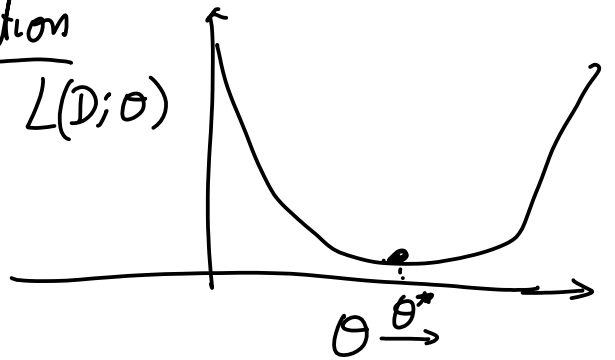


Machine learning as optimization

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

Training data



Choose a model

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

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

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

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

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)  
Test data (proxy for real world performance)

Supervised learning

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

Prediction  $\hat{y}$

Example

2D features  
of each handwritten  
image

Digit  
corresponds  
to the max

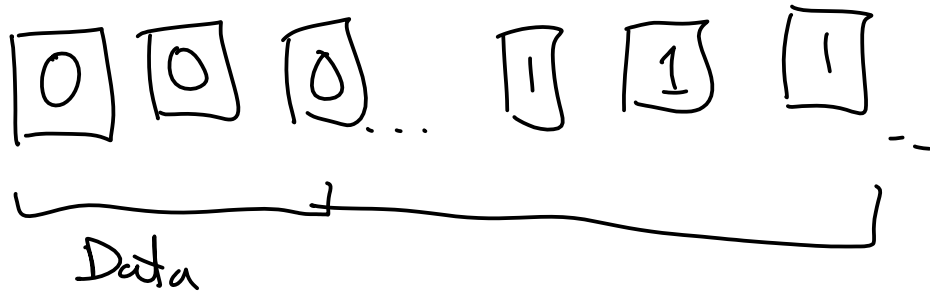
$f(x;\theta)$  here is a model.

---

Probability is a measure/quantification of uncertainty

Sources of uncertainty in machine learning


eg  
Handwritten  
digits

  
Data

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

↳ Incomplete observability

↳ Incomplete modeling

  $\sim$  not taking factors into account

↳ Inherent uncertainty in the physical process


---

Probability theory

Def 1 Sample Space

Eg 2-coin tosses . H, T

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

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

Eg Roll of a die with 6-sides

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

Def 2 Event Space : is the space of potential results of the experiment

Eg Roll of dice

Event: whether we got dice  $\geq 5$

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

$$A \subseteq \Omega$$

Event Space is all possible values of such Events

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

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

For discrete sample space

Event space is the powerset of sample space

$$\text{Event Space} = \mathcal{A} = \mathcal{P}(\Omega) = 2^{\Omega}$$

↑ power set

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

Def 3 Probability measure

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

↑ Event

$$A \subseteq \Omega$$

$$A \in \mathcal{A}$$

↑ Event set

Def 4 (Sample space, Event Space, Probability measure)

Probability space

Def 3 Probability measure is consistent if

Kolmogorov's axioms

- (a)  $P(A) \in [0, 1]$
- (b)  $P(\Omega) = 1$
- (c)  $A_1 \cap A_2 = \emptyset$

e.g. dice

$A_1 = \{5, 6\}$

$A_2 = \{4, 3\}$

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

countable infinite sets

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

$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$  when  $A_n \cap A_{n2} = \emptyset$

countable infinity — If you can assign a natural number to all

the elements

e.g. Natural numbers  $\{1, \dots, \infty\}$

uncountable  $\infty$  — Real numbers

Def 5 Random variable (RV)

