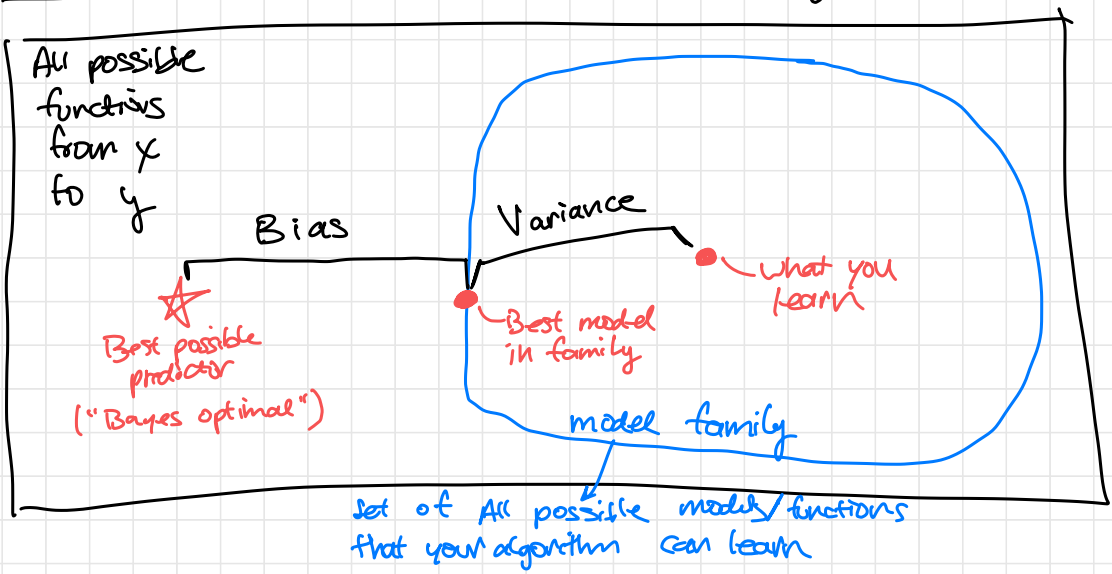
9/7/2023: Bias/Variance, MAP, Normal Equations



Bias: Error because assumptions of ML method used don't exactly match reality

Variance: Error because what you learn ≠ best possible model which is because training data is incomplete

Total error can be decomposed into bias & variance

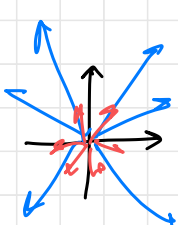
Reduce Bias

Make fewer assumptions
⟺
Make model family bigger

Reduce Variance

make it easier to find best model in family
⟺
make model family smaller

Regularization does this



without reg: w can have very large norm
with reg: w constrained to Ball of small norm

Maximum a Posteriori (MAP) estimation
(extension of maximum likelihood estimation)

Idea: Think about learning as Bayesian inference problem

Think of everything as a random variable

Bayesian probabilistic story:

1. Exists prior distribution over w, $p(w)$
2. Some w gets sampled by nature
3. Dataset D is generated conditioned on w, called $p(D|w)$

Our goal: Infer most likely value of w

i.e. maximize wrt w $p(w|D)$ ← D observed

← most likely w given data

$$p(w|D) = \frac{p(w)\,p(D|w)}{p(D)}$$

$p(D|w)$ = Likelihood = what we maximized during MLE

$p(D)$ doesn't depend on w so ignore it

$p(w)$ ← New

Different choices of $p(w)$ give different regularization terms

For example: Let $p(w) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} e^{-w_j^2/2\sigma^2}$

