Lecture #1 : Primer on probability and statistics

Notation :

- Lower case : e.g. x, y, w, z $\begin{cases} \text{event realizations, data points,} \\ \text{column vectors (row vectors e.g. } w^T) \\ \text{functions, probability measures, etc.} \end{cases}$

- Upper case : e.g X, Y, A, B $\begin{cases} \text{random variables} \\ \text{matrices} \\ \text{some functions (e.g. cdf } F) \end{cases}$

- Caligraphics : $\{$ sets, operators , e.g. $\mathcal{F}$, $\mathbb{E}$ $\}$

Probability space :

$(\underbrace{\Omega}_{\text{sample space}}, \underbrace{\mathcal{F}}_{\sigma\text{-algebra set of all events}}, p) \to$ probability measure

$p : \mathcal{F} \longrightarrow [0, 1]$

, $\underbrace{X}_{\text{random variable}} : \Omega \longrightarrow \mathcal{X}$ $\hookrightarrow$ e.g. $\mathbb{R}^n$

Discrete random variables :

A discrete r.v. X is an event taking values in a finite or countably infinite discrete space $\mathcal{X}$.

We will denote the probability of $X = x$ as $p(X = x)$ or $\boxed{p(x)}$ (more formally : $p(\{\omega \in \Omega : X(\omega) = x\})$)

$\begin{cases} \text{i) } 0 \leq p(x) \leq 1 \\ \text{ii) } \sum_x p(x) = 1 \end{cases}$, here $p(x)$ is called the probability mass function of X.

$x \in \mathcal{X}$

iii) $p(\{\emptyset\}) = 0$ — probability mass ... (pmf)

Continuous random variables: (events that take continuous values)

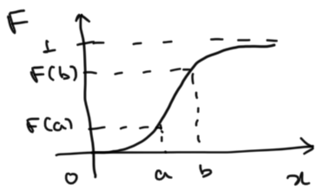
Define the events: $A = (X \leq a)$, $B = (X \leq b)$, $W = (a <$

Observe, $\underbrace{B = A \cup W}$, $\underline{A \cap W = \{\emptyset\}}$

$$p(B) = p(A) + p(W) - \cancel{p(A \cap W)}^{0}$$

$$\Rightarrow \quad p(W) = p(B) - p(A)$$

Define a function $F(x) := p(X \leq x) \rightarrow$ cummulative distribut... function of ... (cdf)



$$p(W) = p(a < X \leq b) = F(b) - F(a)$$

properties of a cdf:

i.) $0 \leq F(x) \leq 1$

ii) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$

iii) $x \leq y \Rightarrow F(x) \leq F(y)$ monotonically non decreasing function

If the cdf is differentiable, define $p(x) := \frac{d}{dx} F(x)$

↳ probability density function (pdf) of X

By definition:

$$p(W) = p(a < X \leq b) = \int_a^b p(x)\, dx$$

Properties of a pdf:

i.) $p(x) \geq 0$

ii) $\int_{-\infty}^{\infty} p(x)\, dx = 1$

iii) $\int p(x)\, dx = p(x \in A)$

⊛ We require that $p(x) \geq 0$, it is possi... that $p(x) \geq 1 \; \forall x$, as long as long as: $\int_{-\infty}^{+\infty} p(x)\, dx = 1$

$x \in A$

Lecture #2

Quantiles :

$$\boxed{F(x) := p(X \leq x)}$$

If the cdf of a continuous random variable X is a monotonically non-decreasing function, then it has an inverse.

Then $F^{-1}(\alpha)$ is the value x_α such that $p(X \leq x_\alpha) = \alpha$.

This probability is called the α-quantile of X.

e.g. $F^{-1}(0.5)$ is called the median of the distribution.

In general, we can use the inverse cdf to compute tail area probabilities.

Mean / Expected value / 1^{st}-order moment :

- Discrete case : $\mu = \mathbb{E}_{x \sim p(x)}[x] := \sum_{x \in \mathcal{X}} x \cdot p(x)$

- Continuous case : $\mu = \mathbb{E}_{x \sim p(x)}[x] = \int_{\mathcal{X}} x \, p(x) \, dx$

Properties of expectation :

- $\mathbb{E}[c] = c$
- $\mathbb{E}[c f(x)] = c \, \mathbb{E}[f(x)] = c \int_{\mathcal{X}} f(x) p(x) dx$
- $\mathbb{E}[f(x) + g(x)] = \mathbb{E}[f(x)] + \mathbb{E}[g(x)]$

Variance / 2^{nd}-order moment :

$$\sigma^2 = \underset{x \sim p(x)}{\mathbb{V}\mathrm{ar}}[x] := \mathbb{E}[(x-\mu)^2] = \int_{\mathcal{X}} (x-\mu)^2 p(x)\, dx$$

