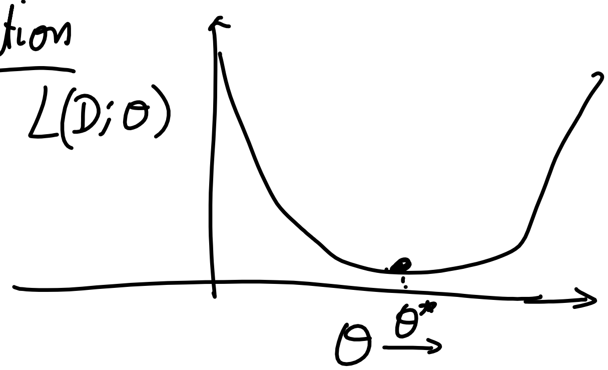
m

Probability theory and machine learning basics

Machine learning as optimization



$$D = \underbrace{\{(\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)\}}_{\text{Training data}}$$

Choose a model

$\hat{y} = f(\underline{x};\theta) =$ either a linear model
or a MLP

$\ell(\hat{y}, y) =$ error or loss for predicting $\hat{y}$ when true value is y

$$L(D;\theta) = \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) = \sum_{i=1}^{n} \ell(f(\underline{x}_i;\theta), y_i)$$

$$\theta^* = \arg\min_{\theta} \underbrace{L(D;\theta)}_{\text{Training loss}}$$

① Classify handwritten digits
↳ 6000 Handwritten images

What matters is the loss/performance on previously unseen images

Data { Training data ✓ (only optimize on training data)
Data { Test data (proxy for real world performance)

Supervised learning

Input $\underline{x}$ $\xrightarrow{f(x;\theta)}$ Label y.

Prediction $\hat{y}$

Example 2D features of each handwritten image

Digit corresponds to the ma

$f(x;\theta)$ here is a a model.

Probability is a measure/quantification of uncertainty

Sources of uncertainty in machine learning

eg Handwritten digits

$$\underbrace{\boxed{0}\ \boxed{0}\ \boxed{0}\ \dots\ \boxed{1}\ \boxed{1}\ \boxed{1}}_{\text{Data}}\ \dots$$

We assume that all data is generated by some data-generating process

↳ Incomplete observability

↳ Incomplete modeling

$\boxed{0}$ ~ not taking factors into account.

↳ Inherent uncertainty in the physical process

Probability theory

Def 1 <u>Sample Space</u>

Eg 2-coin tosses. H, T

$$\underbrace{\{HH, HT, TH, TT\}}_{}$$

Sample Space $= \Omega$ is the set of all possible outcomes

Eg Roll of a dice with 6-sides

$\Omega = \{1, 2, \dots 6\}$

<u>Def 2</u> <u>Event Space</u> : is the space of potential results of the experiment

Eg② Roll of dice

Event: whether we got dice ≥ 5

$A : \{d \geq 5\} = \{5, 6\}$

$A \subseteq \Omega$

<u>Event space</u> is all possible value of such Events

Eg① 2 coin toss $A = \{HH\} \subseteq \Omega$

$A =$ contains 1 tail $= \{HT, TT, TH\} \subseteq \Omega$

For <u>discrete sample space</u>

Event space is the power set of sample space

Event space $= \mathcal{A} = \mathbb{P}(\Omega) = 2^{\Omega}$ ← power set

size of event space $= 2^{|\Omega|}$

<u>Def 3</u> <u>Probability measure</u> $P(A) \in [0,1]$ ← Event

$A \subseteq \Omega$

$A \in \mathcal{A}$ ← Event space

<u>Def 4</u> $\underbrace{\text{(Sample space, Event Space, Probability measure)}}_{\text{Probability space}}$

<u>Def 3</u> Probability measure is consistent if

Kolmogrov's axioms

(a) $P(A) \in [0,1]$

(b) $P(\Omega) = 1$

(c) $A_1 \cap A_2 = \emptyset$

then $P(A_1) + P(A_2) = P(A_1 \cup A_2)$

e.g dice

$A_1 = \{5,6\}$

$A_2 = \{4,3\}$

$A_1 \cup A_2 = \{4,3,5,6\}$

countable infinite sets

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \quad \text{when } A_{n_1} \cap A_{n_2} = \emptyset$$

countable infinity – If you can assign a natural number to all the elements

e.g. Natural numbers $(1, \dots \infty)$

uncountable ∞ – Real numbers

Def 5 Random variable (RV)

e.g 2 coin toss $\Omega = \{\overset{0}{HH}, \overset{1}{HT}, \overset{2}{TH}, \overset{3}{TT}\}$

RV maps an $\underbrace{\text{event}}_{A \subseteq \Omega}$ to a set of **numbers** → Discrete
→ continuous

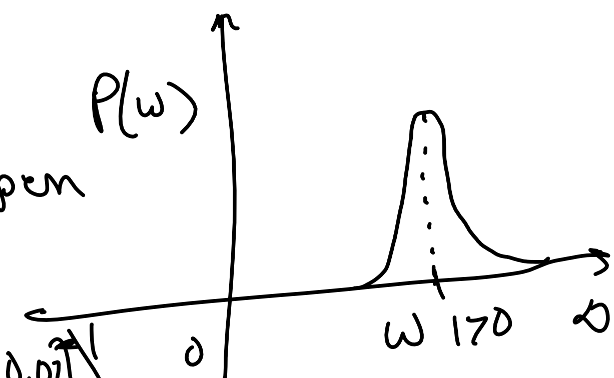
$$X(\{HT, TH, TT\}) = \{1, 2, 3\}$$

e.g. continuous RV : weight of mice

$$\Omega = [0, \infty)$$

closed interval open

$A \in [170, 171]$



Event space of cont. RV $P(w = [170.01, 170.02])$

A = the set of all unions and intersections of countably infinite intervals (open, closed) } Borel set

