Reminders

- Notes (investing time is worth it)
- Office Hours: M + W 12:30 - 2, and by appointment
- Discord
  - DM
  - Pset Questions

Context

Regression : $\underline{x}^{(i)} \in \mathbb{R}^d \quad y^{(i)} \in \mathbb{R}$

Classification : $\underline{x}^{(i)} \in \mathbb{R}^d \quad y^{(i)} \in \{0, \dots, k\}$

- spam or not spam (binary)
- objects in an image
- next word

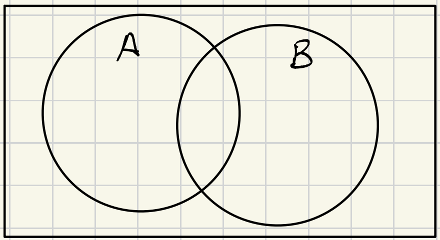
Today: Probability background

Probability

A, B random events

A: it rains tomorrow

B: I carry an umbrella



$Pr(A)$

$Pr(A \cup B)$

$Pr(A \cap B)$

$Pr(A \mid B)$

Properties:

$$0 \le Pr(A) \le 1$$

$$Pr(A \cap B) = Pr(A)\, Pr(B \mid A)$$
$$= Pr(B)\, Pr(A \mid B)$$

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

"complement"

$\neg A$ = not A

$$Pr(\neg A) + Pr(A) = 1 \iff Pr(\neg A) = 1 - Pr(A)$$

A, B indep iff $Pr(A \cap B) = Pr(A)\, Pr(B)$

Random Variables:

$X \sim D$

sampled from

a distribution

For example, $X \sim$ Dice roll

$Pr(X=1) = Pr(X=2) = \ldots = 1/6$

Bayes Rule:

$$Pr(A|B) = \frac{Pr(B|A)\,Pr(A)}{Pr(B)}$$

Proof?

Maximum A Posteriori

Our regression strategy: Choose model that most likely generated data

Classification analogy: Choose label that is most likely given data

Example: $\underline{X}$ = email , $y = \mathbb{1}$[email is spam]

Posterior :

$$\Pr(Y=1 \mid X=x) \quad \text{vs.} \quad \Pr(Y=0 \mid X=x)$$

Compute using Bayes Rule!

$$\Pr(Y=1 \mid X=x) = \frac{\Pr(Y=1)\,\Pr(X=x \mid Y=1)}{\Pr(X=x)} = \frac{\text{prior} \quad \text{likelihood}}{\text{evidence}}$$

Medical Example:

$X \in \{0,1\}$ outcome of test

$Y \in \{0,1\}$ disease

- Rare (1% population)
- 5% FPR
- 10% FNR

Naive Bayes Classifier (with binary features)

assume independent features

Naive assumption: $$\Pr(X = x \mid Y = 1) = \prod_{i=1}^{d} \underbrace{\Pr(X_i = x_i \mid Y = 1)}_{p_i^{(1)} \text{ or } (1 - p_i^{(1)})}$$

For example, $$\Pr\left(X = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \middle| Y = 1 \right) = (1 - p_1)\, p_2\, (1 - p_3)(1 - p_4)\, p_5$$

1. Compute priors $\Pr(Y=1)$, $\Pr(Y=0)$
2. Compute observed probs $p_i^{(1)} = \Pr(X_i = 1 \mid Y = 1)$, $p_i^{(0)} = \Pr(X_i = 1 \mid Y = 0)$
3. Compute likelihoods
$$\Pr(X = x \mid Y = 1) = \prod_{i=1}^{d} \Pr(X_i = x_i \mid Y = 1)$$
4. Predict class with higher posterior
$$\operatorname*{argmax}_{i \in \{0,1\}} \frac{\Pr(X = x \mid Y = i)\, \Pr(Y = i)}{\Pr(X = x)}$$

9/18/2025

Reminders

- Quiz
  - going forward, single answer
  - learning rate strategies
    - manual
    - scheduler
    - adaptive

    NOT analytic (just problem)
  - SGD vs GD
    - speed (RAM vs memory)
- Notes
- Student sessions
  - discuss material or psets
  - advertise on discord
  - incentive: organizer can write 1 quiz question
  - selfie = proof

- Psets 1 and 4 due Monday
- Office hours, let's talk!

