数学基础

微积分：向量和矩阵求导

考虑$\frac{\mathrm{dA}}{\mathrm{dB}}$, A和B都可能是标量、向量或矩阵，共有9种不同的导数

自变量\因变量	标量ϕ	向量f	矩阵F
标量ξ	$\frac{\mathrm{d}\phi}{\mathrm{d}\xi}$	$\frac{\mathrm{d}f}{\mathrm{d}\xi}$	$\frac{\mathrm{d}F}{\mathrm{d}\xi}$
向量x	$\frac{\mathrm{d}\phi}{\mathrm{d}x}$	$\frac{\mathrm{d}f}{\mathrm{d}x}$	$\frac{\mathrm{d}F}{\mathrm{d}x}$
矩阵X	$\frac{\mathrm{d}\phi}{\mathrm{d}X}$	$\frac{\mathrm{d}f}{\mathrm{d}X}$	$\frac{\mathrm{d}F}{\mathrm{d}X}$

微积分：向量和矩阵求导

一、向量f对标量ξ的导数$\frac{\mathrm{d}f}{\mathrm{d}\xi}$

定义1 对于n维向量函数(vector function)

$$f(\xi) = \left(f_1(\xi), f_2(\xi), \cdots, f_n(\xi)\right)^T$$

定义它对ξ的导数为：

$$\frac{\mathrm{d}f(\xi)}{\mathrm{d}\xi} \triangleq \left(\frac{\mathrm{d}f_1(\xi)}{\mathrm{d}\xi}, \frac{\mathrm{d}f_2(\xi)}{\mathrm{d}\xi}, \cdots, \frac{\mathrm{d}f_n(\xi)}{\mathrm{d}\xi}\right)^T$$

也可以对行向量求导：

$$\frac{\mathrm{d}f(\xi)^{\mathrm{T}}}{\mathrm{d}\xi} \triangleq \left(\frac{\mathrm{d}f_1(\xi)}{\mathrm{d}\xi}, \frac{\mathrm{d}f_2(\xi)}{\mathrm{d}\xi}, \cdots, \frac{\mathrm{d}f_n(\xi)}{\mathrm{d}\xi}\right)$$

微积分：向量和矩阵求导

二、矩阵F对标量ξ的导数$\dfrac{\mathrm{d}F}{\mathrm{d}\xi}$

定义2. 对于$n\times m$维矩阵函数(matrix function)

$$F(\xi)=\begin{pmatrix} F_{11}(\xi), & F_{12}(\xi), & \cdots & F_{1m}(\xi) \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1}(\xi), & F_{n2}(\xi), & \cdots & F_{nm}(\xi) \end{pmatrix}=\big(F_{ij}(\xi)\big)_{nm}$$

定义它对ξ的导数为：

$$\frac{\mathrm{d}F(\xi)}{\mathrm{d}\xi}\triangleq\begin{pmatrix} \dfrac{dF_{11}(\xi)}{d\xi}, & \dfrac{dF_{12}(\xi)}{d\xi}, & \cdots & \dfrac{dF_{1m}(\xi)}{d\xi} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{dF_{n1}(\xi)}{d\xi}, & \dfrac{dF_{n2}(\xi)}{d\xi}, & \cdots & \dfrac{dF_{nm}(\xi)}{d\xi} \end{pmatrix}$$

$$=\left(\frac{dF_{ij}(\xi)}{d\xi}\right)_{nm}$$

上述两个定义是统一的！

微积分：向量和矩阵求导

运算性质：

（1）加法运算公式

$$\frac{\mathrm{d}(F \pm G)}{\mathrm{d}\xi} = \frac{\mathrm{d}F}{\mathrm{d}\xi} \pm \frac{\mathrm{d}G}{\mathrm{d}\xi}$$

（2）数乘运算公式（$\lambda(\xi)$ 是标量函数）

$$\frac{\mathrm{d}(\lambda(\xi)F)}{\mathrm{d}\xi} = \frac{\mathrm{d}\lambda(\xi)}{\mathrm{d}\xi}F + \lambda(\xi)\frac{\mathrm{d}F}{\mathrm{d}\xi}$$

（3）乘法运算公式

$$\frac{\mathrm{d}(FG)}{\mathrm{d}\xi} = \frac{\mathrm{d}F}{\mathrm{d}\xi}G + F\frac{\mathrm{d}G}{\mathrm{d}\xi}$$

