No.
Date

Math Foundation of CS182

Entropy: discrete r.v. $\mathcal{X}$ 有限，$|\mathcal{X}| < +\infty$

x 概率为 $p(x)$，信息量为 $\log \frac{1}{p(x)}$

$$\therefore H(X) = \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}\left[\log\frac{1}{p(x)}\right]$$

Joint Entropy: $H(X,Y) = \sum_{x,y\in X,Y} p(x,y)\log p(x,y) = \mathbb{E}_{XY}\left[\log\frac{1}{p(x,y)}\right]$

Conditional Entropy: $H(Y|X) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)}$

$$= -\sum_{x\in X} p(x)\sum_{y\in Y} p(y|x)\cdot\log p(y|x) = \mathbb{E}\left[\log\frac{1}{p(y|x)}\right]$$

△: **Chain Rule** $H(X, X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1})$

△ (Assumption: X_i 事件仅与 $X, X_1, \cdots, X_{i-1}$ 有关)

二元时：$H(X,Y) = H(X) + H(Y|X)$

$\Rightarrow H(Y|X) = -H(X) + H(X,Y)$

Cross Entropy: in Classification：标签分布 $p(x)$，prediction: $q(x)$

则 $H(p,q) = -\sum_x p(x)\log q(x)$

$$= \mathbb{E}_{x\sim p(x)}\left[\log\frac{1}{q(x)}\right]$$

KL-Divergence：两个分布 $p(x)\ q(x)$ 相对熵：

$$D(p(x)\|q(x)) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{q(x)}$$

均不有交换律

$$= -H(p) + H(p,q)$$

△: ML 中常用"减 KL 散度"来限制 θ_{new} 与 θ_{old} 间变化不可过大！

*: $D(p(x)\|q(x)) \geq 0$

Proof: $D = \mathbb{E}_{x\sim p(x)}\left[\log\frac{p(x)}{q(x)}\right] \leq \log\left[\mathbb{E}_{x\sim p(x)}\left[\frac{p(x)}{q(x)}\right]\right]$

$$= \log\sum_{x\in\mathcal{X}} p(x)\cdot\frac{p(x)}{q(x)}$$

欲证 p(x)! 则 $-D = \sum_{x\in X} p(x)\log\frac{q(x)}{p(x)} \le \log\left[\sum_{x} p(x)\cdot\frac{q(x)}{p(x)}\right] = 0$

$\therefore D(p(x)\|q(x)) \ge 0$，取等当且仅当：$p(x)=q(x)$

Correlation: $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} \in [-1,1]$

只可刻化线性相关性！

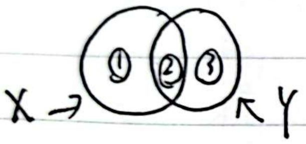
$\rho_{X,Y}=0 \underset{\checkmark}{\overset{\times}{\rightleftarrows}}$ 独立　　但Gaussian间：分布独立 $\Leftrightarrow$ 不相关

Mutual Information: $I(X;Y) =$

$$\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D(p(x,y)\|p(x)p(y))$$

D 刻化两分布距离！$\Rightarrow I(X;Y)$ 刻画 $p(x,y)$ & $p(x)p(y)$ 间距离

$I(X;Y)=I(Y;X)$; $0\le I(X;Y)\le \min\{H(X),H(Y)\}$



①: $H(X|Y)$　② $I(X;Y)$　②+③: $H(Y)$

③: $H(Y|X)$　①+②: $H(X)$　①+②+③: $H(X,Y)$

Convex: proof method: 说明 $\forall x,y\in D,\ \theta\in[0,1]$:

$f(\theta x+(1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$

Linear Algebra:

· Trace 迹: $Tr(A) = \sum_{i=1}^{n} a_{ii}$

则有: $Tr(AB) = Tr(BA)$

matrix inner product: $\langle A,B\rangle = Tr(A^TB) = Tr(B^TA)$

$Tr(aA+bB) = a\,Tr(A) + b\,Tr(B)$

· $(AB)^T = B^TA^T$; AA^T 必是对称矩阵.

