Introduction

对概率的诠释有两大学派，一种是频率派另一种是贝叶斯派。后面我们对观测集采用下面记号：

$$X_{N\times p}=(x_1,x_2,\cdots,x_N)^T,x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$$

这个记号表示有 N 个样本，每个样本都是 p 维向量。其中每个观测都是由 $p(x|\theta)$ 生成的。

频率派的观点

$p(x|\theta)$中的 θ 是一个常量。对于 N 个观测来说观测集的概率为 $p(X|\theta)\mathop{=}\limits *{iid}\prod\limits _{i=1}^{N}p(x*{i}|\theta))$ 。为了求 θ 的大小，我们采用最大对数似然MLE的方法：

$$\theta_{MLE}=\underset{\theta}{argmax}\log p(X|\theta)\underset{iid}{=}\underset{\theta}{argmax}\sum_{i=1}^{N}\log p(x_i|\theta)$$

贝叶斯派的观点

贝叶斯派认为 $p(x|\theta)$ 中的 θ 不是一个常量。这个 θ 满足一个预设的先验的分布 $\theta\sim p(\theta)$ 。于是根据贝叶斯定理依赖观测集参数的后验可以写成：

$$p(\theta|X)=\frac{p(X|\theta)\cdot p(\theta)}{p(X)}=\frac{p(X|\theta)\cdot p(\theta)}{\int\limits_{\theta}p(X|\theta)\cdot p(\theta)d\theta}$$

为了求 θ 的值，我们要最大化这个参数后验MAP：

$$\theta_{MAP}=\underset{\theta}{argmax}p(\theta|X)=\underset{\theta}{argmax}p(X|\theta)\cdot p(\theta)$$

其中第二个等号是由于分母和 θ 没有关系。求解这个 θ 值后计算$\frac{p(X|\theta)\cdot p(\theta)}{\int\limits _{\theta}p(X|\theta)\cdot p(\theta)d\theta}$ ，就得到了参数的后验概率。其中 $p(X|\theta)$ 叫似然，是我们的模型分布。得到了参数的后验分布后，我们可以将这个分布用于预测贝叶斯预测：

$$p(x_{new}|X)=\int\limits_{\theta}p(x_{new}|\theta)\cdot p(\theta|X)d\theta$$

其中积分中的被乘数是模型，乘数是后验分布。

小结

频率派和贝叶斯派分别给出了一系列的机器学习算法。频率派的观点导出了一系列的统计机器学习算法而贝叶斯派导出了概率图理论。在应用频率派的 MLE 方法时最优化理论占有重要地位。而贝叶斯派的算法无论是后验概率的建模还是应用这个后验进行推断时积分占有重要地位。因此采样积分方法如 MCMC 有很多应用。

MathBasics

高斯分布

一维情况 MLE

高斯分布在机器学习中占有举足轻重的作用。在 MLE 方法中：

$$\theta = (\mu, \Sigma) = (\mu, \sigma^2), \theta_{MLE} = \underset{\theta}{argmax} \log p(X|\theta) \underset{iid}{=} \underset{\theta}{argmax} \sum_{i=1}^{N} \log p(x_i|\theta)$$

一般地，高斯分布的概率密度函数PDF写为：

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

带入 MLE 中我们考虑一维的情况

$$\log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x_i-\mu)^2/2\sigma^2\right)$$

首先对 μ 的极值可以得到：

$$\mu_{MLE} = \underset{\mu}{argmax} \log p(X|\theta) = \underset{\mu}{argmax} \sum_{i=1}^{N} (x_i - \mu)^2$$

于是：

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} (x_i - \mu)^2 = 0 \longrightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

其次对 θ 中的另一个参数 σ，有：

$$\begin{aligned} \sigma_{MLE} = \underset{\sigma}{argmax} \log p(X|\theta) &= \underset{\sigma}{argmax} \sum_{i=1}^{N} [-\log \sigma - \frac{1}{2\sigma^2}(x_i - \mu)^2] \\ &= \underset{\sigma}{argmin} \sum_{i=1}^{N} [\log \sigma + \frac{1}{2\sigma^2}(x_i - \mu)^2] \end{aligned}$$

于是：

$$\frac{\partial}{\partial \sigma} \sum_{i=1}^{N} [\log \sigma + \frac{1}{2\sigma^2}(x_i - \mu)^2] = 0 \longrightarrow \sigma^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