→ Different from the power set of all Real numbers

① Probability Mass function (PMF)

is only defined for discrete RV

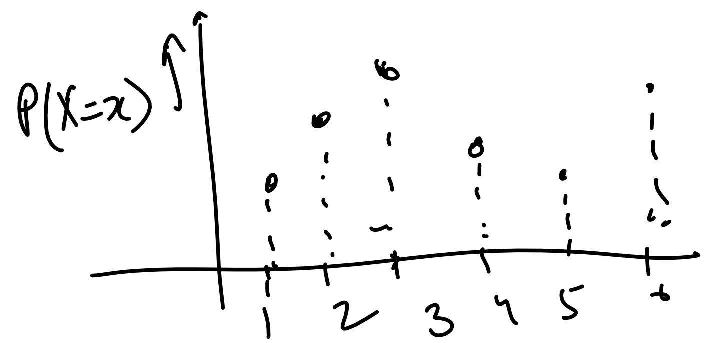
$\Omega = \{1, \dots, 6\}$

$P(X=1) = p_1$

$P(x=2) = p_2$

$\underbrace{P(X=x) = p_x}_{\text{PMF}} \quad : \quad \sum_{x \in \Omega} P(X=x) = 1$

$\boxed{\underbrace{P_X(x)}_{}}$ ↑RV, the value



$\boxed{P(x)}$ Notational short cuts

Probability Density function (PDF)

only defined for continuous RV

$P(X=x) = 0$ if $x \in \mathbb{R}$ X is continuous RV

$$P(a \le X \le b) = P(X \in [a,b]) = \int_a^b \underbrace{f(x)}_{\text{PDF}} \, dx$$

A function $f : \mathbb{R} \to [0, \infty)$ is called PDF if

a) $f(x) \geq 0 \quad \forall \; x \in \mathbb{R}$

b) $\int_{\mathbb{R}} f(x)\,dx = 1$

c) $P(a \leq X \leq b) = \underbrace{\int_a^b f(x)\,dx}$

$$\int_{170.1}^{170.2} 10\,dx = 10(170.2 - 170.1) = 1$$

Common continuous distribution

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

$x = \mu \qquad \exp(0) = 1$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \qquad \sigma < \frac{1}{\sqrt{2\pi}}$$

$$= 2 \qquad \sigma = \frac{1}{2\sqrt{2\pi}}$$

<u>Cumulative density function</u> — Discrete / Continuous

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

$$F_X(x) = P(X \le x) = \sum_{z \le x} \underbrace{P(X = z)}_{\text{PMF}}$$ For discrete RV

Multi variable Probability disribution

Joint probability mass function

$$P(X = x, Y = y) = P((X = x) \cap (Y = y))$$

e.g. 2 coin tosse = X = $\{ \overset{0}{HH}, \overset{1}{HT}, \overset{2}{TH}, \overset{3}{TT} \}$

1 dice roll = Y

$$P(X = 1, Y = 5) = P(X = 1 \ \underline{\underline{\text{and}}} \ Y = 5)$$

$$\underbrace{P\left(\underline{X} = \begin{bmatrix} x \\ y \end{bmatrix}\right)}_{\text{PMF}} = P(X_1 = x \ \underline{\underline{\text{and}}} \ X_2 = y)$$

X_1: first element of Random vector

Random variable

Random vectors $\underline{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

RV

PDF = probability distrubution function

$$P(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b \underbrace{f(x, y)}_{\text{PDF of 2 variable}} \, dx \, dy$$

vectors $\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \underline{x_{1:n/2}} \\ \underline{x_{n/2:n}} \end{bmatrix}$

$\{1, \ldots 6\}$

CDF

↳ Discrete $F_{XY}(x,y) = P(X \le x, Y \le y) = \sum_{a \le x} \sum_{b \le y} P(X = a, Y = b)$

↳ Continuous $F_{XY}(x,y) = P(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,y)\, dy\, dx$

Conditional Probability

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Prob of X given Y

Dec 2022 65-71

Vacc (V) \ Death (D)	No (0)	Yes (1)
No (0)	1.2M	150
Yes (1)	10M	213

$$P(D=1 \mid V=1) = ?$$

$$= \frac{P(D=1, V=1)}{P(V=1)} = \frac{213}{10M + 213}$$

$$P(V=1 \mid D=1) = ? = \frac{P(D=1, V=1)}{P(D=1)}$$

$$= \frac{213}{363}$$

Marginalization

$$\boxed{P(V=1) = P(D=1, V=1) + P(D=0, V=1)}$$

$$P(D=1) = P(D=1, V=0) + P(D=1, V=1)$$

$$P(X) = \sum_{y \in \Omega_Y} P(X, Y = y)$$ } Discrete RV

$$P(X) = \int_{y \in \Omega_Y} f(X, Y = y)\, dy$$ } Continuous RV

$$P(X, Y) = P(Y|X)\,P(X)$$
$$P(X, Y) = P(X|Y)\,P(Y)$$
$$\Rightarrow P(Y|X)\,P(X) = P(X|Y)\,P(Y)$$

$$\boxed{P(Y|X) = \frac{P(X|Y)\,P(X)}{P(Y)}}$$

Bayes Theorem

$\frac{\text{Bloodtest}}{\text{Symptoms}}$ | Disease or Dataset | Parameters

D θ

$$\underbrace{P(D|S)}_{\text{Posterior}} = \frac{\overbrace{P(S|D)}^{\text{Likelihood}}\,\underbrace{P(D)}_{\text{Prior}}}{\underbrace{P(S)}_{\text{Evidence}}}$$