proof: $(AA^T)^T = (A^T)^TA^T = AA^T$

No.
Date

· 逆: $AB=BA=I_n$, 则 $B=A^{-1}$; 若A无逆,则称A是奇异的 (singular), $|A|=0$

$(AB)^{-1}=B^{-1}A^{-1}$, $(A^T)^{-1}=(A^{-1})^T$

· 行列式: $\det(A)$ or $|A|=\prod_{i=1}^{n}\lambda_i$

$A^{-1}=\frac{1}{|A|}A^*$, where $A^*=[C_{ij}]^T$, C_{ij} 是 a_{ij} 的余子式

$C_{ij}=(-1)^{i+j}M_{ij}$, M_{ij} 是子式, 为剔除i行j列剩下n-1方阵的行列式

△: $|A^T|=|A|$, $|\lambda A|=\lambda^n|A|$, $|AB|=|A||B|$, $|A^{-1}|=\frac{1}{|A|}$

· 上三角矩阵: $A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & a_{22} & \vdots \\ 0 & \cdots & a_{nn}\end{bmatrix}$, $\det(A)=\prod_{i=1}^{n}a_{ii}$

· 对角矩阵: $\Lambda=\mathrm{diag}(d_1,\cdots,d_n)$, $|\Lambda|=\prod_{i=1}^{n}d_i$

· $A\in R^{m\times n}$ $\mathrm{row}(A)=\mathrm{span}\{r_1,\cdots,r_m\}$ $\mathrm{null}(A)=\{x\in R^n: Ax=0\}$

$\mathrm{col}(A)=\mathrm{span}\{c_1,\cdots,c_n\}$ $\mathrm{null}(A^T)=\{x\in R^m: A^Tx=0\}$

☆: $\mathrm{rank}(A)+\mathrm{nullity}(A)=n$

· 秩: $\mathrm{rank}(A)=\dim(\mathrm{row}(A))=\dim(\mathrm{col}(A))$, $\mathrm{rank}(A)\le\min(n,m)$

秩最本质: 非0奇异值的个数

$\mathrm{rank}(A^TA)=\mathrm{rank}(A)$ $\mathrm{rank}(AB)\le\min(\mathrm{rank}(A),\mathrm{rank}(B))$

· 范数 norm: $\|x+y\|\le\|x\|+\|y\|$, $\|x\|\ge 0$, $\forall a$ $\|ax\|=|a|\|x\|$

☆: all norms $f: R^n\to R$ are convex 凸 △: 0-norm 不是

△: $\|v\|_\infty=\max\{|v_1|,|v_2|,\cdots,|v_n|\}$

△ Hölder's Inequality: $p,q\ge 0$, $1/p+1/q=1$ 则:

① $\sum_{i=1}^{n}|a_ib_i|\le\|x\|_p\|y\|_q$ ② $\sum_{i=1}^{n}|a_ib_i|\le\|x\|_1\|y\|_\infty$

△ Cauchy Inequality: $\sum_{i=1}^{n}|a_ib_i|\le\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}$

Norm for Matrices

△ 特殊norm: Frobenius Norm: $\|A\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}=\sqrt{Tr(A^TA)}$

Nuclear Norm: $\|A\|_*=\sum_{i=1}^{\min(m,n)}\sigma_i(A)$

KOKUYO

No.
Date

△: 未说范数 $\|x\|_{?}$, 则默认为 2

p-norm: $\|A\|_p = \max_{\|x\|_p \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$, $p=1$: $\max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$

$p=2$: $\|A\|_2 = \sqrt{\lambda_{max}(A^TA)} = \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2}$

$p=\infty$: $\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$

· Dual Norm: 对偶范数 $\|x\|_* = \sup_{\|z\|\le 1} \langle x, z\rangle$

△: $x^T z \le \|x\|_* \|z\|$

· 点到面: $x_0 \to R^n$ 中平面: $H = \{x \in R^n : w^T x + b = 0\}$:

$d = \frac{|w^T x_0 + b|}{\|w\|}$

正交投影: u on v: $w_1 = proj_v(u) = \frac{u \cdot v}{\|v\|^2} v$, $w_2 = u - w_1$

· Quadratic Form 二次型: $Q(x) = x^T \underbrace{A}_{对称} x = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j$

Eg: $Q(x) = x_1^2 + 2x_1x_2 + 3x_2^2 = [x_1\ x_2]\begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

正定: $A \in S_{++}^n$, $A \succ 0$, $\forall x \in R^n \setminus \{\vec{0}\}$, $x^T A x > 0$

