8/31/2023: Classification

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

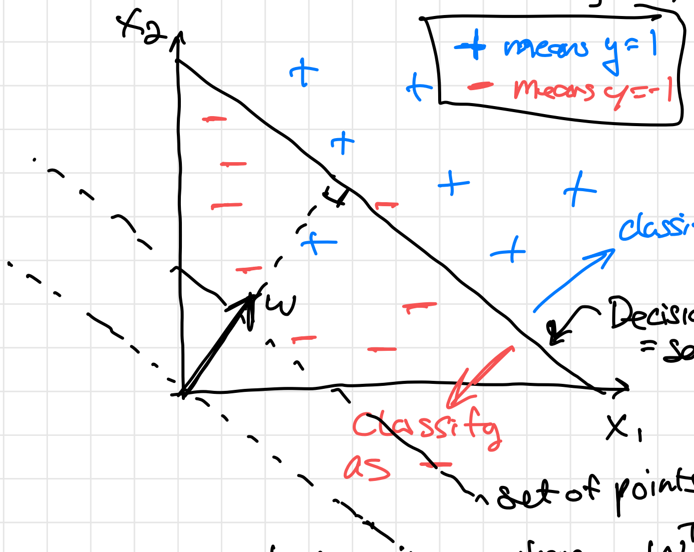
- Tumor: Benign or malignant?
- Email: Spam or not spam?

Binary classification: 2 possible labels

- Handwritten digits: 0, 1, 2, ... 9 "10-way classification problem"
- Image: Bird, snake, dog, ...

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:

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

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

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

Model predictions:

If $w^Tx + b > 0$ predict $y = 1$

If $w^Tx + b \le 0$ predict $y = -1$

Maximum Likelihood Estimation

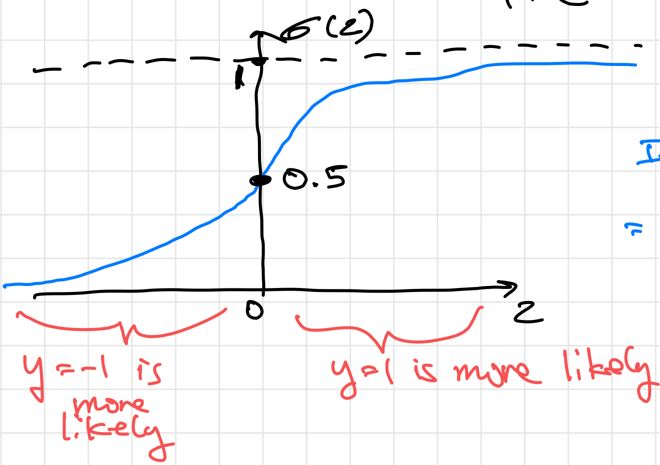
Same idea as linear regression, just need a different probabilistic story

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

omit b for same reason as before

"sigmoid" or "logistic" function

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



Convenient fact:

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

If $y = 1$, clearly true

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

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

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

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: For

Logistic Regression

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

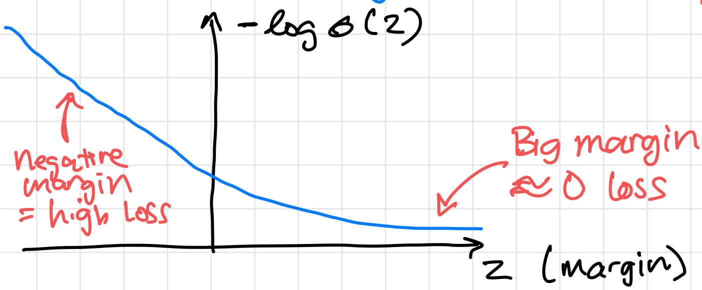
large margin is good

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} -\underbrace{\sigma\left(-y^{(i)} w^T x^{(i)}\right)}_{\text{positive \#}} \cdot \underbrace{y^{(i)}}_{+/-1}\, \underbrace{x^{(i)}}_{\text{vector}}$$

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 $\sigma(\underbrace{-y^{(i)} w^T x^{(i)}}_{})$?

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

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

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

---

Softmax Regression

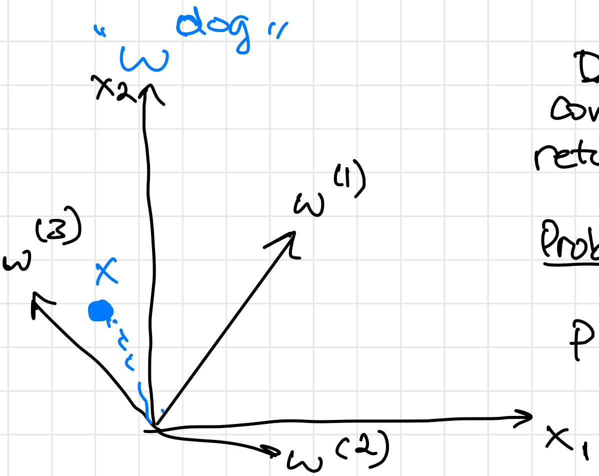
AKA "Multinomial Logistic Regression"

→ 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



"dog"

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

Probabilistic Story:

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

$w^{(1)T}x = 1 \xrightarrow{\exp} \approx 2.7$
$w^{(2)T}x = -3 \rightarrow \approx 0.1$ Normalize $\rightarrow$ $p(y=1|x;w) \approx .27$
$w^{(3)T}x = \boxed{2} \rightarrow \approx 7.4$ $p(y=2|x;w) \approx .01$
$p(y=3|x;w) \approx \boxed{.72}$

Predict $y=3$ $\quad 10.2 \quad\quad 1.0$

Maximum Likelihood Estimation:

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

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

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

(Finishing on 9/5) Now: apply gradient descent w.r.t. $w^{(j)}$

$$\nabla_{w^{(j)}} L(w) = -\frac{1}{n}\sum_{i=1}^{n}\Bigg(\underbrace{\mathbb{1}[y^{(i)}=j]}_{\substack{1 \text{ if } y^{(i)}=j \\ 0 \text{ else}}}x^{(i)}$$

$$- \underbrace{\frac{1}{\sum_{k=1}^{c}\exp(w^{(k)T}x^{(i)})}\cdot \exp(w^{(j)T}x^{(i)})}_{= p(y^{(i)}=j \mid x^{(i)}; w)} \cdot x^{(i)}\Bigg)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big(p(y^{(i)}=j \mid x^{(i)}; w) - \mathbb{1}[y^{(i)}=j]\Big)\cdot x^{(i)}$$

If $y^{(i)} \neq j$: ∇ has positive $\cdot\, x^{(i)}$ $\Rightarrow$ GD subtracts multiple of $x^{(i)}$ from $w^{(j)}$
$\Rightarrow$ makes $w^{(j)T}x^{(i)}$ smaller
$\Rightarrow$ reduces $p(y^{(i)}=j \mid x^{(i)}; w)$

If $y^{(i)}=j$: ∇ has negative $\cdot\, x^{(i)}$ $\Rightarrow$ GD adds multiple of $x^{(i)}$ to $w^{(j)}$