I Probability Theory Review

1. Random variable

X takes a value in $\mathcal{X}$ (set of possible values)

according to the outcome of an event

Associated with r.v. X is a distribution $P(\underbrace{X}_{\text{r.v.}} = \underbrace{x}_{\text{value}})$

that describes the frequency of the events

Example: Discrete r.v. indicator variable $\mathcal{X} = \{0, 1\}$ # of people in a room $\mathcal{X} = \mathbb{Z}_+$

probability mass function (p.m.f.)

$P(X = x)$ = probability of x occurring ($\sum_{x \in \mathcal{X}} P(X = x) = 1$)

$0 \le P(X = x) \le 1$

Continuous r.v. sensor reading $\mathcal{X} \in \mathbb{R}$

probability density function (p.d.f.)

$p(x)$ = Likelihood of x ⇒ not precise the prob.

$p(a \le x \le b) = \int_a^b p(x) \cdot dx$. ($\int p(x)\, dx = 1$, $0 \le p(x)$, $\forall x$)

Notation: $P(X = x) = p(x) = P_{X}(x)$ val r.v.

Example Distribution:

- Bernoulli (coin)

  $\mathcal{X} = \{0, 1\}$, π = prob that 1 occurs

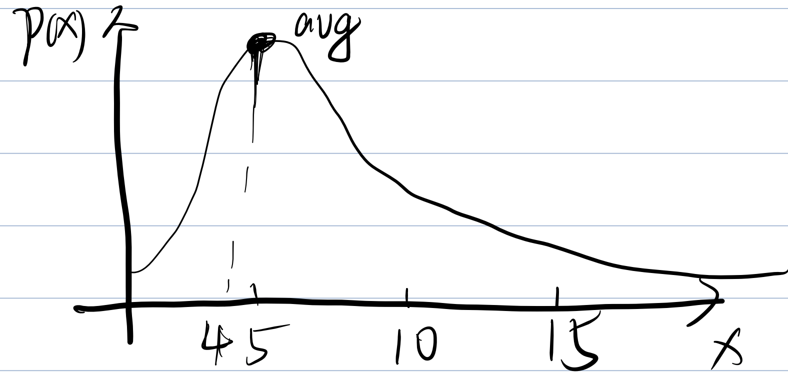
  $$\begin{cases} P_X(0) = 1 - \pi \\ P_X(1) = \pi \end{cases} \Rightarrow P(x) = \pi^{x}(1-\pi)^{1-x}$$

- poisson (# of arrivals in a fixed time period)

  $\mathcal{X} = \mathbb{Z}_+ = \{0, 1, 2, \cdots\}$, λ = average arrival rate

$(\lambda \geq 0)$

$$p(x) = \frac{1}{x!} e^{-\lambda} \cdot \lambda^{x}$$

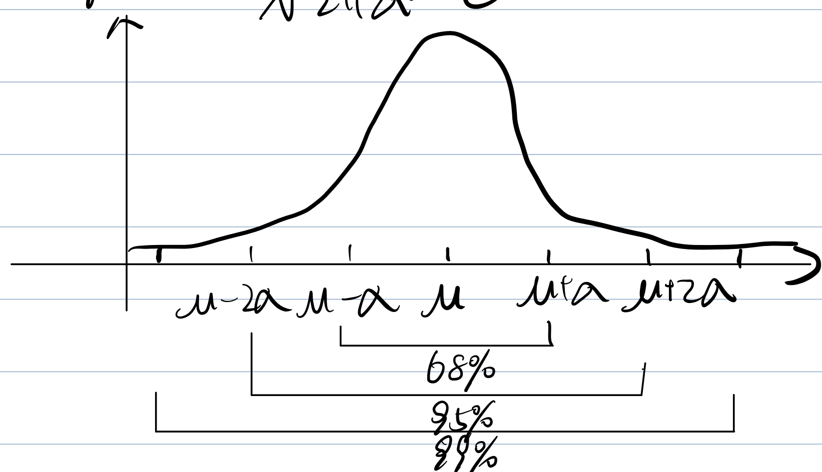


- Normal (Gaussian)

  $\mathcal{X} = \mathbb{R}$, μ = mean $\quad \sigma^2$ = Var

  σ = standard deviation

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Central Limit Theorem (CLT)

sum of N r.v. $\longrightarrow$ Gaussian distribution for large N.

