1/18/2024: Classification

Goal: Predict "class" or "label" for each input
Each class is one of a fixed set of possible classes

- Tumor: Benign vs Malignant?
- Email: Spam vs not spam
- Handwritten digits: 0, 1, 2, ..., 9
- Image: Bird, snake, dog, car, ...

Binary classification (2 possible labels)

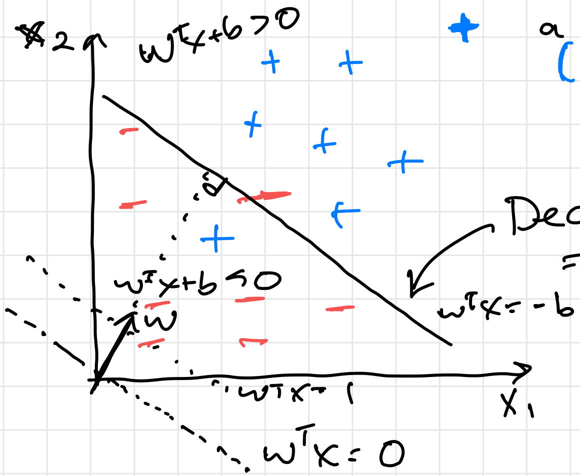
"10-way" classification

Binary classification

- One label $y=1$ "positive"
- One label $y=-1$ "negative"
(sometimes use $y=0$)

"multi-class classification" (>2 possible labels)

Linearity Assumption: Positive & negative points can be separated by a straight line / plane (or almost perfectly separated)



Decision boundary = set of points where

$$w^T x + b = 0$$

parameters

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$

Model prediction:

If $w^T x + b > 0$, predict $y=+1$
If $w^T x + b < 0$, predict $y=-1$

Maximum Likelihood Estimation

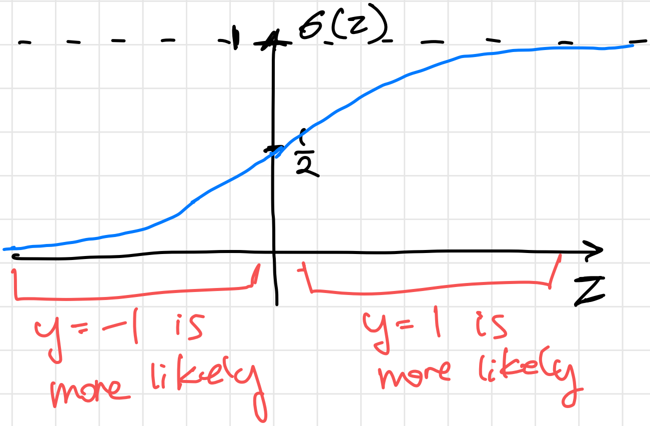
Same idea as linear regression,
Need different probabilistic story

$$p(y=1 \mid x; w) = \frac{1}{1+\exp(-w^T x)} = \sigma(w^T x)$$

"sigmoid" or "logistic" function

omit b b/c can always add a constant feature

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



Convenient fact:

$$p(y \mid x; w) = \sigma(y w^T x)$$

If $y=1$, true by definition

If $y=-1$, $\sigma(-w^T x)$

$$= \frac{1}{1+\exp(w^T x)}$$

$$= \frac{\exp(-w^T x)}{\exp(-w^T x)+1}$$

$$= 1 - \frac{1}{1+\exp(-w^T x)}$$

$$= 1 - p(y=1 \mid x; w)$$

Now: To learn w,
Choose w that maximizes log likelihood of data

$$\log \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; w)$$

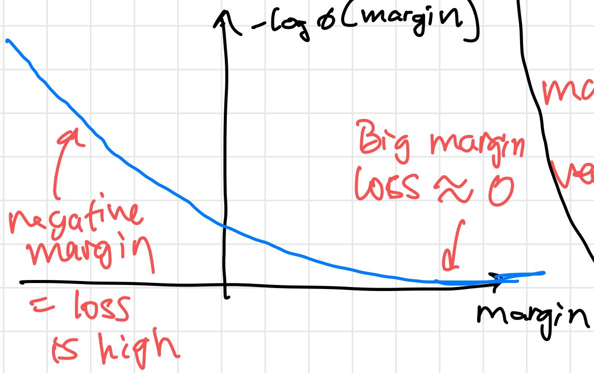
$$= \sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}; w)$$

$$= \sum_{i=1}^{n} \log \sigma(y^{(i)} w^T x^{(i)})$$

Last step: multiply by $-\frac{1}{n}$, swap to minimizing

Final loss: $L(w) = \frac{1}{n} \sum_{i=1}^{n} -\log \sigma(\underbrace{y^{(i)} w^T x^{(i)}}_{\text{"margin"}})$

$-\log \sigma(z)$ function measures "badness" of the margin



large margin is good!

