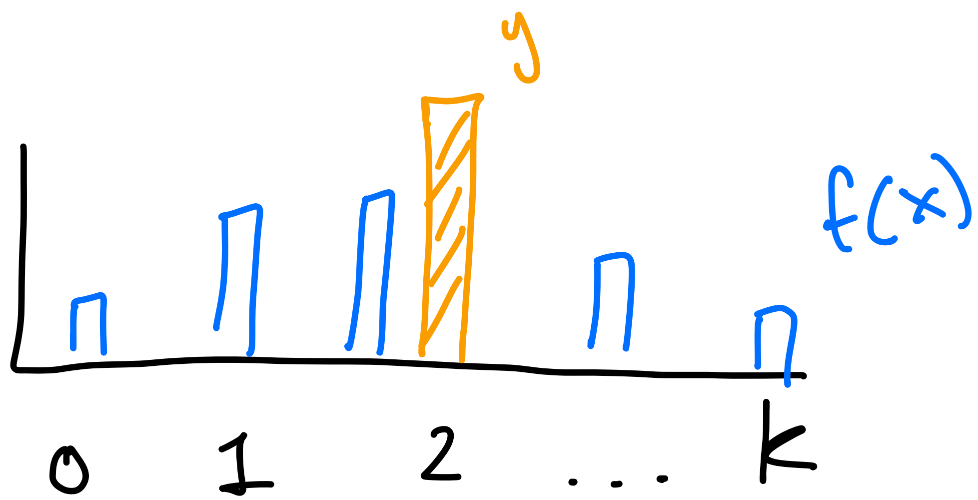
$l_i = -\lceil \log_2 q_i \rceil$



$H(q) = \text{Entropy} = \mathbb{E}[\#\text{ bits}] = \mathbb{E}_{i \sim q}[l_i] = -\mathbb{E}_{i \sim q}[\log_2 q_i] = -\sum_{i=1}^{K} q_i \log_2(q_i)$

$H(p,q) = \text{Cross entropy} = \text{communicate w/ wrong distribution} = -\mathbb{E}_{i \sim p}[\log_2 q_i]$

Loss



Goal: Measure distance between <u>distributions</u> y and $f(x)$

Cross entropy between p and q:

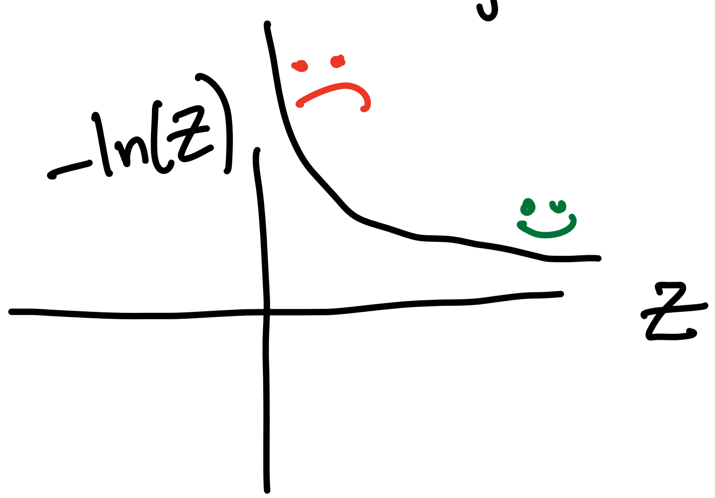
$$H(p,q) = -\mathop{\mathbb{E}}_{j\sim p}\left[-\ln(q(j))\right] = -\sum_{j=1}^{K} p(j)\ln(q(j))$$

When p is "one-hot" i.e., $\exists\ j^*$ s.t. $p(j^*)=1$ then

$$H(p,q) = -\,p(j^*)\ln(q(j^*)) = -\ln(q(j^*))$$

Cross entropy between p and q:

$$H(p,q) = -\sum_{j=1}^{K} p(j) \ln(q(j)) \overset{\text{"one-hot"}}{=} -\ln(q(j^*))$$



<u>Optimization</u>

- Exact is not doable (you'll see on problem), but yet we can still find good weights. How?