2. Joint distribution

distribution over more than 1 r.v.

probability / likelihood of $X=x$ or $Y=y$:

$$P(X=x, Y=y) = p(x,y)$$

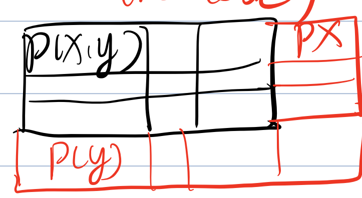
Marginal distribution

distribution over one variable in the joint

$$P(X=x) = \sum_{y \in \mathcal{Y}} p(X=x, Y=y)$$

$$p(x) = \int_{y \in \mathcal{Y}} p(x,y)\,dy$$

"marginalization" (because it is written on the margin of the table)



Conditional distribution

distribution of one r.v. when the value of another r.v. is given

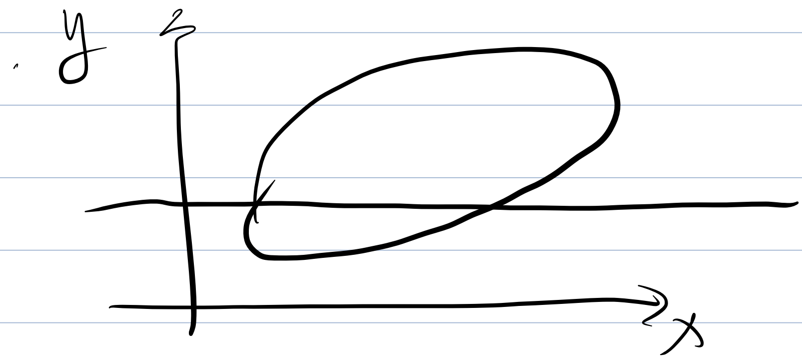
$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

3. Statistical independence

distribution of a r.v. does not change given the value of another r.v.

$X \perp\!\!\!\perp Y$ iff $p(x|y) = p(x)$

② $X \perp Y$ iff $p(x,y) = p(x)\cdot p(y)$



$p(x|y) =$

$p(x)$

range

4. Bayes' Rule

$$p(y|x) = \frac{p(x|y)\cdot p(y)}{p(x)}$$

$$p(x) = \int_y p(x,y)\,dy$$

$$= \int_y p(x|y)\,p(y)\,dy.$$

get $p(y|x)$ only from $p(x|y)\cdot p(y)$.

5. Expectations

Suppose we have a function $f(x)$. r.v. X

On average, what is the value of $f(x)$?

$$E_X[f(x)] = \sum_{x\in\mathcal{X}} f(x)\cdot p(x)$$

$$E_X[f(x)] = \int_{x\in\mathcal{X}} f(x)\cdot p(x)\,dx$$

- mean $E[x] = \int x\, p(x)\, dx = \mu_x$
- Variance $Var[x] = E[(X - E(X))^2] = \sigma_x^2$
$$= E[X^2] - (E(X))^2$$
- Covariance $Cov(x, y) = E_{xy}[(X - E(X))(y - E(y))]$
$$= \int (X - \mu_x)(y - \mu_y)\, p(x, y)\, dx\, dy$$
$$= E_{xy}(xy) - E(x)E(y) = \sigma_{xy}^2$$