例1： 求$x^{\mathrm{T}}Ax$对ξ的导数，其中x是ξ的n维向量函数， A是$n \times n$对称常数矩阵。

微积分：向量和矩阵求导

三、标量ϕ对向量x的导数$\frac{\mathrm{d}\phi}{\mathrm{d}x}$

设函数$\phi(x) = \phi(x_1, \cdots, x_n)$是以$x$为自变量的数量函数
定义3.标量函数ϕ对列向量x的导数为：

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} \triangleq \left(\frac{\partial \phi}{\partial x_1}, \frac{\partial \phi}{\partial x_2}, \cdots, \frac{\partial \phi}{\partial x_n}\right)^T$$

也称为函数ϕ的梯度，记为grad ϕ或$\nabla\phi$
也可以对行向量求导：

$$\frac{\mathrm{d}\phi}{\mathrm{d}x^T} \triangleq \left(\frac{\partial \phi}{\partial x_1}, \frac{\partial \phi}{\partial x_2}, \cdots, \frac{\partial \phi}{\partial x_n}\right)$$

显然：$\frac{d\phi(x)}{dx} = \left(\frac{d\phi(x)}{dx^T}\right)^T$

微积分：向量和矩阵求导

运算性质：

（1）加法运算公式

$$\frac{\mathrm{d}(\varphi \pm \phi)}{\mathrm{d}x} = \frac{\mathrm{d}\varphi}{\mathrm{d}x} \pm \frac{\mathrm{d}\phi}{\mathrm{d}x}$$

（2）乘法运算公式

$$\frac{\mathrm{d}(\varphi\phi)}{\mathrm{d}x} = \frac{\mathrm{d}\varphi}{\mathrm{d}x}\phi + \varphi\frac{\mathrm{d}\phi}{\mathrm{d}x}$$

例2： 求函数$\phi(x) = x^{\mathrm{T}}x = x_1^2 + x_2^2 + \cdots + x_n^2$对$x$的导数。

四、向量f对向量x的导数$\frac{\mathrm{d}f}{\mathrm{d}x}$

设函数$f(x)=\left(f_1(x),\cdots,f_m(x)\right)^{\mathrm{T}}$是$x$的$m$维列向量函数

定义4. $n\times m$阶矩阵函数：

$$\begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_1} \\ \vdots & & \vdots \\ \dfrac{\partial f_1}{\partial x_n} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{pmatrix} \triangleq \left(\frac{\partial f_j}{\partial x_i}\right)_{n\times m}$$

称为m维行向量函数$f(x)^{\mathrm{T}}$对n维列向量x的导数，

记为：$\frac{df(x)^T}{dx}$

微积分：向量和矩阵求导

定义4续. $m\times n$阶矩阵函数：

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \triangleq \left(\frac{\partial f_i}{\partial x_j}\right)_{m\times n}$$

称为m维列向量函数$f(x)$对n维行向量x^T的导数。

记为：$\frac{df(x)}{dx^T}$

显然：$\frac{df(x)^T}{dx} = \left(\frac{df(x)}{dx^T}\right)^T$

微积分：向量和矩阵求导

运算性质：

（1）加法运算公式

$$\frac{\mathrm{d}(f^T \pm g^T)}{\mathrm{d}x} = \frac{\mathrm{d}f^T}{\mathrm{d}x} \pm \frac{\mathrm{d}g^T}{\mathrm{d}x}$$

（2）数乘运算公式

$$\frac{\mathrm{d}(\lambda(x)f^T)}{\mathrm{d}x} = \frac{\mathrm{d}\lambda(x)}{\mathrm{d}x}f^T + \lambda(x)\frac{\mathrm{d}f^T}{\mathrm{d}x}$$

（3）乘法运算公式

$$\frac{\mathrm{d}(f^T g)}{\mathrm{d}x} = \frac{\mathrm{d}f^T}{\mathrm{d}x}g + \frac{\mathrm{d}g^T}{\mathrm{d}x}f$$

注意：我们有$\frac{\mathrm{d}x^T}{\mathrm{d}x} = I$和$\frac{\mathrm{d}x}{\mathrm{d}x^T} = I$

微积分：向量和矩阵求导

例3： (1) 求行向量$x^T A$对x的导数

(2) 求列向量Bx对x^T的导数，其中A、B是常数矩阵，但不一定是方阵。

例 4：求二次型$x^T Ax$对x的导数

例 5：求数量函数$p^T Ax$对x的导数，其中p^T是$1\times n$行向量，A是$n\times n$矩阵， 都是常量。 x是$n\times 1$列向量

微积分：向量和矩阵求导

复合函数微分法

1.数量函数的公式

公式1. 设$\phi = \phi(x)$, $x = x(\xi)$, 则

$$\frac{d\phi}{d\xi} = \frac{d\phi}{dx^T}\frac{dx}{d\xi} = \frac{dx^T}{d\xi}\frac{d\phi}{dx}$$