Remark: $\boxed{\mathbb{E}[x^2] = \mu^2 + \sigma^2}$

$$\sigma^2 := \int_{\mathcal{X}} (x-\mu)^2 p(x)\, dx = \overbrace{\int_{\mathcal{X}} x^2 p(x)\, dx}^{\mathbb{E}[x^2]} + \mu^2 \cancelto{1}{\int p(x)\, dx}$$

$$- 2\mu \underbrace{\int x\, p(x)\, dx}_{\mu}$$

$$= \mathbb{E}[x^2] - \mu^2$$

Standard deviation:

$$\sigma = \mathrm{Std}[x] = \sqrt{\mathbb{V}\mathrm{ar}[x]} = \sqrt{\sigma^2}$$

- Degenerate pdf: $\lim_{\sigma \to 0} \mathcal{N}(\mu, \sigma^2) = \delta(x-\mu)$,

$$\delta(x) = \begin{cases} \infty, & \text{if } x = 0 \\ 0, & \text{if } x \neq 0 \end{cases}$$

Dirac delta

$$\int_{-\infty}^{+\infty} \delta(x)\, dx = 1$$

- Shifting property: $\int_{-\infty}^{+\infty} \underbrace{\delta(x-\mu)}_{\neq 0 \text{, only when } x=\mu} p(x)\, dx = f(\mu)$

- Empirical measure / distribution:

Given some observations $\mathcal{D} := \{x_1, \dots, x_n\}$ we define:

$$p_e(\mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(\mathcal{D}), \quad \delta_{x_i}(\mathcal{D}) \text{ is the Dirac measur}$$

$$\delta_{x_i}(\mathcal{D}) = \begin{cases} 1, & x_i \in \mathcal{D} \\ 0, & x_i \notin \mathcal{D} \end{cases}$$

We can also associate weights to

We can also associate weights to each observation:

$$p_e(\mathcal{D}) = \sum_{i=1}^{n} w_i \, \delta_{x_i}(\mathcal{D}) \quad , \quad 0 \le w_i \le 1 \quad , \quad \sum_{i=1}^{n} w_i = 1$$

Joint distributions: $p(x_1, x_2, \dots, x_d)$, $x \in \mathbb{R}^d$

$$x = (x_1, x_2, \dots, x_d)$$

Covariance:

The covariance between two random variables X, Y measures the degree to which X and Y are linearly related.

$$\text{cov}[X,Y] := \mathbb{E}\left[(X - \mathbb{E}[X]) (Y - \mathbb{E}[Y])^T \right]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

If $x \in \mathbb{R}^d$ is a d-dimensional random vector, then its covariance matrix is a symmetric and positive definite matrix:

$$\text{Cov}[X] = \text{cov}[X,X] = \mathbb{E}\left[(X - \mathbb{E}[X]) (X - \mathbb{E}[X])^T \right]$$

$$\underbrace{\begin{bmatrix} \mathbb{V}ar[x_1] & \text{cov}[x_1, x_2] & \cdots & \text{cov}[x_1, x_d] \\ \vdots & \mathbb{V}ar[x_2] & & \vdots \\ \vdots & & \ddots & \vdots \\ \text{cov}[x_d, x_1] & \cdots & \cdots & \mathbb{V}ar[x_d] \end{bmatrix}}_{d \times d} = \underset{\text{covariance matrix}}{\Sigma} = \underset{\text{precision matrix}}{\Lambda^{-1}}$$

- Covariances can take values between zero and infinity.

Sometimes it is more preferable to work with a normalized measure with a finite upper bound:

Pearson correlation : $\operatorname{corr}[X,Y] := \dfrac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}}$

$$-1 \leq \operatorname{corr}[X,Y] \leq 1.$$

Specifially one can show that $\operatorname{corr}[X,Y] = 1$ iff

$$Y = \alpha X + b \quad \text{, for some } \alpha, b.$$

<u>Independence</u> : X, Y are (unconditionally) independent if

$$X \perp Y \iff p(X,Y) = p(X)\,p(Y)$$

X, Y are conditionally independent if

$$X \perp Y \mid Z \iff p(X,Y \mid Z) = p(X \mid Z)\,p(Y \mid Z)$$

If X, Y are independent : $\operatorname{cov}[X,Y] = 0$ hence X, Y are also uncorrelated.

<u>Caution</u> : The opposite is not always true!

Entropy : $\underset{x \sim p(x)}{H}[x] := -\int p(x) \log p(x)\, dx$

Relative entropy : $x \sim \underline{p(x)}$, $y \sim \underline{q(y)}$

$$\mathbb{KL}[q \| p] \neq$$

$$H[X|Y] = -\int \log \frac{p(x)}{q(y)}\, p(x)\, dx = \mathbb{KL}[p \| q$$

Kullback-Leibler divergence

Mutual information :

$p(x,y)$

$$I(X,Y) = \int_x \int_y p(x,y) \log \frac{\quad}{p(x)p(y)} \, dx \, dy$$

(= 0 if X, Y are independen