机器学习白板推导系列笔记(2.3~2.7)

2. 数学基础

2.3. 多维高斯分布

多维高斯分布的概率密度函数$\boldsymbol{X} \sim N(\mu, \Sigma)$表达式为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{10}$$

其中p为随机变量的维度，$x \in \mathbb{R}^p$，Σ为协方差矩阵

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}_{p\times p} \tag{11}$$

注 一般的Σ是半正定的(对称的)，但在本节的讨论中，假设Σ是正定的(即$\lambda > 0$)

$$\sigma_{ij} = E[x_i x_j] - E[x_i]E[x_j] \tag{12}$$

可以看出$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$是一个二次型，下面对其进行分析，上式也称为向量$\boldsymbol{x}$与$\boldsymbol{\mu}$之间的马氏距离. 特别的，当$\Sigma = I$时，马氏距离等于欧氏距离. 协方差矩阵$\Sigma$是实对称矩阵，那么一定可以存在正交矩阵$U$，使得

$$\Sigma = U\Lambda U^T = U\Lambda U^{-1} \tag{13}$$

所以协方差矩阵可进行对角化(**证明见附录**)，即$\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$，设$U = (u_1, u_2, \cdots, u_p)_{p\times p}$($u_i$为特征列向量)，则

$$\begin{aligned}
\Sigma &= U\Lambda U^T \\
&= (u_1, u_2, \cdots, u_p)\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix} \\
&= (u_1\lambda_1, u_2\lambda_2, \cdots, u_p\lambda_p)\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix} = \sum_{i=1}^{p} u_i\lambda_i u_i^T
\end{aligned}$$

那么$\Sigma^{-1} = (U\Lambda U^T)^{-1} = (U^T)^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T$(正交矩阵性质)，又因为$\Lambda^{-1}$的特征值为$\frac{1}{\lambda_i}(i = 1,2,\cdots,p)$即

$$\Sigma^{-1}=\sum_{i=1}^{p} u_i \frac{1}{\lambda_i} u_i^T \tag{14}$$

将式(14)代入上方的二次型$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$，并将等式记为Δ得

$$\begin{aligned}\Delta &= (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^T\sum_{i=1}^{p}\left(u_i\frac{1}{\lambda_i}u_i^T\right)(\boldsymbol{x}-\boldsymbol{\mu})\end{aligned}$$

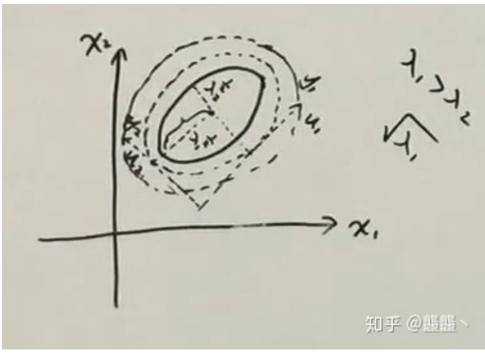
设$(\boldsymbol{x}-\boldsymbol{\mu})_i^T u_i = y_i$，上式可化为

$$\begin{aligned}\Delta &= \sum_{i=1}^{p}\left(y_i\frac{1}{\lambda_i}y_i^T\right) \\ &= \sum_{i=1}^{p}\left(\frac{y_i^2}{\lambda_i}\right)\end{aligned}$$

可以看出y是$x-\mu$在特征向量上的投影长度，设$p=2$(二维)则

$$\Delta=\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}$$

这是个简单的椭圆方程，令Δ取不同的值，可以得到不同的椭圆方程，这些曲线构成了二维高斯分布概率密度函数的等高线，如下图所示.



2.4. 高斯分布的局限性

主要体现在两个方面，一是在参数量上，另一个是高斯分布本身存在问题.

首先来看参数量，式(10)的参数为协方差矩阵Σ，因为其是对角矩阵，因此参数的个数由$p^2\rightarrow\frac{p(p+1)}{2}$，复杂度为$O(p^2)$，当维度$p$增加时，计算量迅速上升，解决办法是将协方差矩阵简化为对角矩阵等.