公式2. 设$\phi = \phi(y)$, $y = y(x)$, 则

$$\frac{d\phi}{dx} = \frac{dy^T}{dx}\frac{d\phi}{dy}$$

$$\frac{d\phi}{dx^T} = \frac{d\phi}{dy^T}\frac{dy}{dx^T}$$

微积分：向量和矩阵求导

公式3. 设$\phi = \phi(x,y)$，$y = y(x)$，则

$$\frac{d\phi}{dx} = \frac{dy^T}{dx}\frac{\partial\phi}{\partial y} + \frac{\partial\phi}{\partial x}$$

$$\frac{d\phi}{dx^T} = \frac{\partial\phi}{\partial y^T}\frac{dy}{dx^T} + \frac{\partial\phi}{\partial x^T}$$

微积分：向量和矩阵求导

2.向量函数的公式

公式4. 设$z = z(y)$, $y = y(\xi)$, 则

$$\frac{dz}{d\xi} = \frac{dz}{dy^T}\frac{dy}{d\xi}$$

公式5. 设$z = z(y)$, $y = y(x)$, 则

$$\frac{dz^T}{dx} = \frac{dy^T}{dx}\frac{dz^T}{dy}$$

$$\frac{dz}{dx^T} = \frac{dz}{dy^T}\frac{dy}{dx^T}$$

微积分：向量和矩阵求导

公式6. 设$z=z(x,y)$，$y=y(x)$，则

$$\frac{dz^T}{dx}=\frac{dy^T}{dx}\frac{\partial z^T}{\partial y}+\frac{\partial z^T}{\partial x}$$

$$\frac{dz}{dx^T}=\frac{\partial z}{\partial y^T}\frac{dy}{dx^T}+\frac{\partial z}{\partial x^T}$$

五、标量ϕ对矩阵X的导数$\frac{\mathrm{d}\phi}{\mathrm{d}X}$

设函数$\phi = \phi(X)$是以$p \times m$矩阵X的$p \times m$个元X_{ij}为自变量的数量函数，简称以矩阵X为自变量的数量函数。

函数$f = X_{11}^3 + (1 + X_{12})X_{11}^2 + (X_{21} + X_{22} + X_{23})X_{11} + X_{21} + X_{22}$

$$= \begin{pmatrix} X_{11} & 1 \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} X_{11} \\ 1 \end{pmatrix} = f(X)$$

就是以

$$\mathrm{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$$

为自变量的函数。

微积分：向量和矩阵求导

定义5. $p \times m$矩阵函数：

$$\begin{pmatrix} \dfrac{\partial \phi}{\partial X_{11}} & \cdots & \dfrac{\partial \phi}{\partial X_{1m}} \\ \vdots & & \vdots \\ \dfrac{\partial \phi}{\partial X_{p1}} & \cdots & \dfrac{\partial \phi}{\partial X_{pm}} \end{pmatrix} = \left(\frac{\partial \phi}{\partial X_{ij}}\right)_{p \times m}$$

称为数量矩阵ϕ对矩阵X的导数。

记为：$\dfrac{d\phi}{dX}$

例 6 求$\phi = x^T A x$ 对矩阵A的导数，其中向量x是定常的。

微积分：向量和矩阵求导

运算性质：

（1）加法运算公式

$$\frac{\mathrm{d}(\phi \pm \varphi)}{\mathrm{d}X} = \frac{\mathrm{d}\phi}{\mathrm{d}X} \pm \frac{\mathrm{d}\varphi}{\mathrm{d}X}$$

（2）乘法运算公式

$$\frac{\mathrm{d}(\phi^T \varphi)}{\mathrm{d}X} = \frac{\mathrm{d}\varphi}{\mathrm{d}X}\phi + \varphi\frac{\mathrm{d}\phi}{\mathrm{d}X}$$

五、标量ϕ对矩阵X的导数$\frac{\mathrm{d}\phi}{\mathrm{d}X}$

设函数$\phi = \phi(X)$是以$p \times m$矩阵X的$p \times m$个元X_{ij}为自变量的数量函数，简称以矩阵X为自变量的数量函数。

函数$f = a^2X_{11} + (X_{21} + X_{12})a + X_{22}$

$$= \begin{pmatrix} a & 1 \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} a \\ 1 \end{pmatrix} = f(X)$$

