MLAPP 读书笔记 - 04 高斯模型 (Gaussian models)

> A Chinese Notes of MLAPP，MLAPP 中文笔记项目
> https://zhuanlan.zhihu.com/python-kivy

记笔记的人：cycleuser

2018年05月16日10:49:49

4.1 简介

本章要讲的是多元高斯分布(multivariate Gaussian),或者多元正态分布(multivariate normal ,缩写为MVN)模型,这个分布是对于连续变量的联合概率密度函数建模来说最广泛的模型了.未来要学习的其他很多模型也都是以此为基础的.

然而很不幸的是,本章所要求的数学水平也是比很多其他章节都要高的.具体来说是严重依赖线性代数和矩阵积分.要应对高维数据,这是必须付出的代价.初学者可以跳过标记了星号的章.另外本章有很多等式,其中特别重要的用方框框了起来.

4.1.1 记号

这里先说几句关于记号的问题.向量用小写字母粗体表示,比如**x**.矩阵用大写字母粗体表示,比如**X**.大写字母加下标表示矩阵中的项,比如X_{ij}.
所有向量都假设为列向量(column vector),除非特别说明是行向量.通过堆叠(stack)D个标量(scalar)得到的类向量记作$[x_1,...,x_D]$.与之类似,如果写**x**=$[x_1,...,x_D]$,那么等号左侧就是一个高列向量(tall column vector),意思就是沿行堆叠x_i,一般写作**x**=$(x_1^T,...,x_D^T)^T$,不过这样很丑哈.如果写**X**=$[x_1,...,x_D]$,等号左边的是矩阵,意思就是沿列堆叠x_i,建立一个矩阵.

4.1.2 基础知识

回想一下本书2.5.2中关于D维度下的多元正态分布(MVN)概率密度函数(pdf)的定义,如下所示:

$$N(x|\mu,\Sigma)\stackrel{*}{=}\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]$$(4.1 重要公式)

此处参考原书图4.1

指数函数内部的是一个数据向量**x**和均值向量$\boldsymbol{\mu}$ 之间的马氏距离(马哈拉诺比斯距离,Mahalanobis distance).对Σ进行特征分解(eigendecomposition)有助于更好理解这个量.$\Sigma = U\Lambda U^T$,其中的U是标准正交矩阵(orthonormal matrix),满足$U^TU = I$,而Λ是特征值组成的对角矩阵.经过特征分解就得到了:

$$\Sigma^{-1} = U^{-T}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{D}\frac{1}{\lambda_i}u_iu_i^T \quad (4.2)$$

上式中的u_i是U的第i列,包含了第i个特征向量(eigenvector).因此就可以把马氏距离写作:

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T(\sum_{i=1}^{D}\frac{1}{\lambda_i}u_iu_i^T)(x-\mu) \quad (4.3)$$

$$= \sum_{i=1}^{D}\frac{1}{\lambda_i}(x-\mu)^Tu_iu_i^T(x-\mu) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \quad (4.4)$$

上式中的$y_i \overset{*}{=} u_i^T(x-\mu)$.二维椭圆方程为:

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1 \quad (4.5)$$

因此可以发现高斯分布的概率密度的等值线沿着椭圆形,如图4.1所示.特征向量决定了椭圆的方向,特征值决定了椭圆的形态即宽窄比.

一般来说我们将马氏距离(Mahalanobis distance)看作是对应着变换后坐标系中的欧氏距离(Euclidean distance),平移μ,旋转U.

4.1.3 多元正态分布(MVN)的最大似然估计(MLE)

接下来说的是使用最大似然估计(MLE)来估计多元正态分布(MVN)的参数.在后面的章节里面还会说道用贝叶斯推断来估计参数,能够减轻过拟合,并且能对估计值的置信度提供度量.

