Chapter 1: Introduction

① Supervised Learning: model $P(y|\vec{x},\theta)$

Unsupervised Learning: model $P(\vec{x}|\theta)$

② Parametric Model: parameter grows with training data

Non-Parametric Model: parameter number is fixed.

③ Overfitting

④ (Cross)-Validation & Early stop

Chapter 2: Probability

① : Some concepts:

probability mass function (pmf) : for discrete random variable.

cumulative distribution function (cdf) $F(q) \triangleq P(X \le q)$ } for continous

probability density function (pdf) $f(q) = \frac{d}{dq} F(q)$ } random variable

joint distribution, marginal distribution

conditional probability & Bayes' Rule : $P(A|B) = \frac{P(A,B)}{P(B)}$, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

independence & conditional independence : $X \perp Y \Rightarrow P(X,Y) = P(X)P(Y)$

$X \perp Y | Z \Rightarrow P(X,Y|Z) = P(X|Z)\,P(Y|Z)$

quantiles : inverse of cumulative distribution function.

② Expectation & Variance:

$E(x) \triangleq \int_x x p(x) dx$, $Var(x) = E[(X-E(x))^2] = E(X^2) - [E(x)]^2$

$E(x)$, $Var(x)$ is not defined if integral is not finite.

③ Some discrete distribution

Name	pmf	Mean	Variance	Remark
1) binomial 二项分布	$Bin(k \mid n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$	$n\theta$	$n\theta(1-\theta)$	
2) Bernoulli 伯努利分布	$Ber(x \mid \theta) = \theta^{I(x=1)}(1-\theta)^{I(x=0)}$	θ	$\theta(1-\theta)$	* $Ber(x \mid \theta) = Bin(k \mid 1,\theta)$
3) multinomial 多项分布	$Mu(\vec{x} \mid n,\vec{\theta}) = \binom{n}{x_1,x_2\cdots x_K}\prod_{j=1}^{K}\theta_j^{x_j}$	$n\vec{\theta}$	$n(diag(\vec{\theta})-\vec{\theta}\vec{\theta}^T)$	
4) Poisson 泊松分布	$Poi(x \mid \lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ	* often used to model counts of rare event.
5) empirical				

④ Some continuous distribution for scalar.

Name	pdf	Mean	Variance	Remark
1) Gaussian / Normal 高斯 / 正态分布	$\mathcal{N}(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	* sums of i.i.d random variabe → Gaussian central limit therom 中心极限定理
2) Student-t 学生-t分布	$\mathcal{T}(x \mid \mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\frac{\nu+1}{2})}$	μ (for $\nu>1$)	$\frac{\nu\sigma^2}{\nu-2}$ (for $\nu>2$)	Can handle outliers better than Gaussian.
3) Laplace 拉普拉斯分布	$Lap(x \mid \mu,b) = \frac{1}{2b}e^{-\frac{\lvert x-\mu\rvert}{b}}$	μ	$2b^2$	more weight on center; can be used to encourage sparsity
4) Gamma λ-分布	$Ga(T \mid shape=a, rate=b) = \frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb}$; $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$	$\frac{a}{b}$	$\frac{a}{b^2}$	when $a=1$, $Expon(x \mid \lambda) = Ga(x \mid 1,\lambda)$
5) Chi-square χ^2-分布	$\chi^2(x \mid \nu) = \frac{1}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}$	ν	2ν	$\chi^2(x \mid \nu) = Ga(x \mid \frac{\nu}{2},\frac{1}{2})$ sum of squared Gaussian.
6) Beta β-分布	$Beta(x \mid a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$; $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$a=b=1 \Rightarrow$ uniform; $a,b<1 \Rightarrow$ bimodal 双峰; $a,b<1 \Rightarrow$ unimodal 单峰

⑤ : Joint distribution

Covariance $\quad cov(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$

Correlation $\quad corr(X,Y) = \dfrac{Cov[X,Y]}{\sqrt{Var(X)Var(Y)}}$ $\quad$ corr $=1 \Rightarrow$ Linear; corr $=0 \Rightarrow$ independent

1) Multi-variate Gaussian

$$\mathcal{N}(\vec{x}|\vec{\mu},\vec{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