就是以

$$\mathrm{X} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$$

为自变量的函数。

微积分：向量和矩阵求导

定义5. $p \times m$矩阵函数：

$$\begin{pmatrix} \frac{\partial \phi}{\partial X_{11}} & \cdots & \frac{\partial \phi}{\partial X_{1m}} \\ \vdots & & \vdots \\ \frac{\partial \phi}{\partial X_{p1}} & \cdots & \frac{\partial \phi}{\partial X_{pm}} \end{pmatrix} = \left(\frac{\partial \phi}{\partial X_{ij}}\right)_{p\times m}$$

称为数量矩阵ϕ对矩阵X的导数。

记为：$\frac{d\phi}{dX}$

例 6 求$\phi = x^T Ax$ 对矩阵A的导数，其中向量x是定常的。

微积分：向量和矩阵求导

运算性质：

（1）加法运算公式

$$\frac{\mathrm{d}(\phi \pm \varphi)}{\mathrm{d}X} = \frac{\mathrm{d}\phi}{\mathrm{d}X} \pm \frac{\mathrm{d}\varphi}{\mathrm{d}X}$$

（2）乘法运算公式

$$\frac{\mathrm{d}(\phi\varphi)}{\mathrm{d}X} = \frac{\mathrm{d}\varphi}{\mathrm{d}X}\phi + \varphi\frac{\mathrm{d}\phi}{\mathrm{d}X}$$

微积分：向量和矩阵求导

矩阵微分法

标量函数对矩阵的导数，即$\frac{d\phi}{dX}$

微分法：

$$d\phi = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial\phi}{\partial X_{ij}}dX_{ij} = tr\left(\frac{\partial\phi}{\partial X^T}dX\right)$$

微分性质

（1）加法运算公式：

$$d(X \pm Y) = dX \pm dY$$
$$d(XY) = d(X)Y + Xd(Y)$$
$$d(X^T) = (dX)^T$$
$$d\,tr(X) = tr(dX)$$

（2）逆：$dX^{-1} = -X^{-1}dX\,X^{-1}$

微积分：向量和矩阵求导

（3）行列式：$d|X| = tr(X^* dX)$
（4）逐元素乘法： $d(X \odot Y) = dX \odot Y + X \odot dY$
（5）逐元素函数： $d\sigma(X) = \sigma'(X) \odot dX$

关于迹的运算
（1）标量的迹：$\xi = tr(\xi)$
（2）转置：$tr(X) = tr(X^T)$
（3）线性：$tr(X \pm Y) = tr(X) \pm tr(Y)$
（4）矩阵乘法：$tr(XY) = tr(YX)$
（5）矩阵乘法/逐元素乘法：$tr\left(X^T (Y \odot Z)\right) = tr\left((X \odot Y)^T Z\right)$
其中X，Y，Z具有相同的尺寸。

微积分：向量和矩阵求导

例 1. $\phi = a^T X b$, 计算$\frac{d\phi}{dX}$. 其中a是m维列向量，X是$m \times n$矩阵，b是n维列向量.

例 2. $\phi = a^T \exp(Xb)$, 计算$\frac{d\phi}{dX}$. 其中a是m维列向量， X是$m \times n$矩阵，b是n维列向量.

例 3. $\phi = tr(Y^T M Y)$, $Y = \sigma(WX)$, 计算$\frac{d\phi}{dX}$. 其中W是$l \times m$矩阵， X是$m \times n$矩阵， Y是$l \times n$矩阵, M是$l \times l$对称矩阵, σ是逐元素函数

例 4. [线性回归]：$l = \|Xw - y\|^2$, 求w的最小二乘估计，即求$\frac{dl}{dw}$的零点。其中y是m维列向量，X是$m \times n$矩阵，w是n维列向量。

信息论

信息是对不确定性的消除。

自信息：$I(a_i) = -\log(a_i)$

性质：设 a_1, a_2 为两个随机事件

（1）若 $P(a_1) \geq P(a_2)$，则 $I(a_1) \leq I(a_2)$

（2）若 $P(a_1) = 1$，则 $I(a_1) = 0$

（3）若 $P(a_1) = 0$，则 $I(a_1) = \infty$

（4）若 a_1, a_2 为独立事件，则 $I(a_1, a_2) = I(a_1) + I(a_2)$

信息论

信息熵是指随机系统的总体信息量，用所有随机事件自信息的统计平均来表示。

$$H(X)=f(p_1,p_2,\ldots,p_N)=-\sum_{x\in X}p(x)\log\ p(x)$$

性质：

1、对称性：$f(p_1,p_2,\ldots,p_N)=f(p_{k(1)},p_{k(2)},\ldots,p_{k(N)})$

2、非负性：$H(X)\geq 0$

3、可加性：$H(X,Y)=H(X)+H(Y|X)$

4、条件减少熵：$H(X|Y)\leq H(X)$

5、最大离散熵定理：$f(p_1,p_2,\ldots,p_N)\leq f(\frac{1}{N},\frac{1}{N},\ldots,\frac{1}{N})=\log N=\log|X|$