值得注意的是，上面的推导中，首先对 μ 求 MLE， 然后利用这个结果求 $\sigma*{MLE}$，因此可以预期的是对数据集求期望时 $\mathbb{E}_{\mathcal{D}}[\mu_{MLE}]$ 是无偏差的：

$$\mathbb{E}_{\mathcal{D}}[\mu_{MLE}] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N} \sum_{i=1}^{N} x_i] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}[x_i] = \mu$$

但是当对 $\sigma*{MLE}$ 求 期望的时候由于使用了单个数据集的 μ_{MLE}，因此对所有数据集求期望的时候我们会发现 σ_{MLE} 是 有偏的：

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[\sigma^2_{MLE}] &= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x_i^2-2x_i\mu_{MLE}+\mu^2_{MLE}) \\ &= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2_{MLE}] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2+\mu^2-\mu^2_{MLE}] \\ &= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2]-\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}-\mu^2] = \sigma^2-(\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}]-\mu^2) \\ &= \sigma^2-(\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}]-\mathbb{E}^2_{\mathcal{D}}[\mu_{MLE}]) = \sigma^2-Var[\mu_{MLE}] \\ &= \sigma^2-Var[\frac{1}{N}\sum_{i=1}^{N}x_i] = \sigma^2-\frac{1}{N^2}\sum_{i=1}^{N}Var[x_i] = \frac{N-1}{N}\sigma^2\end{aligned}$$

所以：

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$$

多维情况

多维高斯分布表达式为：

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

其中 $x,\mu\in\mathbb{R}^{p},\Sigma\in\mathbb{R}^{p\times p}$ ，Σ 为协方差矩阵，一般而言也是半正定矩阵。这里我们只考虑正定矩阵。首先我们处理指数上的数字，指数上的数字可以记为 x 和 μ 之间的马氏距离。对于对称的协方差矩阵可进行特征值分解，$\Sigma=U\Lambda U^{T}=(u*{1},u*{2},\cdots,u*{p})diag(\lambda*{i})(u*{1},u*{2},\cdots,u*{p})^{T}=\sum\limits_{i=1}^{p}u*{i}\lambda*{i}u*{i}^{T}$ ，于是：

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

$$\Delta = (x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i=1}^{p}(x-\mu)^Tu_i\frac{1}{\lambda_i}u_i^T(x-\mu) = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}$$

我们注意到 $y*{i}$ 是 $x-\mu$ 在特征向量 $u*{i}$ 上的投影长度，因此上式子就是 Δ 取不同值时的同心椭圆。

下面我们看多维高斯模型在实际应用时的两个问题

1. 参数 Σ,μ 的自由度为 $O(p^{2})$ 对于维度很高的数据其自由度太高。解决方案：高自由度的来源是 Σ 有 $\frac{p(p+1)}{2}$ 个自由参数，可以假设其是对角矩阵，甚至在各向同性假设中假设其对角线上的元素都相同。前一种的算法有 Factor Analysis，后一种有概率 PCA(p-PCA) 。
2. 第二个问题是单个高斯分布是单峰的，对有多个峰的数据分布不能得到好的结果。解决方案：高斯混合GMM 模型。

下面对多维高斯分布的常用定理进行介绍。

我们记 $x=(x_1, x_2,\cdots,x_p)^T=(x*{a,m\times 1}, x*{b,n\times1})^T,\mu=(\mu*{a,m\times1}, \mu*{b,n\times1}),\Sigma=\begin{pmatrix}\Sigma*{aa}&\Sigma*{ab}\\\Sigma*{ba}&\Sigma*{bb}\end{pmatrix}$，已知 $x\sim\mathcal{N}(\mu,\Sigma)$。

首先是一个高斯分布的定理：

> 定理：已知 $x\sim\mathcal{N}(\mu,\Sigma), y\sim Ax+b$，那么 $y\sim\mathcal{N}(A\mu+b, A\Sigma A^T)$。
>
> 证明：$\mathbb{E}[y]=\mathbb{E}[Ax+b]=A\mathbb{E}[x]+b=A\mu+b$，$Var[y]=Var[Ax+b]=Var[Ax]=A\cdot Var[x]\cdot A^T$。

