1/19/2023: MLE for Linear Regression, Logistic Regression

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

Why not 4?
Why not absolute value

Maximum Likelihood Estimation

→ Posit a probabilistic process that generated our data
→ Find parameters that make observed data seem most likely

Coin Flips (testbed for MLE)

[Heads, tails, heads, heads, heads] ← observed data

p = probability of flipping heads once ← unknown parameter

eg. if $p = \frac{1}{3}$: $\frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$ ← likelihood given value of p

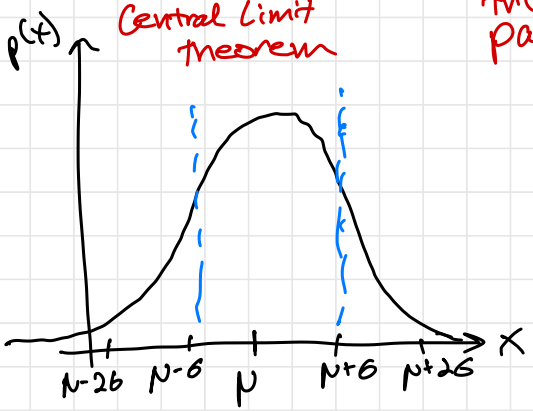
---

Linear Regression: Assume $y^{(i)}$ drawn from Gaussian w/ mean $w^T x^{(i)}$ & variance σ^2 independently

$w^T x^{(i)}$: true parameter

σ^2: constant

Central Limit Theorem



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

$P(A \mid B)$

Likelihood of data

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

↑ "parameterized 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 $L(w)$ is equivalent to maximizing $\log L(w)$

$$\ell(w) = \sum_{i=1}^{n} \left[\log \frac{1}{\sigma\sqrt{2\pi}} + \left(\frac{-(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right)\right]$$

$\ell(w)$: log-likelihood

$\log \frac{1}{\sigma\sqrt{2\pi}}$: constant

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

$\frac{-1}{2\sigma^2}$: neg

maximizing $\ell(w)$ is equivalent to minimizing $L(w)$

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

$\frac{1}{n}$: pos

Linear regression: original loss

Classification

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

- Tumor: Benign or Malignant
- Email: spam vs not spam

Binary classification (2 labels)

- Handwritten digits: is it $0, 1, 2, \ldots, 9$

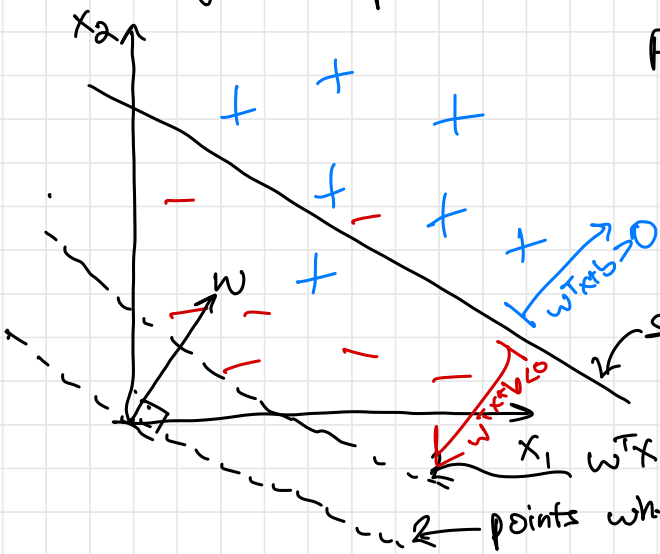
Multi-class classification (>2 labels)

Binary classification terms:

one label is "positive", other is "negative"

$y = 1$ $\quad y = -1$ (or $y = 0$)

Modelling Assumption of linearity



Formally, line is defined by $w^T x + b = 0$

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

Predictions: If $w^T x + b > 0$, predict $y = 1$

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

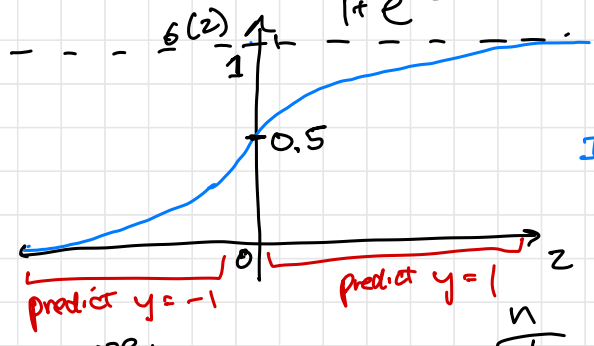
Use MLE to come up with appropriate loss function

(We can omit for same reason as before)

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

↑ "sigmoid" or "logistic" func.

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



Convenient fact:

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

If $y = 1$ true

If $y = -1$: $\sigma(-w^T x) = \frac{1}{1+e^{w^T x}} = \frac{\exp(-w^T x)}{\exp(-w^T x)+1}$

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

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

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

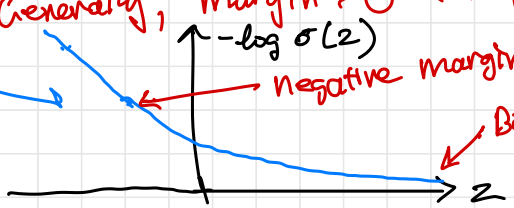
equivalent to minimizing $J(w) = \sum_{i=1}^{n} -\log \sigma(y^{(i)} w^T x^{(i)})$

"margin"

large margin is good

* the loss function for logistic regression

- If $y^{(i)} = 1$, want $w^T x^{(i)}$ large
- If $y^{(i)} = -1$, want $w^T x^{(i)}$ small
- Generally, margin $> 0 \iff$ prediction is correct



Fact: $-\log \sigma(z)$ is convex

Gradient descent time

$J(w)$ is convex (by same rules as linear reg.)

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

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

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

$\sigma(-y^{(i)} w^T x^{(i)})$: positive number; $y^{(i)}$: +/- ?; $x^{(i)}$: vector

If $y^{(i)} = 1$, adding multiple of $x^{(i)}$ to w makes $w^T x^{(i)}$ larger, increases $p(y^{(i)} | x^{(i)}; w)$

If $y^{(i)} = -1$, subtract multiple of $x^{(i)}$ from w

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

If margin large: ≈ 0 (Already doing well)

If margin small: ≈ 1 (Room for improvement)

Prioritization