Mean $= \vec{\mu}$, $\quad$ covariance matrix $= \vec{\Sigma}$

2) Multi-variate student-t

$$\mathcal{T}(\vec{x}|\vec{\mu},\vec{\Sigma},\gamma) = \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\vec{\Sigma}|^{-1/2}}{\gamma^{D/2}\pi^{D/2}} \times \left[1+\frac{1}{\gamma}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right]^{-\frac{\gamma+D}{2}}$$

$\vec{\Sigma}$ is called scale matrix. $\quad$ mean $=\mu$, $\quad$ covariance matrix $= \dfrac{\gamma}{\gamma-2}\Sigma$

$\mathcal{T}(\vec{x}|\vec{\mu},\vec{\Sigma},\gamma)$ has fatter tail than $\mathcal{N}(\vec{x}|\vec{\mu},\vec{\Sigma})$, smaller γ is, fatter tail is.

$$\mathcal{N}(\vec{x}|\vec{\mu},\vec{\Sigma}) = \lim_{\gamma\to+\infty} \mathcal{T}(\vec{x}|\vec{\mu},\vec{\Sigma},\gamma)$$

3) Dirichlet distribution

$$Dir(\vec{x}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})}\prod_{k=1}^{K} x_k^{\alpha_k-1} I(x\in S_K)$$

where: $S_K = \{\vec{x} \mid 0\le x_k\le 1, \sum_k x_k = 1\}$; $B(\vec{\alpha}) = \dfrac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\alpha_0)}$; $\alpha_0 = \sum_{i=1}^{K}\alpha_i$

$$E(x_k) = \frac{\alpha_k}{\alpha_0} \qquad var(x_k) = \frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0^2(\alpha_0+1)}$$

α is an indicator of precision; larger α is, smaller variance is.

⑥ Ramdon variable's transformations

(1) Linear transformation.

if $\vec{x}$ is a random vector, $E(\vec{x}) = \vec{\mu}$, $\vec{y} = A\vec{x}+\vec{b}$, $var(\vec{x}) = \Sigma$

then $E(\vec{y}) = A\vec{\mu}+\vec{b}$, $var(\vec{y}) = A\Sigma A^T$

(2) Gernal Transformation.

if $\vec{y} = f(\vec{x})$, then pdf of $\vec{x}$ is $P_x(\vec{x})$, then the pdf of $\vec{y}$, $P_y(\vec{y}) = P_x(\vec{x}) \left| \frac{d\vec{x}}{d\vec{y}} \right| = P_x(\vec{x}) \left| \det \vec{J}_{\vec{y}\to\vec{x}} \right|$

here $\vec{J}_{\vec{y}\to\vec{x}}$ is the Jacob matrix from $\vec{y}$ to $\vec{x}$ i.e. $\vec{J}(i,j) = \frac{\partial \vec{y}_i}{\partial \vec{x}_j}$

⑦ Central Limit Theorem

⑧ Monte Carlo Approximation

(1) $E(f(x)) = \int f(x)p(x)dx \doteq \frac{1}{S}\sum_{s=0}^{S} f(x_s)$, where x_s is sampled on $p(x)$

(2) Monte Carlo method can be applied to approximate several statistical features, such as $E(x)$, $var(x)$, cdf of x, etc

(3) Accuracy of Monte Carlo: μ is the true expectation while $\hat{\mu}$ is the approximation

$\hat{\mu} - \mu \longrightarrow N(0, \frac{\sigma^2}{s})$, where S is # samples, σ^2 is true variance.

⑨ Information Theory

(1) entropy of a discrete distribution: $H(x) = -\sum_{k=1}^{K} P(x=k)\log_2 P(x=k)$

(2) KL-divergence: $KL(p\|q) = \sum_{k=1}^{K} P_k \log \frac{P_k}{q_k} = -H(p) + H(p,q)$

(3) cross-entropy: $H(p,q) = -\sum_{k=1}^{K} P_k \log q_k$

(4) $KL(p\|q) \geq 0$ with equality iff $p = q$

discrete distribution with maximum entropy is uniform distribution

(5) Mutual information: how one variable tells the information about another

$I(X;Y) = KL(p(X,Y) \| p(X)p(Y)) = \sum_x \sum_y P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}$

(6) conditional entropy: $H(X|Y) = \sum_y P(y) H(X|Y=y)$

(7) pointwise mutual information: $PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$