下面利用这个定理得到 $p(x_a),p(x_b),p(x_a|x_b),p(x_b|x_a)$ 这四个量。

1. $x_a=\begin{pmatrix}\mathbb{I}*{m\times m}&\mathbb{O}{m\times n}\end{pmatrix}\begin{pmatrix}x_a\x_b\end{pmatrix}$，代入定理中得到：

$$\mathbb{E}[x_a]=(\mathbb{I}\quad\mathbb{O})\begin{pmatrix}\mu_a\\mu_b\end{pmatrix}=\mu_a$$

$$Var[x_a]=(\mathbb{I}\quad\mathbb{O})\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}\mathbb{I}\\mathbb{O}\end{pmatrix}=\Sigma_{aa}$$

   所以 $x_a\sim\mathcal{N}(\mu_a,\Sigma_{aa})$。

2. 同样的，$x_b\sim\mathcal{N}(\mu_b,\Sigma_{bb})$。

3. 对于两个条件概率，我们引入三个量：

$$x_{b\cdot a}=x_b-\Sigma_{ba}\Sigma_{aa}^{-1}x_a$$

$$\mu_{b\cdot a}=\mu_b-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a$$

$$\Sigma_{bb\cdot a}=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$$

   特别的，最后一个式子叫做 Σ_{bb} 的 Schur Complementary。可以看到：

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

   所以：

$$\mathbb{E}[x_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\quad\mathbb{I}_{n\times n})\begin{pmatrix}\mu_a\\mu_b\end{pmatrix}=\mu_{b\cdot a}$$

$$Var[x_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\quad\mathbb{I}_{n\times n})\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}-\Sigma_{aa}^{-1}\Sigma_{ba}^T\\mathbb{I}_{n\times n}\end{pmatrix}=\Sigma_{bb\cdot a}$$

   利用这三个量可以得到 $x_b=x*{b\cdot a}+\Sigma*{ba}\Sigma_{aa}^{-1}x_a$。因此：

$$\mathbb{E}[x_b|x_a]=\mu_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}x_a$$

$$Var[x_b|x_a]=\Sigma_{bb\cdot a}$$

   这里同样用到了定理。

4. 同样：

$$x_{a\cdot b}=x_a-\Sigma_{ab}\Sigma_{bb}^{-1}x_b$$

$$\mu_{a\cdot b}=\mu_a-\Sigma_{ab}\Sigma_{bb}^{-1}\mu_b$$

$$\Sigma_{aa\cdot b}=\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

   所以：

$$\mathbb{E}[x_a|x_b]=\mu_{a\cdot b}+\Sigma_{ab}\Sigma_{bb}^{-1}x_b$$

$$Var[x_a|x_b]=\Sigma_{aa\cdot b}$$

下面利用上边四个量，求解线性模型：

> 已知：$p(x)=\mathcal{N}(\mu,\Lambda^{-1}),p(y|x)=\mathcal{N}(Ax+b,L^{-1})$，求解：$p(y),p(x|y)$。
>
> 解：令 $y=Ax+b+\epsilon,\epsilon\sim\mathcal{N}(0,L^{-1})$，所以 $\mathbb{E}[y]=\mathbb{E}[Ax+b+\epsilon]=A\mu+b$，$Var[y]=A \Lambda^{-1}A^T+L^{-1}$，因此：

$$p(y) = \mathcal{N}(A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

引入 $z=\begin{pmatrix}x\y\end{pmatrix}$，我们可以得到 $Cov[x,y]=\mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])^T]$。对于这个协方差可以直接计算：

$$Cov(x, y) = \mathbb{E}[(x - \mu)(Ax - A\mu + \epsilon)^T] = \mathbb{E}[(x - \mu)(x - \mu)^T A^T] = Var[x]A^T = \Lambda^{-1}A^T$$

注意到协方差矩阵的对称性，所以 $p(z)=\mathcal{N}\begin{pmatrix}\mu\A\mu+b\end{pmatrix},\begin{pmatrix}\Lambda^{-1}&\Lambda^{-1}A^T\A\Lambda^{-1}&L^{-1}+A\Lambda^{-1}A^T\end{pmatrix})$。根据之前的公式，我们可以得到：

$$\mathbb{E}[x|y] = \mu + \Lambda^{-1}A^T(L^{-1} + A\Lambda^{-1}A^T)^{-1}(y - A\mu - b)$$

$$Var[x|y] = \Lambda^{-1} - \Lambda^{-1}A^T(L^{-1} + A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1}$$