定理4.1.1(MVN的MLE)

如果有N个独立同分布样本符合正态分布,即$x_i \sim N(\mu, \Sigma)$,则对参数的最大似然估计为:

$$\hat{\mu}_{mle} = \frac{1}{N}\sum_{i=1}^{N}x_i \overset{*}{=} \bar{x} \quad (4.6)$$

$$\hat{\Sigma}_{mle} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T = \frac{1}{N}(\sum_{i=1}^{N}x_ix_i^T) - \bar{x}\bar{x}^T \quad (4.7)$$

也就是MLE就是经验均值(empirical mean)和经验协方差(empirical covariance).在单变量情况下结果就很熟悉了:

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \bar{x} \quad (4.8)$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - x)^2 = (\frac{1}{N}\sum_i x_i^2) - \bar{x}^2 \text{(4.9)}$$

4.1.3.1 证明

要证明上面的结果,需要一些矩阵代数的计算,这里总结一下.等式里面的**a**和**b**都是向量,**A**和**B**都是矩阵.记号$tr(A)$表示的是矩阵的迹(trace),是其对角项求和,即$tr(A) = \sum_i A_{ii}$.

$$\frac{\partial(b^T a)}{\partial a} = b$$

$$\frac{\partial(a^T A a)}{\partial a} = (A + A^T)a$$

$$\frac{\partial}{\partial A} tr(BA) = B^T$$

$$\frac{\partial}{\partial A} \log|A| = A^{-T} \overset{*}{=} (A^{-1})^T$$

$$tr(ABC) = tr(CAB) = tr(BCA)$$

(4.10 重要公式)

上式中最后一个等式也叫做迹运算的循环置换属性(cyclic permutation property).利用这个性质可以推广出很多广泛应用的求迹运算技巧,对标量内积$x^T A x$就可以按照如下方式重新排序:

$x^T A x = tr(x^T A x) = tr(x x^T A) = tr(A x x^T)$(4.11)

证明过程

接下来要开始证明了,对数似然函数为:

$$l(\mu, \Sigma) = \log p(D|\mu, \Sigma) = \frac{N}{2}\log|\Lambda| - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T \Lambda (x_i - \mu) \text{(4.12)}$$

上式中$\Lambda = \Sigma^{-1}$,是精度矩阵(precision matrix)

然后进行一个替换(substitution)$y_i = x_i - \mu$,再利用微积分的链式法则:

$$\frac{\partial}{\partial \mu}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) = \frac{\partial}{\partial y_i} y_i^T \Sigma^{-1} y_i \frac{\partial y_i}{\partial \mu} \quad (4.13)$$

$$= -1(\Sigma_{-1} + \Sigma^{-T}) y_i \quad (4.14)$$

因此:

$$\frac{\partial}{\partial \mu} l(\mu.\Sigma) = -\frac{1}{2}\sum_{i=1}^{N} - 2\Sigma^{-1}(x_i - \mu) = \Sigma^{-1}\sum_{i=1}^{N}(x_i - \mu) = 0 \quad (4.15)$$

$$= -1(\Sigma_{-1} + \Sigma^{-T}) y_i \quad (4.16)$$

所以μ的最大似然估计(MLE)就是经验均值(empirical mean).

然后利用求迹运算技巧(trace-trick)来重写对Λ的对数似然函数:

$$l(\Lambda) = \frac{N}{2}\log|\Lambda| - \frac{1}{2}\sum_i tr[(x_i-\mu)(x_i-\mu)^T\Lambda] \quad (4.17)$$

$$= \frac{N}{2}\log|\Lambda| - \frac{1}{2}tr[S_\mu\Lambda] \quad (4.18)$$

$$(4.19)$$

上式中

$S_\mu \overset{*}{=} \sum_{i=1}^N (x_i-\mu)(x_i-\mu)^T$(4.20)

是以μ为中心的一个散布矩阵(scatter matrix).对上面的表达式关于Λ进行求导就得到了:

$$\frac{\partial l(\Lambda)}{\partial\Lambda} = \frac{N}{2}\Lambda^{-T} - \frac{1}{2}S_\mu^T = 0 \quad (4.21)$$

$$\Lambda^{-T} = \Lambda^{-1} = \Sigma = \frac{1}{N}S_\mu \quad (4.22)$$

因此有:

$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N (x_i-\mu)(x_i-\mu)^T$(4.23)

正好也就是以μ为中心的经验协方差矩阵(empirical covariance matrix).如果插入最大似然估计$\mu = \bar{x}$(因为所有参数都同时进行优化),就得到了协方差矩阵的最大似然估计的标准方程.

