Probability Theory Review

Random variable

The r.v. X takes a value in $\mathcal{X}$ (set of possible values) according to the outcome of an event.

Eg. Event: flip a coin

$\mathcal{X} = \{H, T\}$

X has value H when coin is heads, & T when it's tails

Associated with r.v. X is a distribution $P(X=x)$ that describes the frequency of the events.

$$P(\underset{\substack{\uparrow \\ \text{r.v.}}}{X} = \underset{\substack{\uparrow \\ \text{value}}}{x})$$

Examples: Discrete r.v.

indicator variable

$$\mathcal{X} = \{\underset{\substack{\uparrow \\ \text{didn't happen}}}{0}, \underset{\substack{\uparrow \\ \text{it happened}}}{1}\}$$

\# of people in a room

$$\mathcal{X} = \underset{\substack{\uparrow \\ \text{non-negative integers}}}{\mathbb{Z}_+}$$

probability mass function (pmf)

$P(X=x)$ = probability of x occurring

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

$$0 \le P(X=x) \le 1, \ \forall x \in \mathcal{X}$$

Notation: $P(X=x) = p(x)$ ← value

$= p_X(x)$ ← r.v.

Continuous r.v.

Sensor reading

$\mathcal{X} = \mathbb{R}$

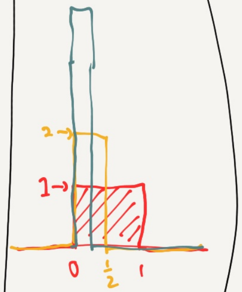
probability density function (pdf)

$p(x)$ = likelihood of x

$$P(a \le x \le b) = \int_a^b p(x)dx$$

$$\int p(x)dx = 1$$

$$0 \le p(x) \quad \forall x \in \mathcal{X}$$



Example Distributions

- Bernoulli (coin)

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

$$\begin{cases} p_X(1) = \pi \\ p_X(0) = 1-\pi \end{cases}$$

$$p(x) = \pi^x (1-\pi)^{1-x}$$

$x=0 \Rightarrow \pi^0 (1-\pi)^1 = 1-\pi$

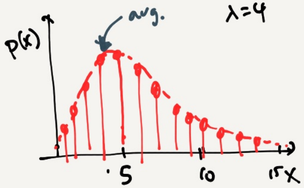
$x=1 \Rightarrow \pi^1 (1-\pi)^0 = \pi$

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

$\mathcal{X} = \mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, λ = average arrival rate $(\lambda \ge 0)$

$$p(x) = \frac{1}{\underbrace{x!}_{\text{x factorial}}} e^{-\lambda} \lambda^x$$

$x \cdot (x-1) \cdot (x-2) \cdots \cdot 1$



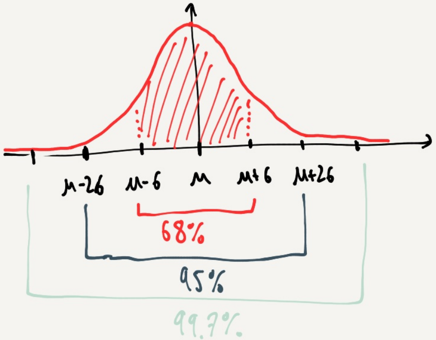
-

Normal (Gaussian)

$\mathcal{X} = \mathbb{R}$, μ = mean σ^2 = variance > 0

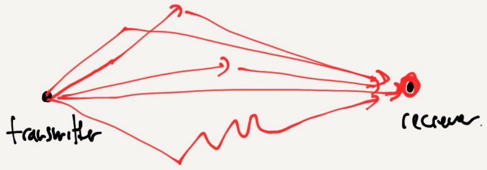
σ = standard deviation

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

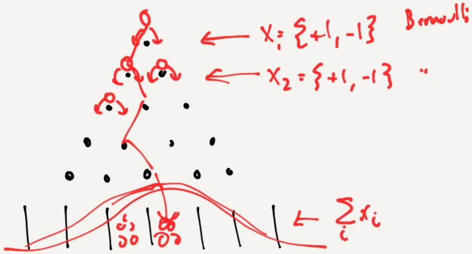


Central Limit Theorem (CLT)

Sum of N r.v. → Gaussian distribution for large N



LLN



Joint distributions

distribution over more than 1 r.v.

probability/likelihood of X=x & Y=y:

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

Example: $\mathcal{X} = \{0,1\}$
$\mathcal{Y} = \{0,1\}$

p(x,y)	Y=0	Y=1	P(x)
X=0	0.08	0.12	0.2
X=1	0.32	0.48	0.8
P(y)	0.4	0.6	

why is it called marginalization?

"margin of table"

Marginal distribution

distribution over one variable in the joint.

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

↑ summing over other r.v.

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

↑ integrating out other r.v.

"marginalization"

Conditional distribution

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

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

r.v. "given" "known value"

$p(x,y) = p(x|y)\,p(y)$

e.g.