Gaussian with mean 0 and variance σ

$$\max_w p(w|D) = \max_w p(w)\,p(D|w)$$

$$= \min_w -\log p(w) - \log p(D|w)$$

$-\log p(D|w)$: original MLE loss

$-\log p(w)$

$$-\log P(w) = -\sum_{j=1}^{d} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{-w_j^2}{2\sigma^2}$$

$$= \text{Constant} + \frac{1}{2\sigma^2}\sum_{j=1}^{d} w_j^2$$

$$= \text{constant} + \boxed{\frac{1}{2\sigma^2}} \|w\|^2$$

$$= \lambda$$

small $\sigma \Rightarrow$ large λ

large $\sigma \Rightarrow$ small λ

Closed form solution for linear regression ("Normal Equations")

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

$$\nabla_w L(w) = \frac{1}{n}\sum_{i=1}^{n} 2\left(w^T x^{(i)} - y^{(i)}\right) x^{(i)} = 0$$

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

define:

$$X = \begin{bmatrix} \text{—} x^{(1)} \text{—} \\ \text{—} x^{(2)} \text{—} \\ \vdots \\ \text{—} x^{(n)} \text{—} \end{bmatrix}$$

"design matrix"

$n \times d$ matrix

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

n-dim vector

$$= X^T y$$

$$= \underbrace{X^T X}_{d\times d} \underbrace{w}_{\in \mathbb{R}^d} = \underbrace{X^T y}_{\in \mathbb{R}^d}$$

($X^T X$: $d \times n$ times $n \times d$; $X^T y$: $d \times n$ times n)

So: $$\boxed{w = (X^T X)^{-1} X^T y}$$

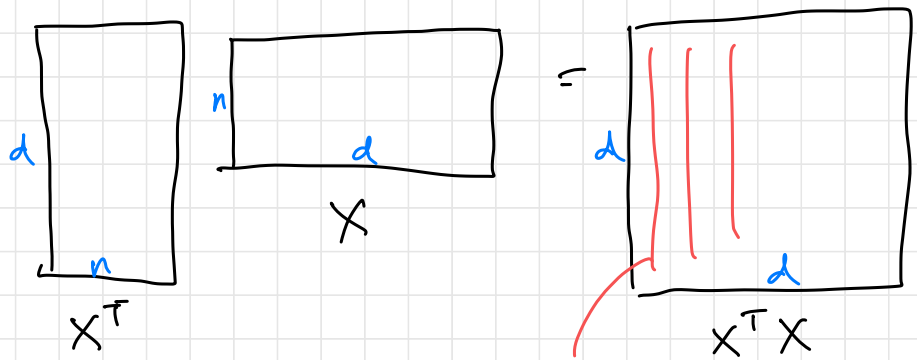
closed form solution!

Question: What if X^TX is not invertible?

Scenario 1: $n < d$ (we have fewer examples than features)

train examples

dimension of x's = # of features



each column is linear combination of columns of X^T

every column of X^TX lies in $\leq n$-dim subspace of $\mathbb{R}^d$

$\text{rank}(X^TX) \leq n < d$

$\Rightarrow (X^TX)^{-1}$ does not exist!

Implication: $X^TXw = X^Ty$ has many solutions

with many features & few examples, easy to find many w to get 0 train loss

we have high variance → Based on training data many equally good w's

→ Some will be better than others on test data

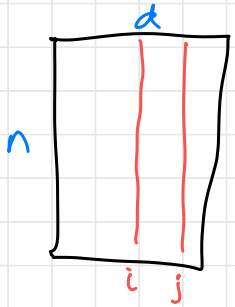
Common rule of thumb: "Always" have more examples than features

In practice, use pseudo-inverse A^+

- $A^+ = A^{-1}$ when A^{-1} exists
- For $Ax = b$, $x = A^+ b$ will be a solution

So: Implement $w = (X^T X)^+ X^T y$

Scenario 2: Duplicated features



Suppose columns i & j are equal

Then $X^T X$ not invertible

Intuitively: also high variance

$w = [w_1 \dots w_i, \dots w_j \dots]$

$+100$ -100

-500 $+500$

many equally good w's hard to find best one

Rule of thumb: Avoid highly correlated / duplicate features