高斯分布本身存在的问题是有些数据不能用其很好的表示，因此在 GMM 中提出了混合模型，使用多个高斯分布进行混合.

2.5. 多维高斯分布的边缘分布和条件分布

已知随机变量$X\sim\mathcal{N}(\mu,\Sigma)$，且$x\in\mathbb{R}^p$，设将$x$分为$m$维的$x_a$和$n$维的$x_b(m+n=p)$那么

$$\mathrm{x}=\begin{pmatrix}x_a\\x_b\end{pmatrix}\quad \mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}\quad \Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix} \tag{15}$$

接下来要求边缘分布$p(x_a)$以及条件分布$p(x_b|x_a)$($p(x_b)$和$p(x_a|x_b)$由对称性就可以求出).在此之前，需要介绍一个**定理**

定理 若一随机变量$X\sim\mathcal{N}(\mu,\Sigma)$，另一随机变量$y=Ax+b$，那么$Y\sim\mathcal{N}(A\mu+b,A\Sigma A^T)$，证明省略

那么就可以对x_a做如下变换

$$x_a=\underbrace{(I_m\ 0)}_{A}\underbrace{\begin{pmatrix}x_a\\x_b\end{pmatrix}}_{x}+\underbrace{0}_{b}$$

先求边缘分布$p(x_a)$，根据定理很容易得出$E[x_a]=(I_m\ 0)\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}=\mu_a$，$Var[x_a]=(I_m\ 0)\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}I_m\\0\end{pmatrix}=\Sigma_{aa}$，因此

$$X_a\sim\mathcal{N}(\mu_a,\Sigma_{aa}) \tag{16}$$

下面求条件分布$p(x_b|x_a)$，计算较为复杂且推导带有很多技巧性，为了推导方便先做以下几个符号假设

$$\begin{aligned}x_{b\cdot a}&=x_b-\Sigma_{ba}\Sigma_{aa}^{-1}x_a\\\mu_{b\cdot a}&=\mu_b-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a\\\Sigma_{bb\cdot a}&=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\end{aligned} \tag{17}$$

注 上式并没有特殊的含义，纯粹是前人为了推导出条件分布函数根据经验构造的

下面先求$x_{b\cdot a}$的概率密度函数，对$x_{b\cdot a}$进行适当的变换

$$x_{b\cdot a}=(-\Sigma_{ba}\Sigma_{aa}^{-1}\ I)\begin{pmatrix}x_a\\x_b\end{pmatrix}$$

由定理易得$E[x_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\ I)\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}=\mu_b-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a=\mu_{b\cdot a}$，$Var[x_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\ I)\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}(-\Sigma_{ba}\Sigma_{aa}^{-1})^T\\I\end{pmatrix}=(-\Sigma_{ba}\Sigma_{aa}^{-1}\ I)\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}-\Sigma_{aa}^{-1}\Sigma_{ba}^T\\I\end{pmatrix}=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}=\Sigma_{bb\cdot a}$. 因此

$$X_{a\cdot b}\sim\mathcal{N}(\mu_{b\cdot a},\Sigma_{bb\cdot a}) \tag{18}$$

得到式(18)后，就可以利用式(17)中的 1 式得$x_b=x_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}x_a$. 那么$p(x_b|x_a)=p(x_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a)$，接下来证明随机变量$x_{b\cdot a}$和$x_a$相互独立，根据高斯随机变量的性质，相互独立和不相关等价，因此只需要证明$Cov(x_{b\cdot a},x_a)=E[x_{b\cdot a}x_a]-E[x_{b\cdot a}]E[x_a]=0$.

$$
\begin{aligned}
E[x_{b\cdot a}x_a]-E[x_{b\cdot a}]E[x_a] &= E[(x_b-\Sigma_{ba}\Sigma_{aa}^{-1}x_a)x_a]-E[X_a]E[x_b-\Sigma_{ba}\Sigma_{aa}^{-1}x_a] \\
&= E[x_a x_b]-\Sigma_{ba}\Sigma_{aa}^{-1}E[x_a^2]-E[x_a]E[x_b]+\Sigma_{ba}\Sigma_{aa}^{-1}E^2[x_a] \\
&= E[x_a x_b]-E[x_a]E[x_b]-\Sigma_{ba}\Sigma_{aa}^{-1}\{E[x_a^2]-E^2[x_a]\} \\
&= \Sigma_{ba}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{aa} \\
&= 0
\end{aligned}
$$