If $y^{(i)} = 1$, want $w^T x^{(i)} > 0$

If $y^{(i)} = -1$, want $w^T x^{(i)} < 0$

margin $> 0 \iff$ prediction is correct

very large margin = very far from decision boundary (in correct direction)

Minimize $L(w)$ with gradient descent

- Fact: this $L(w)$ is also convex

Gradient $\nabla_w \frac{1}{n} \sum_{i=1}^{n} -\log \sigma\left(y^{(i)} w^T x^{(i)}\right)$

$$= \frac{1}{n} \sum_{i=1}^{n} -\sigma\left(-y^{(i)} w^T x^{(i)}\right) \cdot \nabla \left[\underbrace{y^{(i)}}_{\text{constants}} w^T \underbrace{x^{(i)}}_{\text{constants}}\right]$$

Fact: $\frac{d}{dz} -\log \sigma(z) = -\sigma(-z)$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{-\sigma\left(-y^{(i)} w^T x^{(i)}\right)}_{\text{scalar}} \cdot \underbrace{y^{(i)}}_{\substack{\text{scalar} \\ (+1/-1)}} \underbrace{x^{(i)}}_{\text{vector} \in \mathbb{R}^d}$$

If $y^{(i)} = 1$: gradient for example i is [negative number] $\cdot \, x^{(i)}$

During G.D., we add multiple of $x^{(i)}$ to w

Increases $w^T x^{(i)}$, increases $p(y^{(i)} = 1 \mid x^{(i)}; w)$

If $y^{(i)} = -1$, reverse happens

✓

What about $\sigma(-y^{(i)} w^T x^{(i)})$?

$= \sigma(-\text{margin})$

If margin large, $\sigma(-\text{margin}) \approx 0$ we're already doing well, no need to update

If margin small, $\sigma(-\text{margin}) \approx 1$

we're getting this example wrong! Need real update to w

Method name: Logistic Regression

Multi-class Classification

Method: Softmax Regression (AKA "Multinomial Logistic Regression")

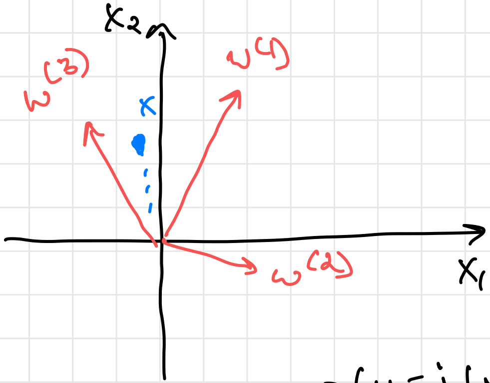
Similar to Logistic Regression but number of classes $C > 2$

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

Model will have $C \times d$ parameters

$$w^{(1)}, w^{(2)}, \dots, w^{(C)} \in \mathbb{R}^d$$

$w^{(j)T} x$ measures how much x "looks like class j"



Decision Rule: For input x

Compute $w^{(1)T}x, \dots, w^{(C)T}x$

Return j with largest value of $w^{(j)T}x$

Probabilistic Story:

$$p(y=j \mid x; W) = \frac{\exp(w^{(j)T}x)}{\sum_{k=1}^{C} \exp(w^{(k)T}x)}$$

all the $w^{(j)}$'s

$$\begin{array}{l} w^{(1)T}x = 1 \\ w^{(2)T}x = -3 \\ w^{(3)T}x = \boxed{2} \end{array} \xrightarrow{\text{exp}} \begin{array}{r} \approx 2.7 \\ \approx 0.1 \\ + \approx 7.4 \\ \hline 10.2 \end{array} \xrightarrow{\text{normalize}} \begin{array}{l} p(y=1 \mid x; w) \approx .27 \\ p(y=2 \mid x; w) \approx .01 \\ p(y=3 \mid x; w) \approx \boxed{.72} \end{array}$$

Maximum Likelihood Estimation

minimize $-\frac{1}{n} \cdot \log$ likelihood

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} -\log p(y^{(i)} \mid x^{(i)}; w)$$

$$= \frac{1}{n} \sum_{i=1}^{n} -w^{(y^{(i)})T} x^{(i)} + \log \sum_{k=1}^{c} \exp(w^{(k)T} x)$$