半正定: $A \in S_+^n$, $A \succeq 0$, $\forall x \in R^n$, $x^T A x \ge 0$

判定正定矩阵: ① 特征值均 >0 ② $\forall x \neq \vec{0}$, $x^T A x > 0$

Eg: A^TA 正定, 因为: $x^T A^T A x = (Ax)^T Ax = \|Ax\|^2 \ge 0$ for $\forall x \neq \vec{0}$

· 特征值: eigenvalue: $Ax = \lambda x$, $(\lambda I - A)x = 0$ $\exists x \in R^n, x \neq \vec{0}$

$\Rightarrow |\lambda I - A| = 0$, $p(\lambda) = |\lambda I - A| \Rightarrow$ 特征多项式, x 是特征向量

而 null$(A - \lambda I)$ 为 λ 对应的特征空间

✲ Skill: 施密特正交化① 相似对角化② 特征值分解③

正交对角化④ 奇异值分解 (SVD)⑤

No.

Date

① 一组线性无关的向量 ⇒ 一组正交向量，（且模均为1）

设原：$\alpha_1, \cdots, \alpha_n$. ⇒ $\beta_1, \cdots, \beta_n$. α_2在β_1上投影

先令 $\beta_1 = \alpha_1 / \|\alpha_1\|_2$，然后 $\varepsilon_2 = \alpha_2 - \langle \alpha_2, \beta_1 \rangle \cdot \beta_1$

$\beta_2 = \varepsilon_2 / \|\varepsilon_2\|_2$.

之后：$\varepsilon_i = \alpha_i - \sum_{j=1}^{i-1} \langle \alpha_i \cdot \beta_j \rangle \beta_j$，$\beta_i = \varepsilon_i / \|\varepsilon_i\|_2$

A有n个线性无关特征向量，才可对角化

② 相似对角化*：A方阵 ⇒ PDP^{-1}，D为对角矩阵

△：相似定义：对于A.B来说，若∃可逆P, s.t.，$A = PBP^{-1}$，则AB相似

step 1. $\det(A - \lambda I) = 0$，得 $\lambda_1, \cdots, \lambda_j$ （$j \le n$，λ_i可能为多重根）

step 2. 对于λ_i，求 nullity $(A - \lambda I)$（并且施密特正交化），得系ξ_i.

step 3. 构造 $D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_j \end{bmatrix}$，注意$\lambda_i$重数为k，有k个$\lambda_i$.

$P = [\xi_1, \cdots, \xi_j] \in \mathbb{R}^{n \times n}$，注意$\xi_i$可能含不止一个列向量.

③. 特征值分解：就是对角化

④. 正交对角化：△：只有对称矩阵可以；相似对角化强制施密特正交化

⑤. 奇异值分解：可以操作非方阵：欲 $A = U \Sigma V^T$，$A \in \mathbb{R}^{m \times n}$

其中：$U \in \mathbb{R}^{m \times m}$，$\Sigma \in \mathbb{R}^{m \times n}$，$V \in \mathbb{R}^{n \times n}$

step 1：计算 A^TA，解 $\det(A^TA - \lambda I) = 0$，出 $\sigma_1^2 \ge \sigma_2^2 \cdots \ge \sigma_r^2 > \sigma_r^2 = \cdots = \sigma_n^2 = 0$

解出对应特征向量 $v_1, \cdots, v_r$，并且归一化

step 2：$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ 用0填至$\mathbb{R}^{m \times n}$，$V = [v_1, \cdots, v_r]$.

则 $U = [u_1, \cdots, u_r]$，$u_i = \frac{1}{\sigma_i} A v_i$

KOKUYO

扫描全能王 创建

No.

Date

Optimization: $x \backslash y$	scalar	vector	matrix
矩阵求导: Scalar	dy/dx	$\frac{\partial \vec{y}}{\partial x}$ ②	$\frac{d\vec{Y}}{dx}$ ③
vector	$\frac{\partial y}{\partial \vec{x}}$ ①	$\frac{\partial \vec{y}}{\partial \vec{x}}$ ④	
matrix	$\frac{\partial y}{\partial X}$		

② $= \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{pmatrix}$ ① $= \left[\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_n}\right]$ ④: $\begin{bmatrix} \frac{\partial y_1}{\partial x_1}, & \cdots, & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial y_n}{\partial x_1}, & \cdots, & \frac{\partial y_n}{\partial x_n} \end{bmatrix}$