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}}\,\overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(D)}_{\text{Evidence}}}$$

$$P(D|S) \propto P(S|D)\,P(D)$$

$$S \approx f(D) \Rightarrow D \approx f^{-1}(S)$$

Probabilistic

Statistical Independence

$X \perp\!\!\!\perp Y$ are independent iff $P(X, Y) = P(X)P(Y)$

$$P(Y|X) = P(Y)$$

$$P(X|Y) = P(X)$$

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2) \ldots P(X_n)$$

Identical Random Variables

Identically distributed RV

X_1 is RV with $(\Omega, \mathcal{F}, P_{X_1})$ as the Probability space

Probability measure

X_2 is RV with $(\Omega, \mathcal{F}, P_{X_2})$ as the Prob. space

Sample space

Event space

$$P_{X_1} \equiv P_{X_2}$$

Def IID: Identically Independentally distributed

$X_1, X_2 \ldots X_n$ are RV

1. Independent
2. Identically distributed

Example

$$D = \{\underbrace{(x_1, y_1)}_{z_1}, \underbrace{(x_2, y_2)}_{z_2} \ldots \underbrace{(x_n, y_n)}_{z_n}\}$$

$$z_1 \perp\!\!\!\perp z_2 \qquad z_1 \perp\!\!\!\perp z_n$$

$$\underbrace{P(D|\theta)}_{\text{likelihood}} = P(z_1|\theta)\, P(z_2|\theta) \ldots P(z_n|\theta)$$

$$\overbrace{P(X, Y|Z) = P(X|Z)\, P(Y|Z)}^{\text{Conditional independence}}$$

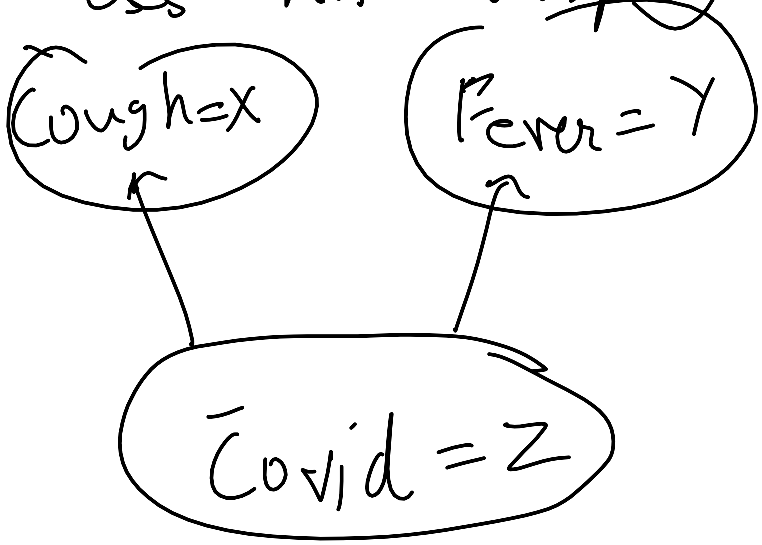
$X \perp\!\!\!\perp Y$ given Z

$\Downarrow$

Does not imply

$$P(X\,Y) = P(X)P(Y)$$

$$X \perp\!\!\!\perp Y$$



$P(Y|X) = \text{high}$ does not imply X causes Y

Can think of $P(Y|x)$ as non-linear correlation
x

Expectation

$$\mathbb{E}_X[g(X)] = \sum_{x \in \Omega_X} P(X=x)\, g(x)$$

Discrete RV

$$\mathbb{E}_X[g(X)] = \int_{x \in \Omega_x} f_X(x)\, g(x)\, dx$$

Cont RV

Sample mean

$X_1, X_2 \ldots X_n$

$$\mu(x_1, \dots x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_n = \mathbb{E}_X[X]$$

Example Dice

Sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$

Event $F = 2^{\Omega}$, $P(X = x) = \frac{1}{6}$

Expectation $\mathbb{E}_X[X] = 3.5$

$$\mathbb{E}_X[X] = \sum_{x \in \Omega} P(X = x)\, x$$

Sample mean

$$\mu(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Event space $= \{ \{\},$
$\{1\}, \{2\},$
$\{1, 2\}, \{2, 3\} \}$

$\}$

$$\boxed{\mathcal{F} = 2^{\Omega}}$$

Probability measure

$$P_X(X = x) = \frac{1}{6}$$

$$\mathbb{E}_X[X] = \sum_{x \in \Omega} P(X = x)\, x = \sum_{x \in \Omega} \frac{1}{6} x$$

Variance

$$\mathbb{V}_x[g(x)] = \mathbb{E}_x\left[\left(g(x) - \underbrace{\mathbb{E}_x[g(x)]}_{\text{expectation}}\right)^2\right]$$

Vector Variance

Covariance Matrix

$$\mathbb{V}_{\underline{x}}[\underline{x}] = \mathbb{E}_x\left[\underbrace{\left(\underline{x} - \underbrace{\mathbb{E}_x[\underline{x}]}_{\text{mean}}\right)\left(\underline{x} - \mathbb{E}_x[\underline{x}]\right)^T}_{\text{outer product}}\right]$$

Machine learning as optimization

$D = \{(\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)\}$ Dataset

$\hat{y}_1 = f(\underline{x}_1; \theta)$ model

$l(y_1, \hat{y}_1)$ loss function

$$\theta^* = \arg\min_{\theta} \underbrace{\sum_{i=1}^{n}}_{\text{sum over the dataset}} l(\underset{\substack{\uparrow \\ \text{Training label}}}{y_i}, \underbrace{f(\underline{x}_i;\theta)}_{\text{predicted label}})$$

$$P(y_i \mid \underline{x}_i, \theta) = \frac{1}{Z} \exp\left(-l(y_i, f(\underline{x}_i;\theta))\right)$$

