Bayes Theorem

Law of total probability: Let $B_1, B_2, \ldots, B_K$ be a disjoint partition of Ω with $B_i \in \mathcal{A}$ for all i, and $A \in \mathcal{A}$. Then

$$P(A) = \sum_{i=1}^{K} P(A|B_i) \cdot P(B_i) = \sum_{i=1}^{K} P(A \cap B_i)$$

Bayes Formula:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_i P(A|B_i) \cdot P(B_i)} = \frac{P(A \cap B_i)}{P(A)}$$

Eg: prob(poor reviews | accept), P(accept | poor reviews)

Example: COVID Testing

- Assume that 1% of all humans have COVID

- 90% of people with COVID test positive
  ("true positive")

- 8% of people without COVID test positive.
  ("false positive")

- Given that a person tested positive, what is the likelihood that they have COVID?

$$P(\text{covid} \mid \text{+ve test}) = \frac{P(\text{+ve test} \mid \text{covid}) \cdot P(\text{covid})}{P(\text{+ve} \mid \text{covid}) \cdot P(\text{covid}) + P(\text{+ve} \mid \text{no covid})\, P(\text{no covid})}$$

$$= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.08 \times 0.99}$$

$$\approx 10\%$$

Independence

Def: Consider a probability space $(\Omega, \mathcal{A}, P)$.

Two events A, B are called independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

Observation: A is independent of $B \Leftrightarrow P(A|B) = P(A)$

A family of events $(A_i)_{i \in I}$ is called independent if for all finite subsets $J \subseteq I$ we have

$$P\Big(\bigcap_{i \in J} A_i\Big) = \prod_{i \in J} P(A_i)$$

(Family is called pairwise independent if $\forall i, j \in I$:
$P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$. This does not imply that the family of events is independent)

Def: Two random variables $X: \Omega \to \Omega_1$, $Y: \Omega \to \Omega_2$ are called independent if their induced σ-algebras $\sigma(X), \sigma(Y)$ are independent: $\forall A \in \sigma(X), B \in \sigma(Y)$: $P(A \cap B) = P(A) \cdot P(B)$.

Notation for independence

(events) $A \perp\!\!\!\perp B$

(RV) $X \perp\!\!\!\perp Y$

$\cancel{\boxed{A \perp B \quad X \perp Y}}$

Don't do this

this is orthogonal

<u>Key concept</u>:

- <u>Prob</u>: CLT, sum of independent RVs.
- <u>Fairness</u>: $\hat{Y} \perp\!\!\!\perp S$ enforce while learning.
- <u>Learning theory</u>: assumption that samples are indep. w.r.t. each other.

Expectation (discrete case)

Consider a discrete random variable $X: \Omega \to \mathbb{R}$ (that is, the image $X(\Omega)$ is at most countable)

Def: $(\Omega, \mathcal{A}, P)$ be a probability space, $S \subset \mathbb{R}$ at most countable, $X: \Omega \to S$ random variable.

If $\sum_{r \in S} |r| \cdot P(X=r) < \infty$, then

$$E(X) := \sum_{r \in S} r \cdot P(X=r)$$ is called the expectation of X.

(sometimes written as EX, $\mathbb{E}X$ or $\mathbb{E}(X)$)

Examples:

- Toss a coin. $\Omega = \{\text{head}, \text{tail}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$

  $p(\text{head}) = p$, $p(\text{tail}) = 1-p$. $0 < p < 1$

  $X: \Omega \to \{0,1\}$, head $\mapsto 0$, tail $\mapsto 1$.

$$E(X) = 0 \cdot \underbrace{P(X=0)}_{p} + 1 \cdot \underbrace{P(X=1)}_{1-p}$$

$$= 1-p$$

- Test error of a classifier.

  $\hat{y} = f(x)$ where x is input, $f(\cdot)$ classifier

  y is the output.

  $$e = (\hat{y} - y)^2 = (f(x) - y)^2$$

$$\min_f \underbrace{E_x(e)}$$

Def: A random variable is called "centered" if $E(x) = 0$.

<u>Important properties:</u>

- Linear: $$E\left(\underset{\mathbb{R}}{a} \cdot \overset{RV}{X} + \underset{\mathbb{R}}{b} \cdot \overset{RV}{Y}\right) = a \cdot E(X) + b \cdot E(Y)$$

- X, Y independent $\Rightarrow E(X \cdot Y) = E(X) \cdot E(Y)$

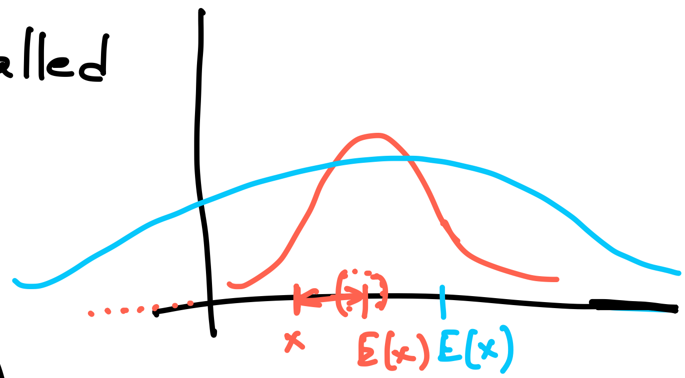
Variance, Covariance, Correlation
(discrete case)

Def: $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ discrete random variables with $E(X^2) < \infty$, $E(Y^2) < \infty$.