6. Conditional Expectation

$$E_{x|Y}[x] = \int x\, p(x|y)\, dx$$

known

function of y

(integrating out x)

$$E_{x|Y}[f(x)] = \int f(x) \cdot p(x|y) \cdot dx$$

II. Linear Algebra

1. Column Vector : $x \in R^d$ $\quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ (usually column centric)

matrix : $A \in R^{m \times n}$ $\quad A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix}$

a_i is the ith column of A.

inner product $x^T y = \sum_i x_i y_i$ (similarity between x and y)

norm $\|x\| = \sqrt{x^T x}$

distance $d(x, y) = \|x - y\|$

outerproduct $xy^T = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & \ddots & \vdots \\ x_n y_1 & \cdots & x_n y_n \end{bmatrix}$

matrix-vector multiplication

① $y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_d \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \sum_i x_i a_i \in R^m$

$A \in R^{m\times d}$, $x \in R^d$

coefficient vector

linear combination of the colums in A

just linear combination of A, coefficient given by x

② $y = A^T x = \begin{bmatrix} - a_1^T - \\ \vdots \\ - a_m^T - \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix} = \begin{bmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{bmatrix} \in R^m$

$A \in R^{d\times m}$, $x \in R^d$

inner products between columns of A & vector x.

Matrix - matrix multiplication.

① $\underset{R^{m\times d}}{A} \cdot \underset{R^{d\times n}}{B} = A \begin{bmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ Ab_1 & \cdots & Ab_n \\ | & & | \end{bmatrix} \in R^{m\times n}$

"= A multiplied with each column of B"

② $\underset{R^{d\times m}}{A^T} \underset{R^{d\times n}}{B} = \begin{bmatrix} - a_1^T - \\ \vdots \\ - a_m^T - \end{bmatrix} \begin{bmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{bmatrix}$

$= [a_i^T b_j]_{ij} \in R^{m\times n}.$

"= All the inner products between columns of A & B"

③ $\underset{R^{m\times d}}{A} \underset{R^{n\times d}}{B^T} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix} \begin{bmatrix} - b_1^T - \\ \vdots \\ - b_n^T - \end{bmatrix} = \sum_i a_i b_i^T \in R^{m\times n}.$

"= Sum of outerproducts of columns of A & B"

2. vector r.v.

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}, \quad \mathcal{X} = R^d$$

notation $p(x_1, x_2, \ldots, x_d) = p(\underset{\text{vector}}{X})$

$$\int p(x)\,dx = 1 \iff \int \cdots \int p(x_1, x_2, \ldots, x_d)\,dx_1 \cdots dx_d = 1$$

mean vector

vector

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} = E[X] = \int_{R^d} x\,p(x)\,dx$$

vector

covariance matrix

$$\mathrm{cov}(X) = E[(X - EX)(X - EX)^T] = E[XX^T] - E(X)E(X)^T$$

2d eg.

$$= E\begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) \\ (x_1 - \mu_1)(x_2 - \mu_2) & (x_2 - \mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix}$$

covariance of variables

variance of each r.v.

3. multivariant Gaussian

$\mathcal{X} = \mathbb{R}^d$ mean $\mu \in \mathbb{R}^d$

cov matrix $\Sigma \in S^{d}_{++}$ ← positive definite (symetric) matrix

$d \times d$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\underbrace{\|x-\mu\|^2_\Sigma}} = \mathcal{N}(x|\mu,\Sigma)$$

Mahalanolas distance:

$$\|x-\mu\|^2_\Sigma = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

determinant $= |\Sigma| =$ "volume of Gaussian"

Special cases: Σ is a diagonal matrix $= \begin{bmatrix} \alpha_1^2 & & 0 \\ & \alpha_2^2 & \\ 0 & & \ddots \ \alpha_d^2 \end{bmatrix}$

$$p(x) = \prod_{i=1}^{d} \mathcal{N}(x_i|\mu_i, \alpha_i^2)$$

= product of univariate Gaussian

joint

product

marginal

i.e. d independent univariate Gaussians on each dim.