Review

$x^{(i)} \in \mathbb{R}^d$ $\quad y^{(i)} \in \{0, 1, \dots, k-1\}$

$k = 2$ for now

classification

Today: Build classification model with parameters

Naive Bayes:

Choose $\underset{y \in \{0,1\}}{\operatorname{argmax}} \Pr(Y = y \mid X = x)$

Naive assumption is independence

$$\Pr(X = x \mid Y = y) = \prod_{i=1}^{d} \Pr(X_i = x_i \mid Y = y)$$

supervised (labeled data)

but no parameters

Model & Loss

Idea: Adapt linear regression

Attempt #1:

$$f(x) = \langle w, x \rangle$$

$$\mathcal{L}(w) = | y - \langle w, x \rangle |$$

Attempt #2:

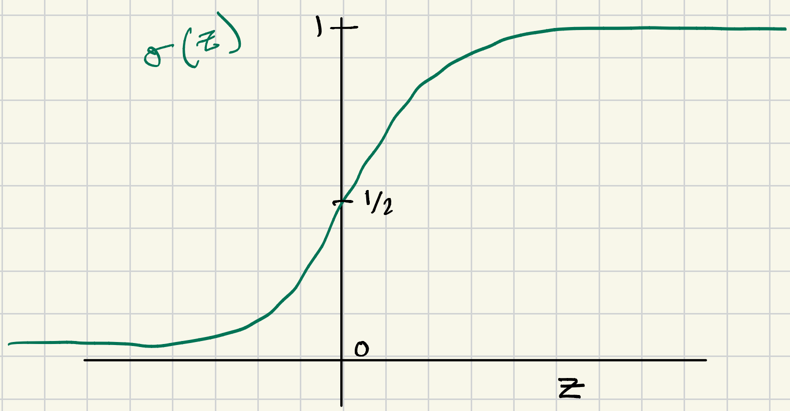
$$f(x) = \langle w, x \rangle$$

$$\mathcal{L}(w) = (y - \langle w, x \rangle)^2$$

Attempt #3:

$$f(x) = \sigma(w, x) \quad \text{where}$$

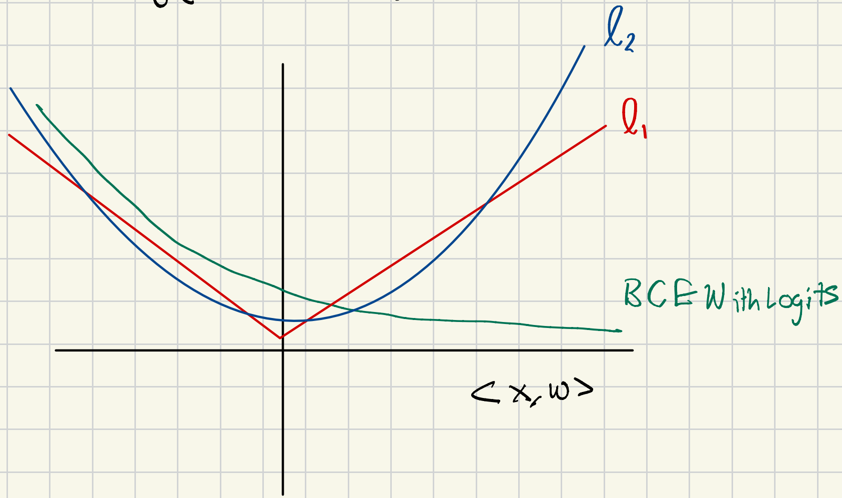
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Binary Cross Entropy Loss

$$\mathcal{L}(w) = -\left[y \log(\sigma(\langle w,x\rangle)) + (1-y)\log(1-\sigma(\langle w,x\rangle))\right]$$