Z ← normalization factor

$$l(y_i, f(\underline{x}_i, \theta)) = -Z \log P(y_i \mid \underline{x}_i, \theta)$$

$$P(D \mid \theta) = \prod_{i=1}^{n} P((\underline{x}_i, y_i) \mid \theta)$$

Assumption

$(\underline{x}_i, y_i) \perp\!\!\!\perp (\underline{x}_j, y_j)$

when $i \neq j$

Take log on both sides

$$\log P(D \mid \theta) = \sum_{i=1}^{n} \log P((\underline{x}_i, y_i) \mid \theta)$$

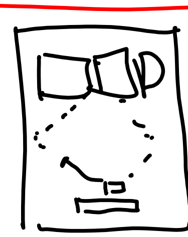
$\log(xy)$
$= \log x + \log y$

Take away 1

Check if your data samples are IID

$$\underbrace{P(\theta \mid D)}_{\text{Posterior}} = \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}} \, \overbrace{P(\theta)}^{\text{Prior}}}{P(D)}$$

Not IID example



Images $\underline{x}_i$ → Button press

labels y_i

$(\underline{x}_t, y_t) \curvearrowright (\underline{x}_{t+1}, y_{t+1})$

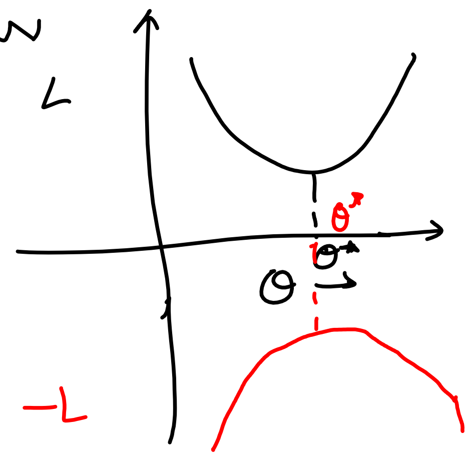
Observation Minimizing loss function
$\equiv$ Maximizing the likelihood

$$L(D;\theta) = \sum_{i=1}^{n} \ell\left(y_i, f(\underline{x}_i, \theta)\right)$$

$$\theta^* = \arg\min_{\theta} L(D;\theta)$$

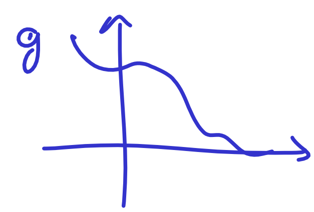
(Take negative sign of the loss function
argmin changes to arg max

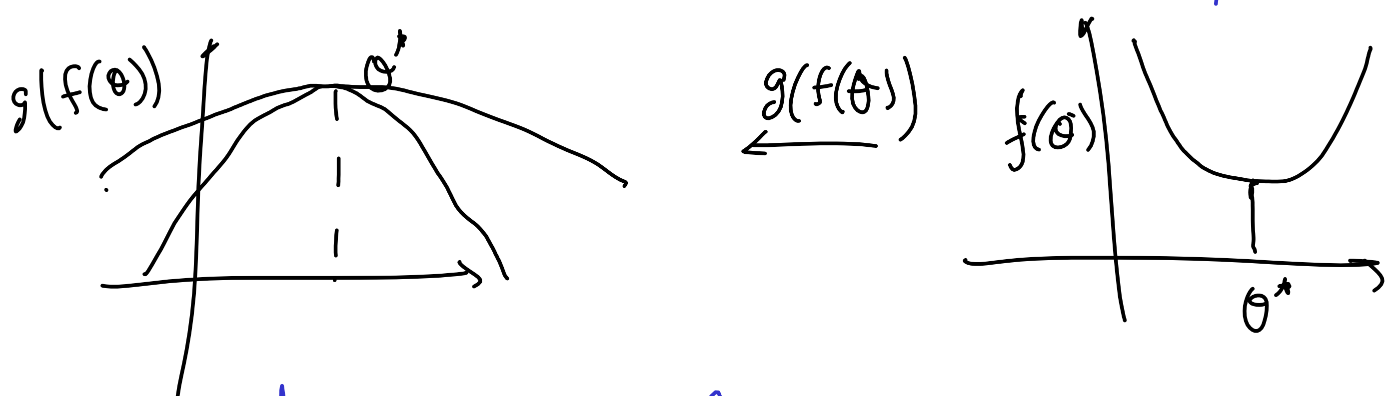
$$\theta^* = \arg\max_{\theta} \left\{-L(D;\theta)\right\}$$