因此$p(x_b|x_a)=p(x_b)$，那么由定理可知$E[x_b|x_a]=\mu_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}x_a$，$Var[x_b|x_a]=\Sigma_{bb\cdot a}$

$$x_b|x_a \sim \mathcal{N}(\mu_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}x_a,\Sigma_{bb\cdot a}) \tag{19}$$

2.6. 多维高斯分布求联合概率分布

本节跟上一节不同，是已知边缘概率分布和条件概率分布求联合概率分布．已知$p(x)=\mathcal{N}(\mu,\Lambda^{-1})$，$p(y|x)=\mathcal{N}(Ax+b,L^{-1})$，求$p(y)$和$p(x|y)$，($\Lambda^{-1}$为$precision\ matrix=(covariance\ matrix)^{-1}$).

第一步先求$p(y)$，设$y=Ax+b+\epsilon$且$\epsilon\sim\mathcal{N}(0,L^{-1})$，与$x$相互独立.$E[y]=E[Ax+b+\epsilon]=E[Ax+b]+E[\epsilon]=A\mu+b$，$Var[y]=Var[Ax+b+\epsilon]=Var[Ax+b]+Var[\epsilon]=A\Lambda^{-1}A^T+L^{-1}$，因此

$$Y\sim\mathcal{N}(A\mu+b,A\Lambda^{-1}A^T+L^{-1}) \tag{20}$$

第二步求$p(x|y)$，构造新的随机变量$z=\begin{pmatrix}x\\y\end{pmatrix}$，很容易可以得到$E[z]=\begin{pmatrix}\mu\\A\mu+b\end{pmatrix}$，$Var[z]=\begin{pmatrix}Cov(x,x) & Cov(x,y)\\Cov(y,x) & Cov(y,y)\end{pmatrix}=\begin{pmatrix}\Lambda^{-1} & Cov(x,y)\\Cov(y,x) & L^{-1}+A\Lambda^{-1}A^T\end{pmatrix}$，那么

$$
\begin{aligned}
Cov(x,y) &= E[(x-E[x])\cdot(y-E[y])^T] \\
&= E[(x-\mu)\cdot(Ax+b+\epsilon-A\mu-b)^T] \\
&= E[(x-\mu)\cdot(Ax-A\mu+\epsilon)^T] \\
&= E[(x-\mu)\cdot(Ax-A\mu)^T+(x-\mu)\cdot\epsilon^T] \\
&= E[(x-\mu)(x-\mu)^T A^T]+E[(x-\mu)\epsilon^T]
\end{aligned}
$$

又因为ϵ与x相互独立，因此$E[(x-\mu)\epsilon^T]=E[x-\mu]\cdot E[\epsilon^T]=0$，$E[(x-\mu)(x-\mu)^T A^T]=E[(x-\mu)(x-\mu)^T]A^T=\Lambda^{-1}A^T$，由对称性得$Cov(y,x)=A\Lambda^{-1}$，因此

$$Z\sim\mathcal{N}\left(\begin{pmatrix}\mu\\A\mu+b\end{pmatrix},\begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T\\A\Lambda^{-1} & L^{-1}+A\Lambda^{-1}A^T\end{pmatrix}\right) \tag{21}$$

有了联合概率分布，由式(19)得

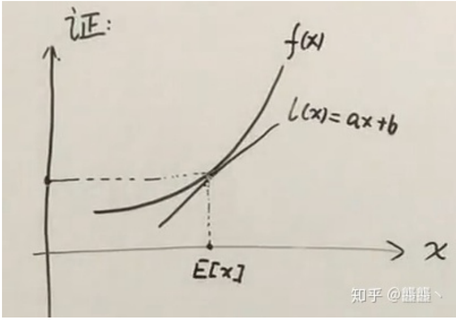
$$x|y\sim N(\mu+\Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}(y-A\mu-b),\Lambda^{-1}-\Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1}) \tag{22}$$