$$\overset{y=1}{=} -\log(\sigma(\langle w,x\rangle))$$



Optimization

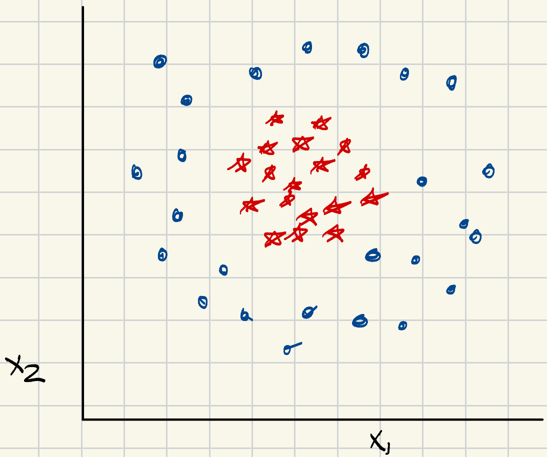
Goal : Compute $\nabla \mathcal{L}(w)$

similar to linear regression?

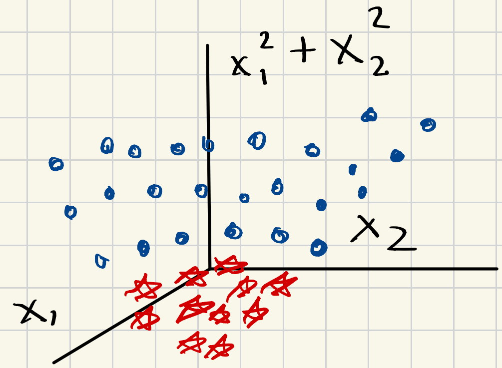
Theorem: $\nabla \mathcal{L}(w) = X^\top(\sigma(Xw) - y)$

can we solve for w^*?

Non-linear Transformation



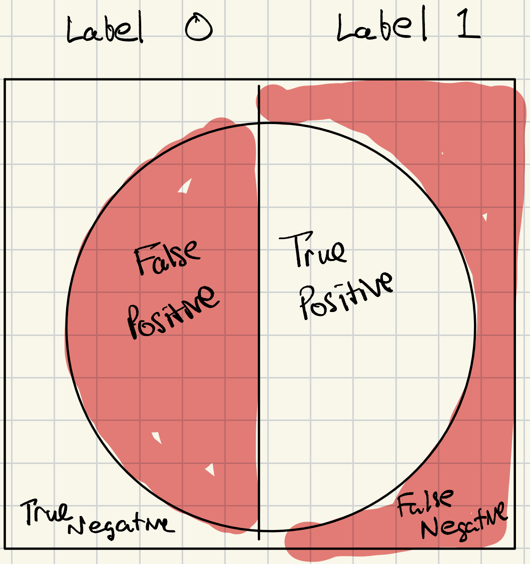
not linearly separable



linearly separable

Measuring Error in Binary Classification

$$\text{error rate} = \frac{\#\,\text{incorrect}}{\#\,\text{predictions}}$$



TPR =

FPR =

Precision =

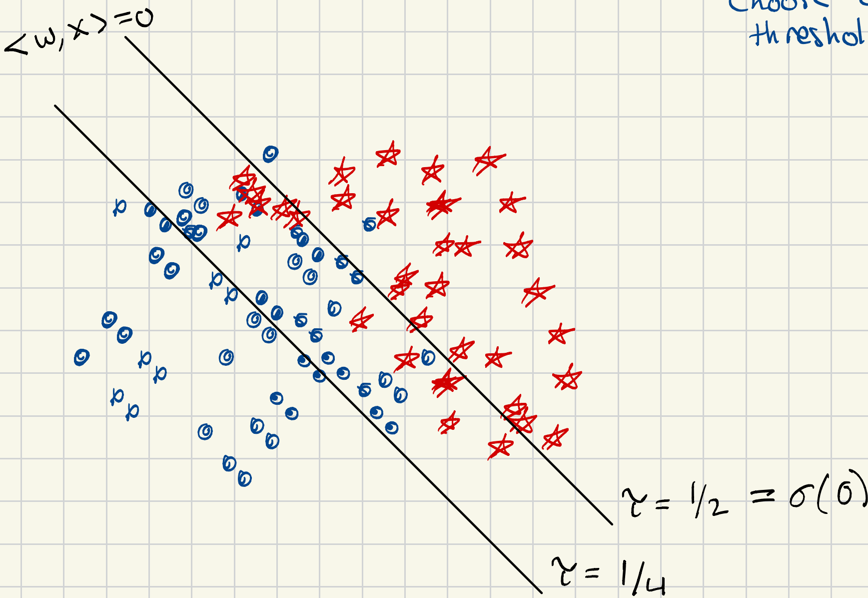
A Hidden Lever...