Monotonically decreasing function
$g(f(x)) := -f(x)$

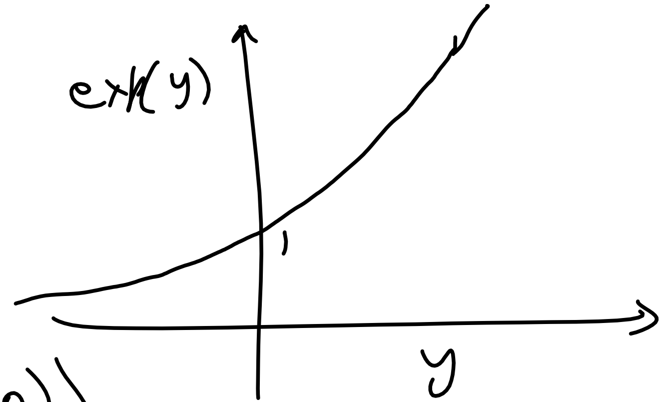
$g(y_2) < g(y_1)$ if $y_2 > y_1$



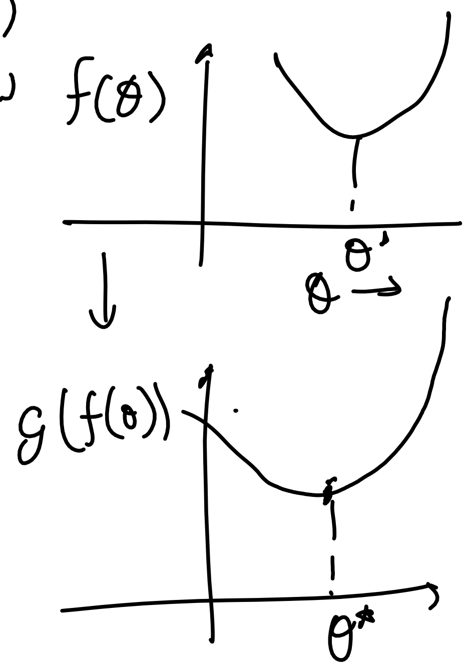


Monotonically increasing function $\exp(y)$

$$\theta^* = \arg\max_{\theta} \; -L(D;\theta)$$



$$\theta^* = \arg\max_{\theta} \; \underbrace{\exp(-L(D;\theta))}$$



$$\theta^* = \underbrace{\arg\max_{\theta} \; P(D|\theta)}_{\text{estimation}}$$

is called **Maximum Likelihood estimation** (MLE)

Maximum a-POSTERIORI estimation (MAP)

$$\theta^* = \arg\max_{\theta} \; \underbrace{P(\theta|D)}_{\text{Posterior}}$$

$$= \arg\max_{\theta} \; \underbrace{P(D|\theta)}_{\text{Likelihood}} \; \underbrace{P(\theta)}_{\text{Prior}}$$

By Bayes theorem

$$\theta^* = \arg\min_{\theta} \; \underbrace{\{-\log P(D|\theta)\}} + \underbrace{\{-\log P(\theta)\}}$$

$\log(xy) = \log x + \log y$

$$= \arg\min_{\theta} \; L(D;\theta) + \underbrace{\lambda R(\theta)}$$ ← Regularizer

Examples of Regularizers

① $R(\underline{\theta}) = \underbrace{\|\underline{\theta}\|_2^2}$ L-2 norm
L-2 regularizer

② $R(\underline{\theta}) = |\underline{\theta}|_1$ L-1 norm
L-1 regulariyer

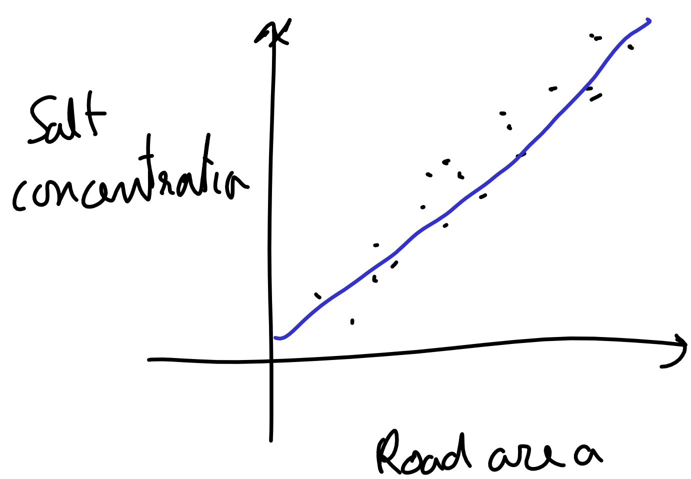
L-2 norm $\underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix}$; $\|\theta\|_2 = \left(\theta_1^2 + \theta_2^2 + \dots \theta_m^2\right)^{1/2}$

L-1 norm $\underline{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$; $\|\theta\|_1 = \left(|\theta_1| + |\theta_2| + \dots |\theta_m|\right)$

$\boxed{\text{L-p norm}}$ $\underline{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$; $\|\theta\|_p = \left(|\theta_1|^p + |\theta_2|^p + \dots |\theta_m|^p\right)^{1/p}$

Example Least square regression

$\underline{m} = \begin{bmatrix} m \\ c \end{bmatrix}$ $\underline{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$ $\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$



$L(D : \underline{m}) = \|\underline{y} - \underline{X}\,\underline{m}\|_2^2$

Parameters

$R(\underline{m}) = \|\underline{m}\|_2^2$

Regularized Least square regression

$$\underline{m}^* = \underset{\underline{m}}{\arg\min} \; \overbrace{\|\underline{y} - \underline{X}\,\underline{m}\|_2^2 + \lambda \|m\|_2^2}^{f(\underline{m})}$$

λ ← some positive scalar

$$\underline{m}^* = (\underline{X}^T \underline{X})^{-1} X^T \underline{y}$$ Not regularized

$$f(\underline{m}) = (\underline{y} - X\underline{m})^T (\underline{y} - X\underline{m}) + \lambda \underline{m}^T \underline{m}$$

$$= m^T (\underline{X}^T \underline{X} + \lambda \underline{I}) \underline{m} - 2 y^T X \underline{m} + \underline{y}^T y$$

