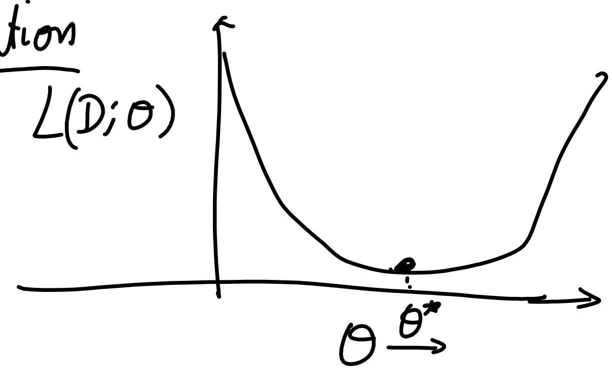
Probability theory and machine learning basics

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

$l(\hat{y}, y) =$ 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 $\begin{cases} \text{Training data ✓ (only optimize on training data)} \\ \text{Test data (proxy for real world performance)} \end{cases}$

Probability is a measure/quantification of uncertainty

Sources of uncertainty in machine learning

eg
Handwritten digits $\underbrace{\boxed{0}\ \boxed{0}\ \boxed{0} \ldots \boxed{1}\ \boxed{1}\ \boxed{1}}_{\text{Data}}$ --

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

<u>Def 1</u> <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

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

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

Eg② Roll of dice

Event: whether we got dice ≥ 5

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

$A \subseteq \Omega$

Event space 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 discrete sample space

Event space is the powerset of sample space

Event space $= \mathcal{A} = \underbrace{\mathbb{P}(\Omega)}_{\text{power set}} = 2^{\Omega}$ $\quad$ size of event space $= 2^{|\Omega|}$

Def 3 Probability measure $P(\underbrace{A}_{\text{Event}}) \in [0,1]$ $\quad$ $A \subseteq \Omega$

$A \in \underbrace{\mathcal{A}}_{\text{Event space}}$

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

Def 3 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 numbes $(1, \dots \infty)$

uncountable ∞ - Real numbers

Def 5 Random variable (RV)

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

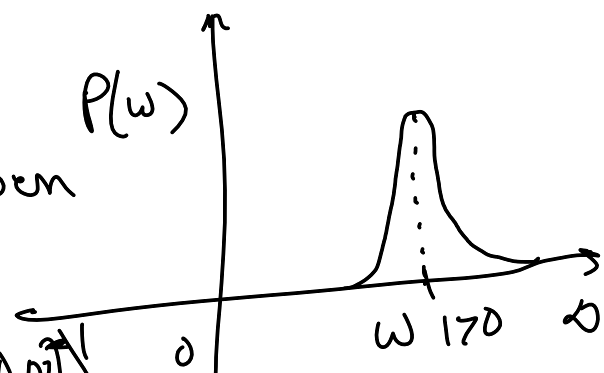
RV maps an $\underbrace{\text{event}}_{A \subseteq \Omega}$ to a set of numbers → Discrete / → continuous

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

e.g. continuous RV : weight of mic

$\Omega = [0, \infty)$

↑ 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

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

$P(X=1) = p_1$

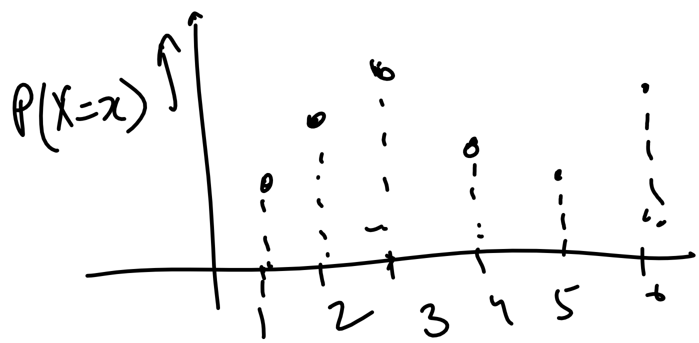
$P(X=2) = p_2$

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

$\underbrace{\boxed{P_X(x)}}$

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

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

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

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

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

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

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

Cumulative density function — 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 $= Y$

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

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

first element of Random vector

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

Random variable

RV

PDF = probability distrubution function

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

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

$\{1, \ldots 6\}$

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