③: $\left[\frac{\partial Y_{i,j}}{\partial x}\right]$

求导结果有两种维度布局，分子布局：求导结果维度以分子为主，分母反之。~~matrix~~ vector 对 scalar 求导：分子；scalar 对 vector/matrix：分母

常见：$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$，$\frac{\partial x^T A x}{\partial x} = (A+A^T)x$

如 $L = \|b-Ax\|_2^2$，$= (b-Ax)^T(b-Ax) = b^Tb - \underbrace{b^TAx - x^TA^Tb}_{①} + \underbrace{x^TA^TAx}_{②}$

$\frac{\partial ①}{\partial x} = -(b^TA)^T - A^Tb = -2A^Tb$ $\quad \frac{\partial ②}{\partial x} = (A^TA + (A^TA)^T)x = 2A^TAx$

多元微分：∇(nabla)：$\nabla f = \frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right]^T$

$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$ 第 i,j 元素：$\frac{\partial^2 f}{\partial x_i \partial x_j}$

Convex: $f(x)$ 凸 iff: ① $\forall x, y \in \mathbb{R}^n$, $\forall \theta \in [0,1]$,
$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$
② $\forall x \in \mathbb{R}^n$, $\nabla^2 f(x) \succeq 0$
③ $\forall x, y \in \mathbb{R}^n$, $f(y) \ge f(x) + \nabla f(x)^T (y-x)$.

等价

$\Downarrow$

半正定，特征值都大于/等于0

Campus

扫描全能王 创建

No.

Date

Taylor: $f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x)$

中值: $\exists\, \theta \in [0,1]$, $z = \theta x + (1-\theta) y$.

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x).$$

or. $f(y) = f(x) + \nabla f(z)^T (y-x)$

Jensen's Inequality: $f(\mathbb{E}(X)) \le \mathbb{E}(f(X))$

$\Rightarrow$ 优化角度: $f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$, $\forall x, y \in \mathbb{R}^n$, $\theta \in [0,1]$

Convex Problem: 原. $\begin{cases} \min\limits_{x} & f_0(x) \\ \text{s.t.} & f_i(x) \le 0,\ h_i(x) = 0. \end{cases}$

则 Lagrange: $\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{n} \nu_i h_i(x).$

若 $f_0(x)$. $f_i(x)$ 为凸, $h_i(x)$ 为线性函数, 则是凸优化问题

Eg: $\min : x_1^2 + x_2^2$, 且: $x_2 \le \alpha$, $x_1 + x_2 = 1$.

由原 $\Rightarrow$ Lagrange 问题 (对偶问题):

Dual: $\max\limits_{\lambda, \nu}\ g(\lambda, \nu)$ s.t., $\lambda \succeq 0$, $\mathcal{L}(x, \lambda, \nu)$ 取中 $g(\lambda, \nu)$

※: 永远凹函数

KKT Condition: ① $f_i(x) \le 0$. $h_i(x) = 0$. ② $\lambda \succeq 0$ ($L(X, \mu, \lambda)$ 取中 $g(\lambda, \mu)$)

③ $\lambda_i f_i(x) = 0$ ④ $\nabla_x \mathcal{L}(x, \lambda, \nu) = 0$

Eg: $\mathcal{L}(x_1, x_2, \mu, \lambda) = x_1^2 + x_2^2 + \lambda(x_2 - \alpha) + \mu(1 - x_1 - x_2)$, $\lambda \ge 0$

Dual: $g(\lambda, \mu) = \min\limits_{x_1, x_2} \mathcal{L}(x_1, x_2, \lambda, \mu)$, $\boxed{\max\limits_{\mu, \lambda} g(\lambda, \mu) \quad \text{s.t. } \lambda \ge 0.}$

※: ① 若原问题为凸优化问题, 且必KKT得最优解

② KKT条件. 未必充分 ③ 可能发生"有最优解, 但KKT无解"情况

KOKUYO

No.

Date . .

△. 凸函数是否严格是个重要话题!

Eg: T/F: If MSE has a unique minimizer (i.e. argmin), then MAE must also have a unique minimizer. False!

MSE: strictly convex; MAE, not strictly convex

(反之,成立! 事实上 MSE总有minimizer)

5