信息论

联合熵：一对离散随机变量 (X,Y) 的联合熵定义为：

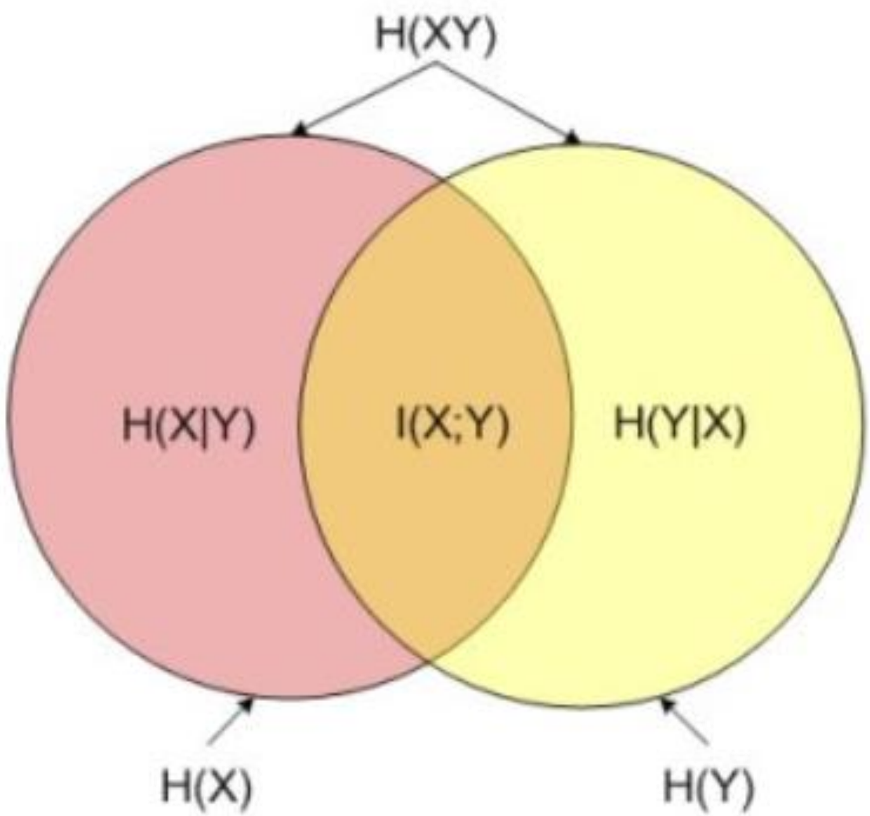
$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

条件熵：$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x)$$

信息论

互信息： $I(X;Y) = H(X) - H(X|Y)$ 或 $I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$



I(X;Y)=H(X)-H(X|Y)

I(X;Y)=H(Y)-H(Y|X)

I(X;Y)=H(X)+H(Y)-H(XY)

信息论

相对熵（KL散度）：

设$p(x)$、$q(x)$是关于随机变量X的两个概率分布，则p相对于q的相对熵是:

$$D_{KL}(p\|q)=\sum_{x}p(x)\log\frac{p(x)}{q(x)}$$

性质：1、如果 $p(x)$和 $q(x)$ 两个分布相同，那么相对熵等于0

2、$D_{KL}(p\|q)\neq D_{KL}(q\|p)$

3、$D_{KL}(p\|q)\geq 0$

信息论

交叉熵（Cross entropy）：

设$p(x)$、$q(x)$是关于随机变量X的两个概率分布，使用分布 $q(x)$ 表示目标分布 $p(x)$ 的困难程度：

$$H(p,q) = -\sum_i p(x_i)\log q(x_i)$$

性质：1、$D_{KL}(p,q) = H(p,q) - H(p)$

信息论

在机器学习中，

（1） 希望学到的模型的分布和真实分布一致：$P(model) \simeq P(real)$

（2）但是真实分布不可知，假设训练数据时从真实数据中独立同分布采样的：$P(train) \simeq P(real)$

（3） 因此，我们希望学到的模型分布至少和训练数据的分布一致：$P(train) \simeq P(model)$

交叉熵可以用来计算学习模型分布与训练分布之间的差异

$$L(Y, P(Y|\mathbf{X})) = -\mathbf{log}\, P(Y|\mathbf{X}) = -\frac{1}{N}\sum_{i=1}^{N}(1-y_i)\log\ (1-p(\mathbf{x_i})) - y_i \log p(\mathbf{x}_i)$$

THE END