8/31/2023: Classification

Goal: Predict "label" or "class" for each input from a discrete set of options

- Tumor: Benign or malignant?
- Email: Spam or not spam?

Binary classification: 2 possible labels

- Handwritten digits: 0, 1, 2, ... 9
- Image: Bird, snake, dog, ...

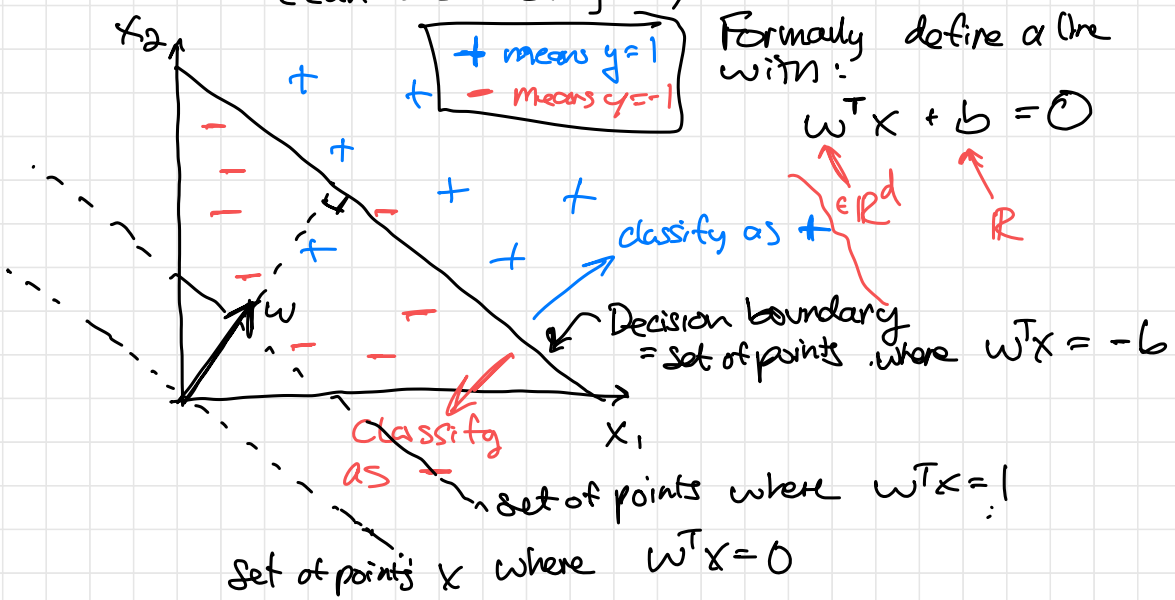
"10-way classification problem"

multi-class classification: >2 possible labels

Binary classification

- One label is $y=1$ "positive"
- Other label is $y=-1$ "negative"

(can also use $y=0$)



Formally define a line with:

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

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

Decision boundary = set of points where $w^T x = -b$

Set of points where $w^T x = 1$

Set of points x where $w^T x = 0$

Machine learning goal: Learn $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that the decision boundary correctly separates +'s and −'s

Model predictions:

If $w^T x + b \geq 0$ predict $y = 1$

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

Maximum Likelihood Estimation

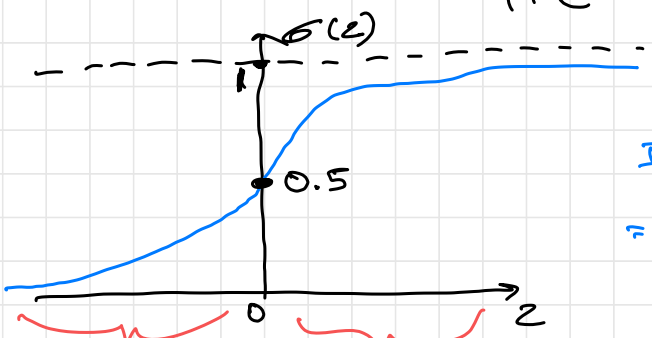
Same idea as linear regression, just need a different probabilistic story