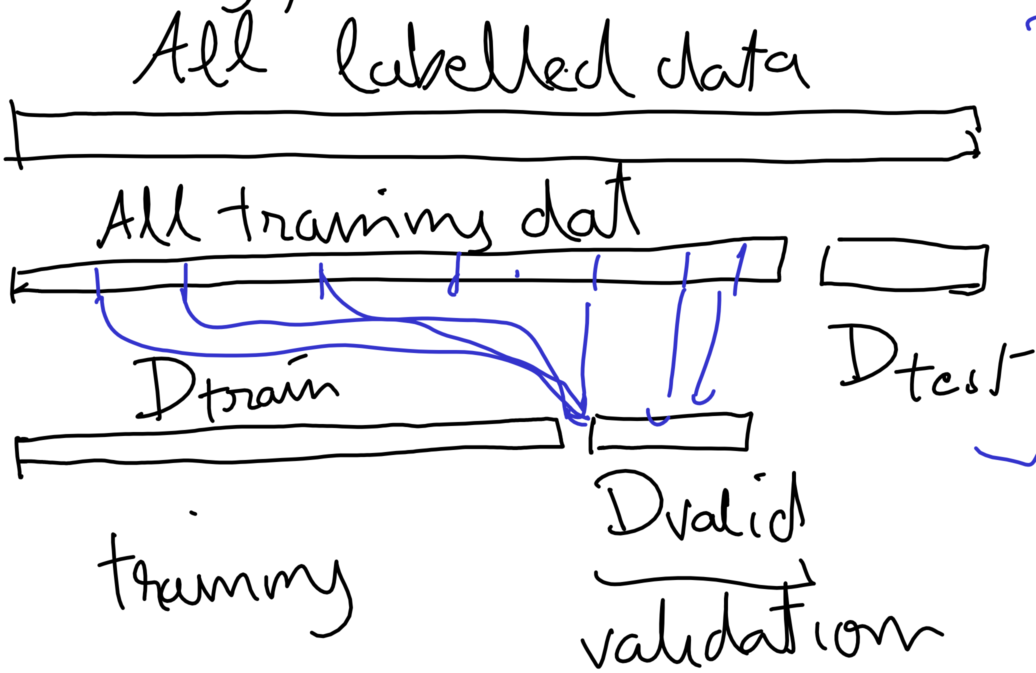
$$\frac{\partial f}{\partial \underline{m}} = 0$$

$$\Rightarrow \underline{m} = (\underline{X}^T \underline{X} + \lambda I)^{-1} \underline{X}^T y$$ Regularized Least square solution

Take away 2

Regularizers in regularized regression can be interpreted as Priors in Bayes Theorem

Training, test and validation split



Always split uniformly random distribution

$$L_{train}(D;\theta) = \sum_{(x,y)\in D_{train}} \ell(y; f(x;\theta))$$

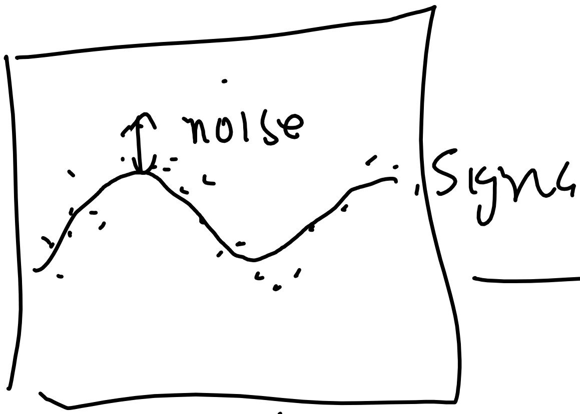
What assumption do we need to make sure

$$L_{train}(D_{train};\theta) \approx L_{test}(D_{test};\theta)$$

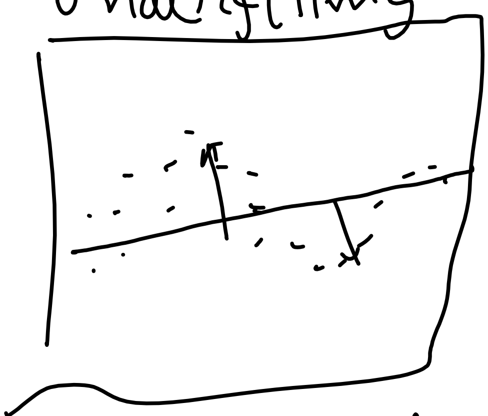
$$D_{train} = \{(\underline{x}_1^{tr}, y_1^{tr}) \ldots (\underline{x}_n^{tr}, y_n^{tr})\}$$

$$P((x_i^{tr}, y_i^{tr}) \mid \theta) = P((x_j^{te}, y_j^{te}) \mid \theta)$$

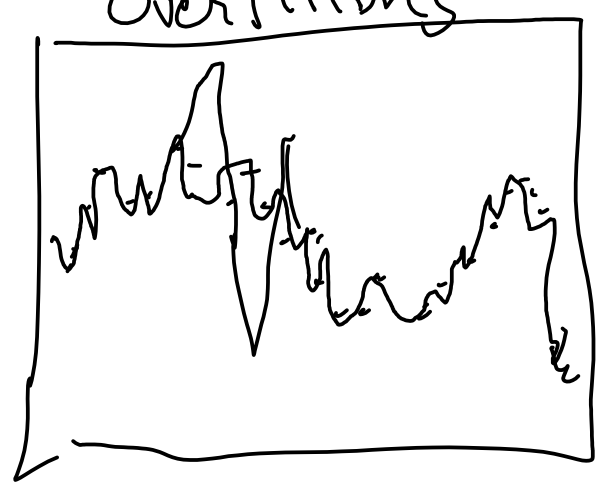
$$D_{test} = \{ (x_1^e, y_1^e) \dots (x_n^e, y_n^e) \}$$