2.7. 杰森不等式(Jensen's Inequality)

杰森不等式经常被用于机器学习的推导中，在数学上的描述是若一个函数$f(x)$是凸函数，那么

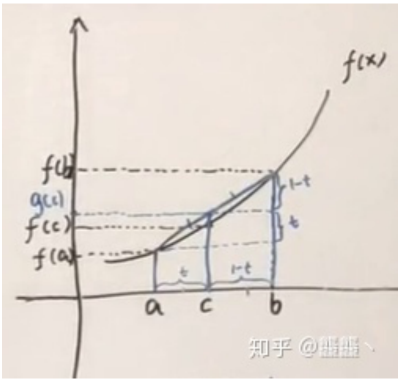
$$E[f(x)] \ge f(E[x]) \tag{23}$$

证明上式的方法有许多，下面采用最简洁的构造法



在数轴上取特定点$E[x]$并做切线$l(x) = ax + b$，因此$f(E[x]) = l(E[x]) = aE[x] + b$，又因为$f(x)$是凸函数对于任意$x$都有$f(x) \ge l(x)$，对上不等式两边取期望得$E[f(x)] \ge E[l(x)] = E[ax + b] = aE[x] + b = f(E[x])$，证毕.

但在机器学习中一般不使用式(23)，通常使用其变式，下面进行推导



如上图所示，在x轴上任取两点$a, b(a > b)$，并且$c \in [a, b]$，那么$c = a + \mu(b - a)(\mu \in [0,1])$，令$t = 1 - \mu \in [0,1]$上式可以化为$c = ta + (1 - t)b$，然后连接$f(a)$与$f(b)$两点作为新的直线$g(x)$，因为 f(x)为凸函数，所以$g(c) \ge f(c)$，且

$$c = ta + (1 - t)b \Rightarrow \frac{c - a}{b - c} = \frac{1 - t}{t}$$

注 上图中标注有误

由初中相似三角形知识可知

$$\frac{f(b) - g(c)}{g(c) - f(a)} = \frac{t}{1 - t} \Rightarrow g(c) = tf(a) + (1 - t)f(b)$$

那么得到一个非常重要且常用的式子

$$tf(a) + (1 - t)f(b) \ge f(ta + (1 - t)b) \tag{24}$$

附录

1) 特征值 & 特征向量

设方阵$A_{n\times n}$，如果存在常数λ和n维向量$x_{n\times 1}$使得

$$Ax=\lambda x$$

那么称x,λ为A的特征向量和特征值. 该式子可以理解为向量x在A的变换下得到λx，只是大小做了伸缩，特征向量提供了复杂的矩阵乘法到简单的数乘之间的转换. 上面证明中用到了该性质$\mathrm{A}^{-1}\mathrm{x}=\frac{1}{\lambda}\mathrm{x}$，下面简单进行证明：已知$Ax=\lambda x$，那么$\frac{1}{\lambda}Ax=x\Rightarrow\frac{1}{\lambda}x=A^{-1}x$.

同时$det(A)=\prod_i\lambda_i$，$tr(A)=\sum_i\lambda_i=\sum_i a_{ii}$.

2) 矩阵对角化

先介绍一下矩阵相似的概念，若存在可逆矩阵P，使得$P^{-1}AP=B$，那么称矩阵A和B相似，记为$A\sim B$，相似矩阵有很多性质，其中最重要的是两者的行列式，秩相等，且有相同的特征值. 特别的，当矩阵B为对角矩阵$\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$时，称矩阵$A$可被对角化，即$P^{-1}AP=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$(其充要条件是矩阵$A$有$n$个线性无关的特征向量).

下面介绍正交矩阵与实对称矩阵的概念，若方阵$AA^T=A^TA=I$，则称A为正交矩阵，易得$A^T=A^{-1}$. 若方阵A是实矩阵，且$a_{ij}=a_{ji}$，即$A^T=A$那么称A为实对称矩阵. 实对称矩阵有很特殊的性质，其特征值为实数且一定存在正交矩阵P使得$P^TAP=P^{-1}AP=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$，即

$$A=P\Lambda P^{-1}=P\Lambda P^T$$

对应于式(13)