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

RV maps an event to a set of numbers  $\left\{ \begin{array}{l} \text{Discrete} \\ \text{Continuous} \end{array} \right.$

$$A \subseteq \Omega$$

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

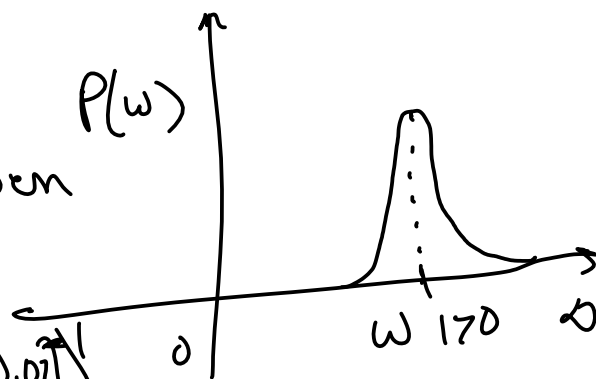
e.g. continuous RV : weight of mic

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

closed interval

open

$$A \in [170, 171]$$



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

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

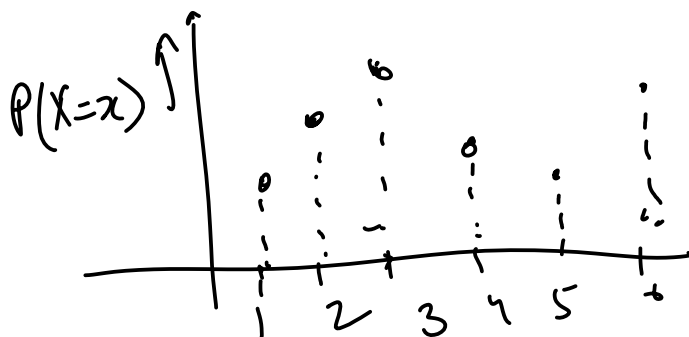
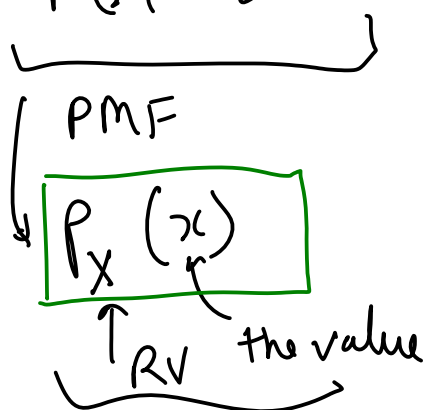
→ Different from the power set of all Real numbers

① Probability Mass function (PMF)  
is only defined for discrete RV  
 $\Omega = \{1, \dots, 6\}$

$$P(X=1) = p_1$$

$$P(X=2) = p_2$$

$$P(X=x) = p_x \quad \quad \quad \sum_{x \in \Omega} P(X=x) = 1$$



$P(x)$  Notational short cuts

Probability Density function (PDF)  
only defined for continuous RV

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

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

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

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

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

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

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

Common continuous distribution

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

$$x = \mu$$

$$\exp(0) = 1$$

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

$$= 2$$

$$\sigma < \frac{1}{\sqrt{2\pi}}$$

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

Cumulative density function

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

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

Discrete

Continuous



$$F_X(x) = P(X \leq x) = \sum_{z \leq x} \underbrace{P(X=z)}_{\text{PMF}} \quad \left. \vphantom{\sum} \right\} \text{For discrete RV}$$

Multi variable probability distribution

Joint probability mass function

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

e.g. 2 coin tosses =  $X = \{H^0H^1, H^1T^0, T^0H^1, T^1H^1\}$   
 1 dice roll =  $Y$

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

$$P(\underline{X} = \begin{bmatrix} x \\ y \end{bmatrix}) = P(\underbrace{X_1 = x}_{\text{first element of Random vector}} \text{ and } \underbrace{X_2 = y}_{\text{Random variable}})$$