Underfitting



Overfitting



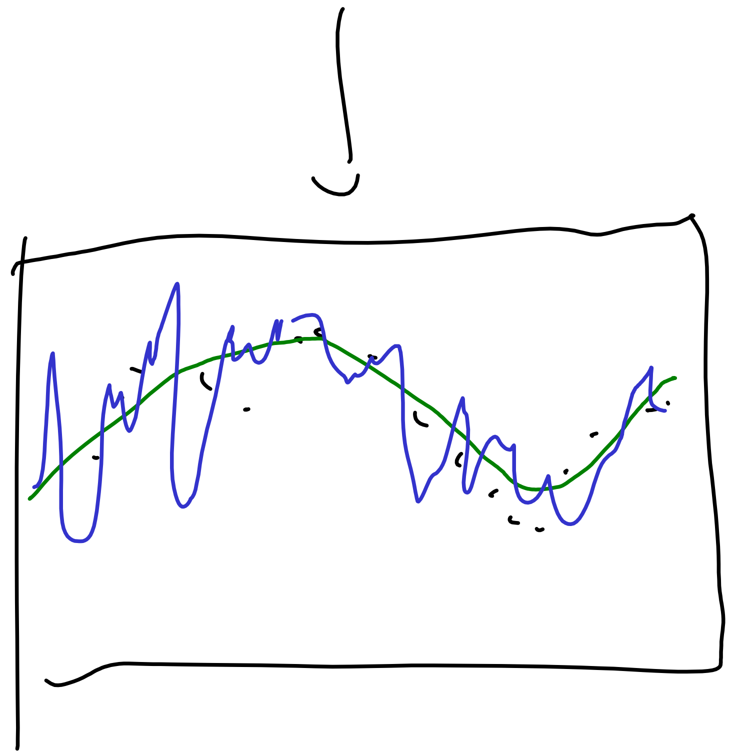
Linear model

The model is not flexible enough to fit the data

fourier series

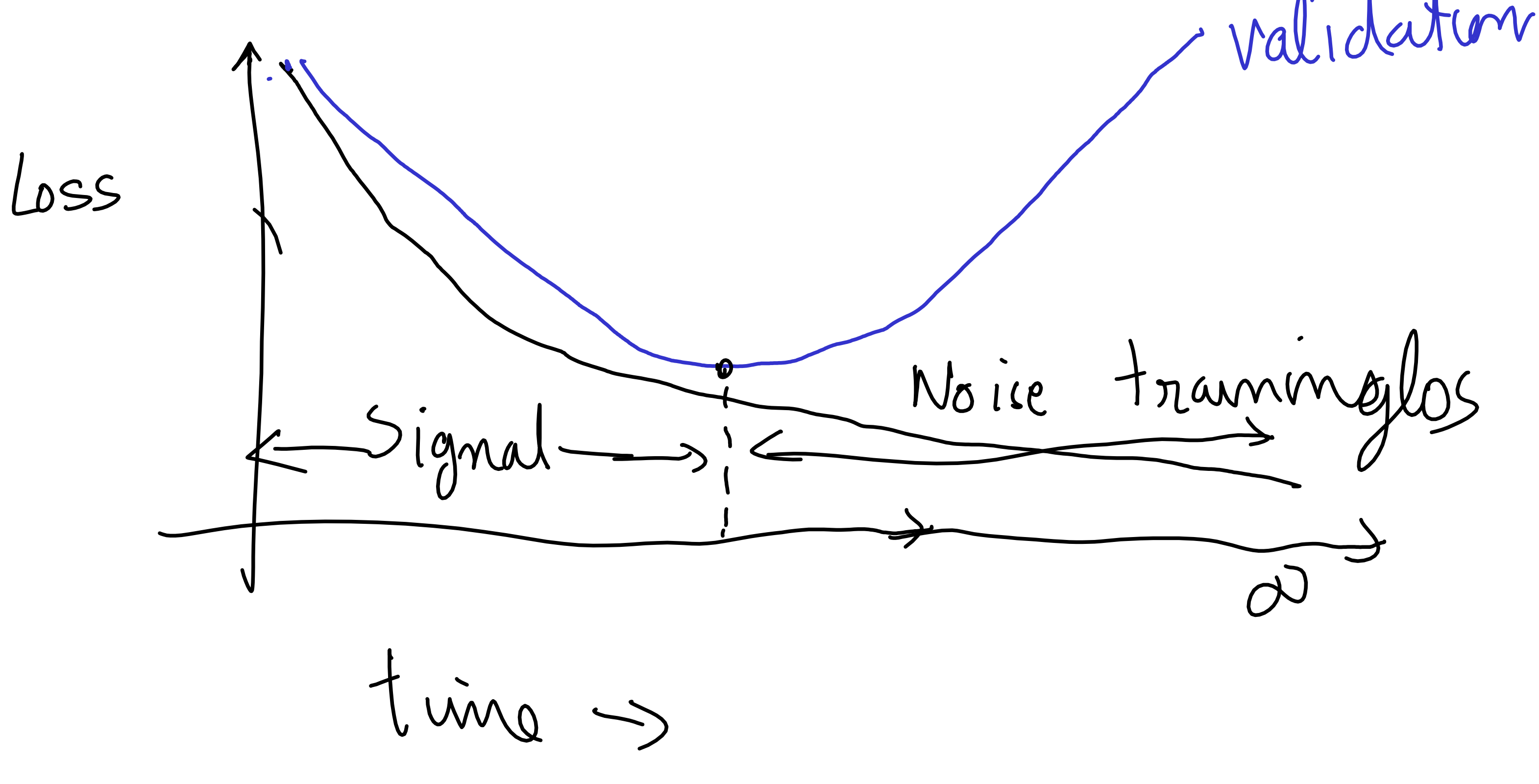
100 frequency

The model is too flexible



① Neural networks typically are <u>too flexible</u>

mostly prone to overfitting

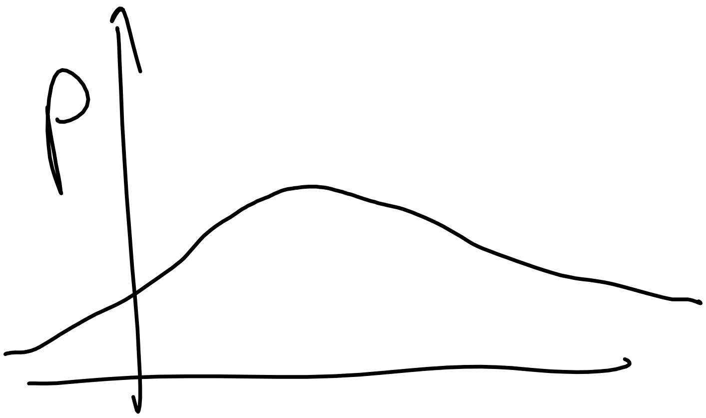


$D_{train} \sim P$

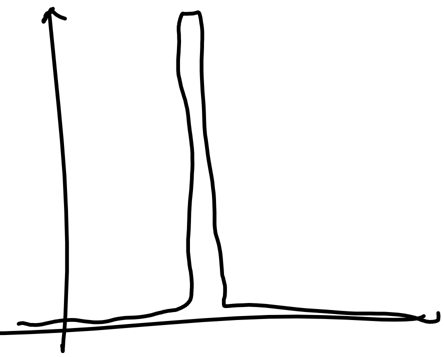
$D_{valid} \sim P$

from sample

P



$P =$

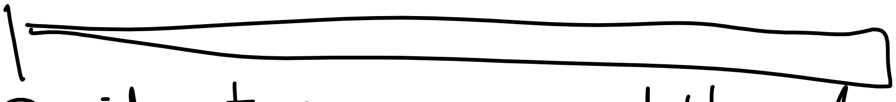


By targetting minimum validation loss, we stop at a point where noise cannot be further eliminated

