Probability Theory Review.

- Probability Space:

  (1) triple: (Ω, F, P)

  Ω : outcome space

  F : $F \subseteq P(\Omega)$ event space

  P : probability measure : $E \in F \rightarrow P(E) \in [0,1]$

  (2) restriction of F

  *: not all the event space has measure.

  ① $\Omega \in F$, $\Phi \in F$

  ② F is closed under (countable) unions.

  ③ F is closed under complement.

  (3) Properties of F:

  ① $\forall \alpha \in F$: $P(\alpha) \geq 0$

  ② $P(\Omega) = 1$

  ③ $\forall \alpha, \beta \in F$, $\alpha \cap \beta = \Phi$: $P(\alpha \cap \beta) = P(\alpha) + P(\beta)$

- Random Variables:

  (1) not variables → functions.

  outcome → $X \in R$.

  (2) indicator variable

  *: the difference between Ω and F.

- Joint distribution & Marginal Distribution.

(1) Joint distribution:

$$P(X=a, Y=b) = \delta.$$

(2) Marginal distribution:

$$P(X) = \sum_{b \in Val(Y)} P(X, Y=b)$$

- Conditional Distributions:

(1) $P(X=a \mid Y=b) = \frac{P(X=a, Y=b)}{P(Y=b)}$

knowing some events are true.

(2) $P(X \mid Y) \rightarrow$ knowing Y.

- Independence:

(1) machine learning:

data → independent.

$$P(X) = P(X \mid Y)$$

$$P(X, Y) = P(X) \cdot P(Y)$$

(2) conditional independence:

$$P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$$

*: Naive Bayes.

- Chain Rule & Bayes Rule

(1) Chain Rule:

$P(X_1 . X_2 \cdots X_n) = P(X_1) \cdot P(X_2 | X_1) \cdot P(X_3 | X_1, X_2) \cdots P(X_n | X_1 . X_2, \cdots X_{n-1})$

*: calculating joint probability.

(2) Bayes Rule:

$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)}$

not knowing $P(Y)$:

$P(Y) = \sum_X P(Y|X) P(X)$ (total probability)

- Probability Distribution:

(1) discrete & continuous

unified: measure theory.

(2) discrete:

probability mass function.

$P(X=a) = p_i$ $\sum p_i = 1$.

(3) continuous:

probability density function.

f: non-negative. integrable.

$\int f(x) dx = 1$.

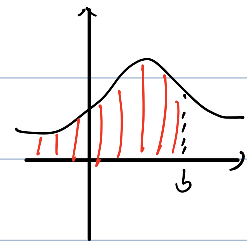
$P(a \le X \le b) = \int_a^b f(x) dx$.

$P(X=c) = \int_c^c f(x) dx = 0$.

e.g. uniform distribution over $[a.b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

cumulative distribution function:



$$F(b) = P(X \le b) = \int_{-\infty}^{b} f(x)\,dx$$

(4) Joint distribution:

$$P(a_1 \le X_1 \le b_1,\ a_2 \le X_2 \le b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2)\, dx_1\, dx_2.$$

(5) Conditional distribution:

$$f(y|x) = \frac{f(x,y)}{f(x)}$$

e.g. $P(a \le Y \le b \mid X = c) = \int_a^b \frac{f(x,y)}{f(x)}\,dy = \int_a^b \frac{f(c,y)}{f(c)}\,dy.$

- Expectations & Varience:

(1) discrete: $E(X) = \sum x\, P(X=a)$ ※: first moment

continuous: $E(X) = \int_{-\infty}^{+\infty} x f(x)\,dx$

(2) linearity:

$$E(ax+by) = aE(X) + bE(y)$$

(3) Independent:

$$E(XY) = \sum_x \sum_Y P(X=a, y=b)\, xy$$

$$= \sum_x P(X=a) \cdot x \cdot \sum_y P(y=b)\, y$$

$$= E(X) \cdot E(Y). \quad \Rightarrow E(XY) = E(X)E(Y)$$

(4) Varience: ※: second moment.

$$Var(X) = E[(X-E(X))^2] = \sigma^2 \quad \Rightarrow \sigma = \sqrt{Var(X)}$$

$$Var(X) = E(X^2) - E^2(X)$$

$Var(ax+b) = a^2 Var(x).$

(5) Independent:

$Var(X+Y) = Var(X) + Var(Y)$

Covarience: (closely related)

$Cov(X,Y) = E(X-E(X))(Y-E(Y))$

∘ Important Distribution:

(1) Bernoulli:

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

*: 2- classification tasks.

logistic regression

(2) Poisson:

fixed arrival rate λ

$P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$

(3) Gaussian: *: normal distribution.

approximate → binomial distribution.

*: noise → Gaussian white noise.

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

multi-variate: (μ, Σ)

$\mu \in R^k$ Σ: covariance matrix $\in R^{k \times k}$

$\Sigma_{ij} = Cov(X_i, X_j).$

$f(x) = \frac{1}{\sqrt{2\pi^k |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

- Efficient Manipulation:

(1) log trick:

product → sum.

*: Likelihood Function:

$$L(\theta)=\prod_{i=1}^{m}(h_\theta(x^{(i)}))^{y^{(i)}}[1-h_\theta(x^{(i)})]^{1-y^{(i)}}$$

$$\log(L(\theta))=\sum_{i=1}^{m}y^{(i)}\log(h_\theta(x^{(i)}))+\sum_{i=1}^{m}[1-y^{(i)}]\log(1-h_\theta(x^{(i)}))$$

(2) delayed normalization.

(3) Jensen's Inequality:

f: convex function.

$f(E(x))\leq E(f(x))$

*: bound → exact value.