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

PDF = probability distribution function

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

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



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

CDF

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

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

Conditional Probability

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

Prob of X given Y

Dec 2022 65-71

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

$$P(D=1|V=1) = ?$$

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

$$P(V=1|D=1) = ? = \frac{P(D=1, V=1)}{P(D=1)} = \frac{213}{363}$$

Marginalization

$$\left[ \begin{aligned} P(V=1) &= P(D=1, V=1) + P(D=0, V=1) \\ P(D=1) &= P(D=1, V=0) + P(D=1, V=1) \end{aligned} \right]$$

$$P(X) = \sum_{y \in \Omega_y} P(X, Y=y)$$

} Discrete RV

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

} Continuous RV

$$\begin{aligned} P(X, Y) &= P(Y|X) P(X) \\ P(X, Y) &= P(X|Y) P(Y) \end{aligned} \Rightarrow P(Y|X) P(X) = P(X|Y) P(Y)$$

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

Bayes Theorem

Bloodtest

Symptoms

Disease

Likelihood on

$$P(D|S) = \frac{P(S|D) P(D)}{P(S)}$$

Posterior (under  $P(D|S)$ )

Evidence (under  $P(S)$ )

Prior (under  $P(D)$ )

Dataset

$D$

Parameters

Likelihood Prior

$$P(\theta|D) = \frac{P(D|\theta) P(\theta)}{P(D)}$$

Posterior (under  $P(\theta|D)$ )

Evidence (under  $P(D)$ )

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

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

## Probability Statistical Independence

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

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

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

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

## Identical Random Variables

Identically distributed RV Probability measure

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

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

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

Sample space  
Event space

Def IID: Identically Independently distributed

$X_1, X_2, \dots, X_n$  are RV

↳ (1) Independent

(2) Identically distributed

Example

$$D = \{ (\underbrace{x_1, y_1}_{Z_1}), (\underbrace{x_2, y_2}_{Z_2}) \dots (\underbrace{x_n, y_n}_{Z_n}) \}$$

$$Z_1 \perp\!\!\!\perp Z_2$$

$$Z_1 \perp\!\!\!\perp Z_n$$

$$P(D|\theta) = P(z_1|\theta) P(z_2|\theta) \dots P(z_n|\theta)$$

likelihood

Conditional independence



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

$$X \perp\!\!\!\perp Y \text{ given } Z$$



Does not imply

Cough = X

Fever = Y

Covid = Z

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

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

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

Can think of  $\rho(Y|X)$  as non-linear correlation  
 $\rho$

### Expectation

Discrete RV  $\nearrow$

$$\mathbb{E}_x[g(x)] = \sum_{x \in \Omega_x} P(X=x) g(x)$$

Cont RV  $\nearrow$

$$\mathbb{E}_x[g(x)] = \int_{x \in \Omega_x} f_x(x) g(x) dx$$

### Sample mean

$X_1, X_2, \dots, X_n$

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


---

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_n = E_x[X]$$

Example Die

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

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

Expectation  $E_x[X] = 3.5$

---

$$E_x[X] = \sum_{x \in \Omega} P(X=x) x$$

Sample mean

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

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

$$\text{Event space} = \left\{ \emptyset, \right. \\ \left. \{1\}, \{2\}, \right. \\ \left. \{1, 2\}, \{2, 3\} \right\}$$

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

Probability measure

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

$$E_X[X] = \sum_{x \in \Omega} P(X=x) x = \sum_{x \in \Omega} \frac{1}{6} x$$



# Variance

$$V_x[g(x)] = E_x \left[ \left( g(x) - \underbrace{E_x[g(x)]}_{\text{expectation}} \right)^2 \right]$$

## Vector Variance Covariance Matrix

$$V_{\underline{x}}[\underline{x}] = E_x \left[ \underbrace{\left( \underline{x} - \underbrace{E_x[\underline{x}]}_{\text{mean}} \right) \left( \underline{x} - E_x[\underline{x}] \right)^T}_{\text{outer product}} \right]$$

## Machine learning as optimisation

$D = \{ (x_1, y_1) \dots (x_n, y_n) \}$  Dataset

$\hat{y}_i = f(x_i; \theta)$  model

$l(y_i, \hat{y}_i)$  loss function

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

$$P(y_i | \underline{x}_i, \theta) = \frac{1}{\underbrace{Z}_{\substack{\uparrow \\ \text{normalization factor}}}} \exp(-l(y_i, f(\underline{x}_i; \theta)))$$

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