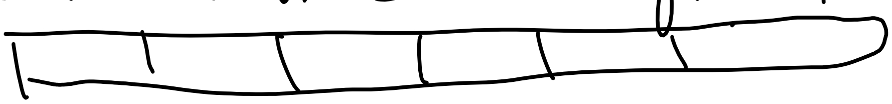
---

Cross validation (For smaller datasets)

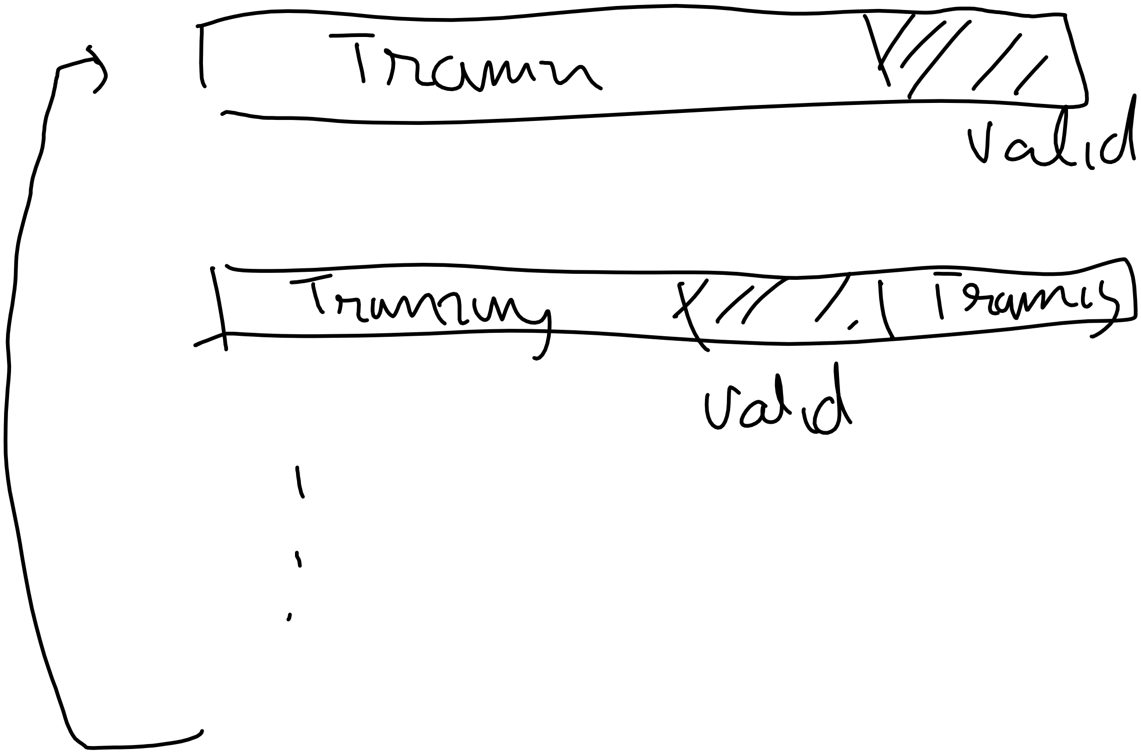
All training dat



Divide it in 5 different parts



4/5 for Training and 1 for validatu

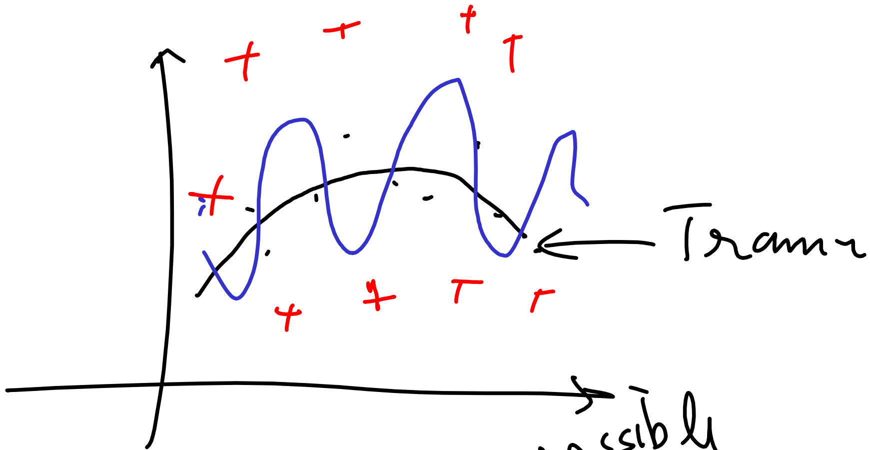


Why is it so hard to choose a model?

- Is overfitting always going to happen
- Can we just choose the biggest NN for all problems?
- Is there a "correct" model for all problems?

The answer is <u>NO</u>

NO FREE LUNCH THEOREM



Averaged over all possible datasets

all models have the same accuracy