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

omit b for same reason as before

σ "sigmoid" or "logistic" function

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



$y = -1$ is more likely

$y = 1$ is more likely

Convenient fact:

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

If $y = 1$, clearly true

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

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

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

maximize log-likelihood:

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

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

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

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

Final Loss function: $L(w) = \frac{1}{n}\sum_{i=1}^{n} -\log\sigma\left(y^{(i)} w^T x^{(i)}\right)$

For Logistic Regression

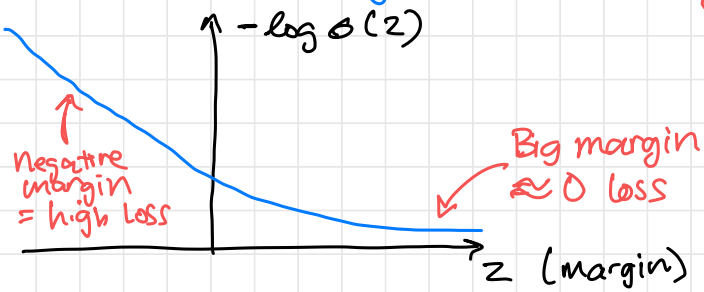
"margin" large margin is good

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

$-\log\sigma(z)$ is a function that measures "badness" of our margin



Minimize $L(w)$ with gradient descent

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

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

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

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

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

positive # · (+/−1) · vector

If $y^{(i)}=1$: gradient has [negative #] $\cdot x^{(i)}$

add multiple of $x^{(i)}$ to w

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

If $y^{(i)} = -1$: gradient has (positive #) $\cdot x^{(i)}$
subtract multiple of $x^{(i)}$ from w
decreases $w^T x^{(i)}$, increase $p(y^{(i)} = -1 \mid x^{(i)}, w)$

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

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

If margin large: $\simeq 0$ ← we're already doing well on this example ⇒ don't need to update

If margin small: $\simeq 1$ ← we're getting this example wrong need to update w

---

Softmax Regression

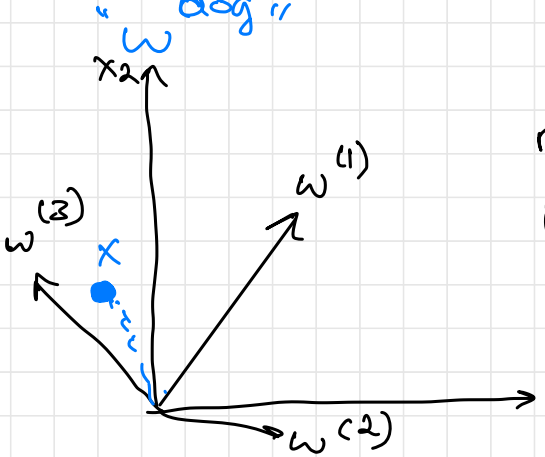
AKA "Multinomial Logistic Regression"

→ Similar to logistic regression but for >2 classes

We now have C classes, $x^{(i)} \in \mathbb{R}^d$

Model will have $C \times d$ parameters $w^{(1)}, \ldots, w^{(C)} \in \mathbb{R}^d$

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



Decision Rule: "dog"
Compute $w^{(1)T} x, \ldots, w^{(C)T} x$
return j with largest $w^{(j)T} x$

Probabilistic Story:

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

$w^{(1)T}x = 1 \xrightarrow{\exp} \approx 2.7$

$w^{(2)T}x = -3 \xrightarrow{\exp} \approx 0.1$

$w^{(3)T}x = 2 \xrightarrow{\exp} \approx 7.4$

Normalize $\longrightarrow$ $p(y=1|x;w) \approx .27$, $p(y=2|x;w) \approx .01$, $p(y=3|x;w) \approx .72$

Predict $y=3$ | 10.2 | 1.0

Maximum Likelihood Estimation:

minimize negative log-likelihood ($* \frac{1}{n}$)

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

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