$P(X=0 \mid Y=0) = \frac{0.08}{0.4} = 0.2$

$P(X=1 \mid Y=0) = \frac{0.32}{0.4} = 0.8$

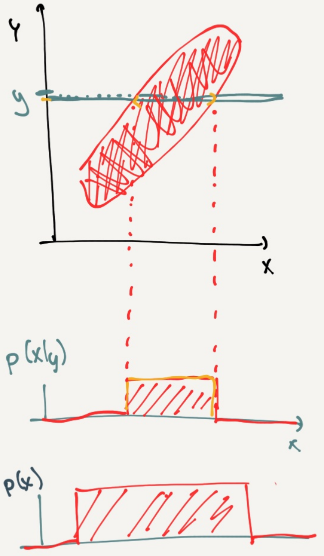
Statistical independence

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

⇒ ① $X \perp\!\!\!\perp Y$ iff $p(x|y) = p(x)$

② $X \perp\!\!\!\perp Y$ iff $p(x,y) = p(x)p(y)$ ← joint is product of marginals.

in our example $X \perp\!\!\!\perp Y$.



Bayes' Rule

$$p(x,y) = p(x|y)p(y)$$
$$p(x,y) = p(y|x)p(x)$$

$$p(y|x)p(x) = p(x|y)p(y)$$

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

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

given only p(x|y) & p(y), we can "invert" the conditioning to get p(y|x)

$$\Rightarrow \boxed{p(y|x) = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\,dy}}$$

Y = class label (digit)

X = feature value

$p(X|Y)$ = distribution of features for class Y

$p(Y)$ = prob. of class Y.

⇒ Bayes' Rule ⇒ $p(\underset{\uparrow\ \text{class}}{Y}|\underset{\uparrow\ \text{feature}}{X})$

Expectations

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

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

$$E_X[f(x)] = \sum_{x\in\mathcal{X}} f(x)p(x)$$ ⇐ weighted average of function values.

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

- **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**: $cov(x,y) = E_{xy}[(X-EX)(y-Ey)]$
  $= \int (x-\mu_x)(y-\mu_y)p(x,y)\,dx\,dy$
  $= E_{xy}(xy) - E(x)\cdot E(y) = \sigma_{xy}^2$

Conditional Expectation

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

known

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

Brief Linear Algebra Review

column vector: $x \in \mathbb{R}^d$ $\quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ (usually we are column-centric)

matrix: $A \in \mathbb{R}^{m \times n}$

\# of rows \# of columns

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_n \\ | & & | \end{bmatrix}$$

a_i is the ith column of A.

---

inner product: $x^T y = \sum_i x_i y_i$ (similarity btwn vectors x & y)

length (norm): $\|x\| = \sqrt{x^T x} = \sqrt{\sum_i x_i y_i}$

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 of all pairwise products of elements in x & y.

matrix-vector multiplication

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

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

coefficient vector

← linear combo of the columns of A (w/ entries in x as the coeff.)

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

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

inner products btwn columns of A & vector x.

Matrix-matrix multiplication

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

"A multiplied w/ each column of B"

② $$\underset{\overset{\uparrow}{A\in\mathbb{R}^{d\times m}}}{A^T}\,\underset{\overset{\uparrow}{\mathbb{R}^{d\times n}}}{B} = \begin{bmatrix} — a_1^T — \\ \vdots \\ — a_m^T — \end{bmatrix}\begin{bmatrix} | & & | \\ b_1 & \cdots & b_n \\ | & & | \end{bmatrix} = \left[a_i^T b_j\right]_{ij} \in \mathbb{R}^{m\times n}$$

"all the innerproducts btwn columns of A & B"

③ $$\underset{\overset{\uparrow}{\mathbb{R}^{m\times d}}}{A}\,\underset{\overset{\uparrow}{B\in\mathbb{R}^{n\times d}}}{B^T} = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{bmatrix}\begin{bmatrix} — b_1^T — \\ \vdots \\ — b_d^T — \end{bmatrix} = \sum_{i=1}^{d} a_i b_i^T \in \mathbb{R}^{m\times n}$$

"sum of outerproducts of columns of A & B"

Vector r.v.

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

Notation

$$\underset{}{p(x_1, x_2, x_3, \cdots, x_d) = \underset{\overset{\uparrow}{\text{vector}}}{p(x)}}, \qquad \int p(x)\,dx = 1$$

$$\downarrow$$

$$\int\cdots\int p(x_1, x_2, \dots)\,dx_1\,dx_2\cdots = 1$$

mean vector

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} = E[x] = \int_{\mathbb{R}^d} \overset{\text{vector}}{\overset{\downarrow}{x}}\,\underbrace{p(x)}_{\overset{\uparrow}{\text{number}}}\,dx \in \mathbb{R}^d$$

covariance matrix

$$\operatorname{cov}(x) = E\left[(x - Ex)(x - Ex)^T\right] = E\left[xx^T\right] - E[x]\,E[x]^T$$

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

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

covariances btwn. dim. of r.v.

variances of each dim. of r.v.

Multivariate Gaussian

$\mathcal{X} = \mathbb{R}^d$ mean $\mu \in \mathbb{R}^d$ cov matrix $\Sigma \in \mathbb{S}^d_{++}$

d×d

positive definite (symmetric) matrix

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

Mahalanobis distance:

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

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

Special cases: Σ is a diagonal matrix $= \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$

$$p(x) = \prod_{i=1}^{d} N(x_i | \mu_i, \sigma_i^2) = \text{product of univariate Gaussians.}$$

joint product marginals

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

(TBC in tutorial)