Then $\quad \mathrm{Var}(X) := E\left((X - E(X))^2 \right)$

is called the <u>variance of X</u>.

and $\sqrt{\mathrm{Var}(X)} =: \sigma_X$ is called the <u>standard deviation</u>.



moderate variance
high variance.

$\mathrm{Cov}(X,Y) := E\left((X-E(X)) \cdot (Y-E(Y)) \right)$

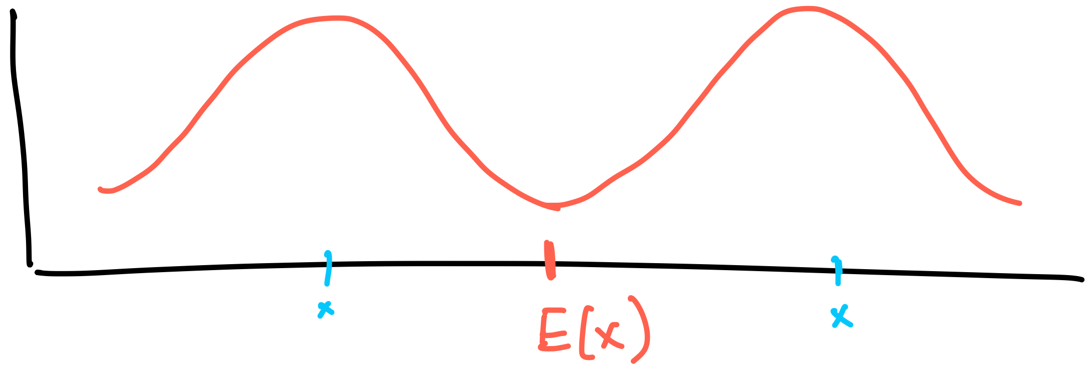
is called the <u>covariance</u> of X and Y.

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,1]$$ is called the <u>correlation coefficient</u>.

If $\mathrm{Cov}(X,Y) = 0$, then X and Y are called <u>uncorrelated</u>.

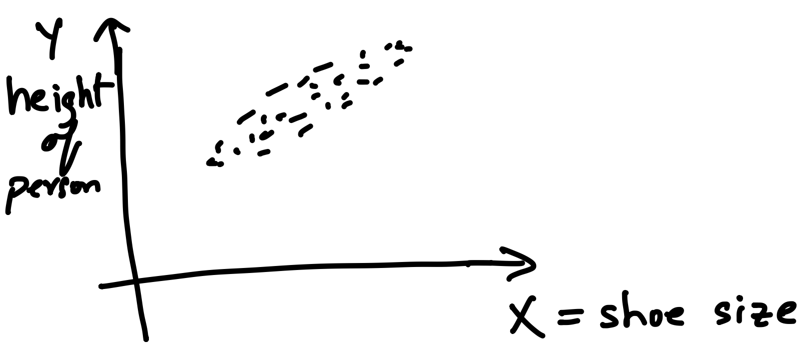
More generally, for $k \in \mathbb{N}$ we define the terms $E(X^k)$ ("k-th moment"),

$E\left((X-E(X))^k\right)$ ("k-th centered moment")



Intuition about covariance

$$Cov(X,Y) = E\left((X-E(X))\cdot(Y-E(Y))\right)$$



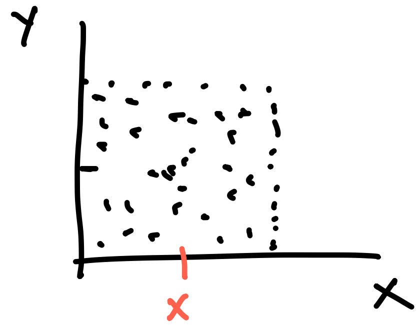
positive, large covariance $\rho \approx 0.9$

negative cov. (large in absolute value) $\rho \approx -0.9$

cov ≈ 0 (uncorrelated) $\neq$ independent.

⚠ uncorrelated $\not\Rightarrow$ independence

$\Leftarrow$



independence

Properties

- $\mathrm{Var}(X) = E(X^2) - (E(X))^2$
- $\mathrm{Cov}(X,Y) = E(X\cdot Y) - \underbrace{E(X)\cdot E(Y)}_{= E(X\cdot Y) \text{ if } X \perp\!\!\!\perp Y}$
- $E(aX+b) = a\,E(X) + b$
- $\mathrm{Var}(a\cdot X+b) = a^2\cdot \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$
- X, Y are independent $\Rightarrow$ $\mathrm{Cov}(X,Y) = 0$

  $\not\Leftarrow$
- X, Y independent $\Rightarrow$ $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.