If incorrect predictions, change threshold

Let $\quad p^{(i)} = f(x^{(i)}) = \sigma(\langle w, x^{(i)} \rangle)$

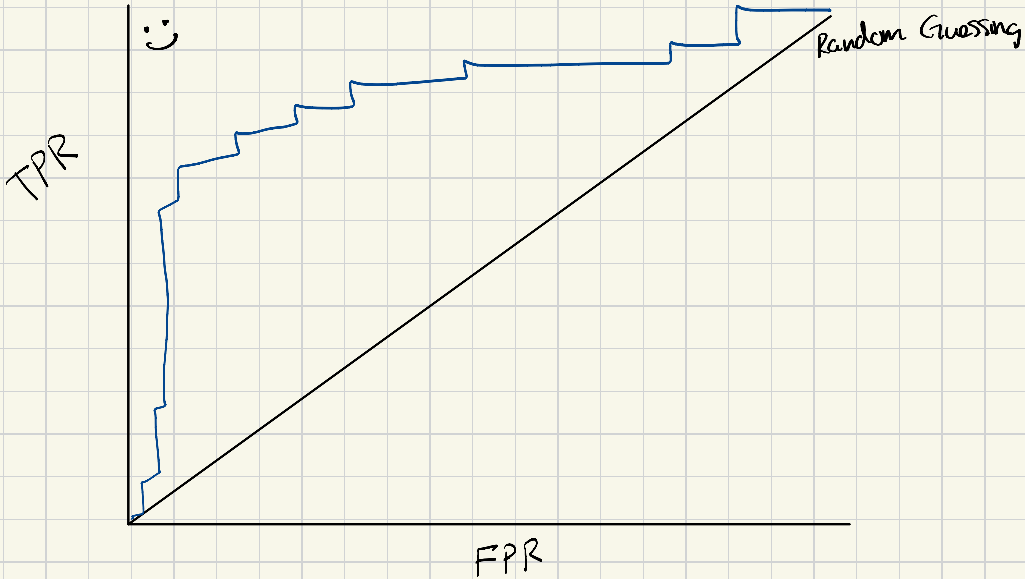
Generally, $\quad \hat{y}^{(i)} = p^{(i)} > 1/2$

Choose any threshold!



Each τ gives a different FPR and TPR!

Receiver Operating Characteristic Curve



Increasing τ leads to...

Area Under Curve (AUC)
gives single number
of performance

Multiple Classes

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

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

$$\begin{bmatrix} \langle w_1, x \rangle \\ \langle w_2, x \rangle \\ \vdots \\ \langle w_k, x \rangle \end{bmatrix}$$

make probabilities

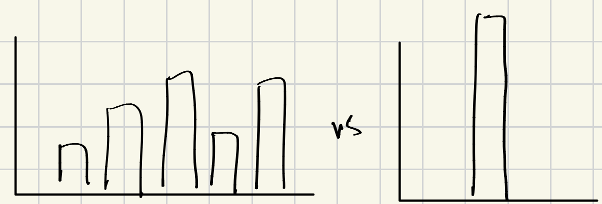
Attempt #1: Sigmoid

Attempt #2: Softmax

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix} \to \begin{bmatrix} e^{z_1} / \sum_\ell e^{z_\ell} \\ e^{z_2} / \sum_\ell e^{z_\ell} \\ \vdots \\ e^{z_k} / \sum_\ell e^{z_\ell} \end{bmatrix}$$

Cross Entropy Loss

$$\mathcal{L}(w) = - \sum_{\ell=1}^{k} \mathbb{1}[y = \ell] \log [f_j(x)]$$



Computing the Gradient ... :)