$$L(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^{\boxed{2}}$$ ← why not 4 or absolute value

Maximum Likelihood Estimation

→ posit a probabilistic process that generates our data

→ find parameters that make observed data seem most likely

Coin flip

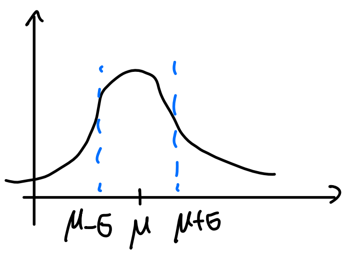
[H, T, H, H, H] ← observed data

p = probability of flipping heads once ← unknown parameter

e.g. if $p = \frac{1}{3}$, $(\frac{1}{3})^4 (\frac{2}{3})$ ← likelihood function of p

---

Linear regression: Assume $y^{(i)}$ drawn from $\underbrace{\text{Gaussian}}_{\text{CLT}}$ w/ mean $\underbrace{\boxed{w^T x^{(i)}}}_{\text{parameter}}$ and variance $\underline{\sigma^2}$ constant



$$P(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Likelihood of data

$$\mathcal{L}(w) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; w)$$ ← parametrized by

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right)$$

maximize $\mathcal{L}(w)$ is equivalent to maximize $\log \mathcal{L}(w)$

$$\log \mathcal{L}(w) = \sum_{i=1}^{n} \left[\underbrace{\cancel{\log \frac{1}{\sigma\sqrt{2\pi}}}}_{\text{constant}} + \left(\frac{-(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right) \right]$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})^2$$

maximize $\log \mathcal{L}(w) \Leftrightarrow$ minimize $L(w)$

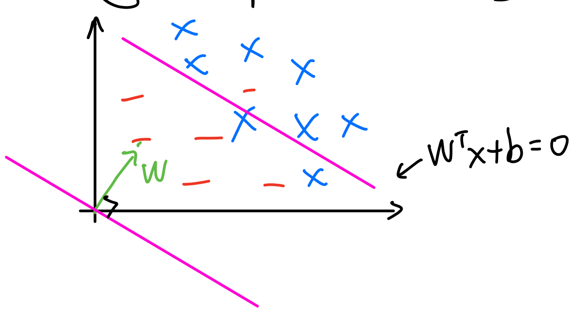
Classification

Goal: predict "label/class" from a discrete set of options

Binary classification : $(1,-1)$, $(1,0)$

Multi-class classification

Modelling assumption of linearity



predictions : if $w^T x + b > 0$, predict $y = 1$

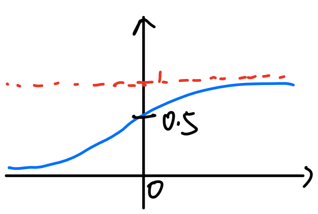
if $w^T x + b < 0$, predict $y = -1$

Use MLE to come up with appropriate loss function

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

sigmoid/ logistic function

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



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

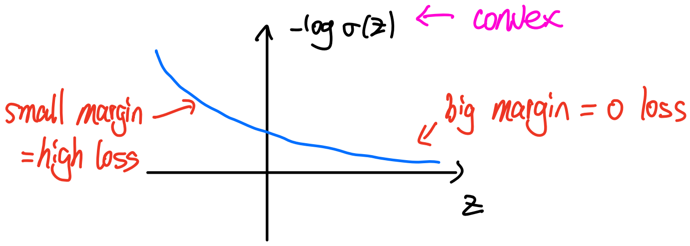
maximize

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

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

equivalent to minimise $J(w) = \sum_{i=1}^{n} -\log \sigma(\underbrace{y^{(i)} w^T x^{(i)}}_{\text{"margin"}})$

margin $> 0 \iff$ prediction is correct



Gradient descent

$J(w)$ is convex

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

$$\nabla J(w) = \sum_{i=1}^{n} -\underbrace{\sigma(-y^{(i)} w^T x^{(i)})}_{\text{pos number}} \cdot \underbrace{y^{(i)}}_{\pm 1} \underbrace{x^{(i)}}_{\text{vector}}$$

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

If $y^{(i)} = 1$, add a multiple of $x^{(i)}$ to w makes $w^T x^{(i)}$ larger

increases $P(y^{(i)} | x^{(i)}; w)$

If $y^{(i)} = -1$, subtract ...

$\sigma(-\text{margin})$ $\begin{cases} \text{if large} & \approx 1 \\ \text{if small} & \approx 0 \end{cases}$

doing well

for improvement