4.1.4 高斯分布最大熵推导(Maximum entropy derivation of the Gaussian)*

在本节,要证明的是多元高斯分布(multivariate Gaussian)是适合于有特定均值和协方差的具有最大熵的分布(参考本书9.2.6).这也是高斯分布广泛应用到一个原因,均值和协方差这两个矩(moments)一般我们都能通过数据来进行估计得到(注:一阶矩（期望）归零，二阶矩（方差）),所以我们就可以使用能捕获这这些特征的分布来建模,另外还要尽可能少做附加假设.

为了简单起见,假设均值为0.那么概率密度函数(pdf)就是:

$p(x) = \frac{1}{Z}\exp(-\frac{1}{2}x^T\Sigma^{-1}x)$(4.24)

如果定义$f_{ij}(x) = x_i x_j, \lambda_{ij} = \frac{1}{2}(\Sigma^{-1})_{ij} i,j \in \{1,\ldots,D\}$,就会发现这个和等式9.74形式完全一样.这个分布（使用自然底数求对数）的（微分）熵为:

$$h(N(\mu, \Sigma)) = \frac{1}{2}\ln[(2\pi e)^D |\Sigma|] \quad (4.25)$$

接下来要证明有确定的协方差Σ的情况下多元正态分布(MVN)在所有分布中有最大熵.

定理 4.1.2

设$q(x)$是任意的一个密度函数,满足$\int q(x)x_i x_j = \Sigma_{ij}$.设$p = N(0, \Sigma)$.那么$h(q) \leq h(p)$.

证明.(参考(Cover and Thomas 1991, p234)).
(注:KL是KL 散度(Kullback-Leibler divergence),也称相对熵(relative entropy),可以用来衡量p和q两个概率分布的差异性(dissimilarity).更多细节参考2.8.2.)

$$\begin{aligned} 0 \leq KL(q||p) &= \int q(x)\log\frac{q(x)}{p(x)}dx && (4.26)\\ &= -h(q) - \int q(x)\log p(x)dx && (4.27)\\ &\overset{*}{=} -h(q) - \int ps(x)\log p(x)dx && (4.28)\\ &= -h(q) + h(p) && (4.29) \end{aligned}$$

等式4.28那里的星号表示这一步是关键的,因为q和p对于由$\log p(x)$编码的二次形式产生相同的矩(moments).

4.2 高斯判别分析(Gaussian discriminant analysis)

多元正态分布的一个重要用途就是在生成分类器中定义类条件密度,也就是:
$$p(x|y = c, \theta) = N(x|\mu_c, \Sigma_c) \quad (4.30)$$

这样就得到了高斯判别分析,也缩写为GDA,不过这其实还是生成分类器(generative classifier），而并不是辨别式分类器（discriminative classifier）,这两者的区别参考本书8.6.如果\Sigma_c$$是对角矩阵,那这就等价于朴素贝叶斯分类器了.

此处参考原书图4.2

从等式2.13可以推导出来下面的决策规则,对一个特征向量进行分类:
$$\hat{y}(x) = \arg\max_c[\log p(y = c|\pi) + \log p(x|\theta_c)] \quad (4.31)$$

计算x 属于每一个类条件密度的概率的时候,测量的距离是x到每个类别中心的马氏距离(Mahalanobis distance).这也是一种最近邻质心分类器(nearest centroids classifier).

例如图4.2展示的就是二维下的两个高斯类条件密度,横纵坐标分别是身高和体重,包含了男女两类人.很明显身高体重这两个特征有相关性,就如同人们所想的,个子高的人更可能重.每个分类的椭圆都包含了95%的概率质量.如果对两类有一个均匀分布的先验,就可以用如下方式来对新的测试向

量进行分类:

$$\hat{y}(x) = \arg\max_c (x-\mu_c)^T \Sigma_c^{-1} (x-\mu_c) \quad (4.32)$$

4.2.1 二次判别分析(Quadratic discriminant analysis,QDA)

对类标签的后验如等式2.13所示.加入高斯密度定义后,可以对这个模型获得更进一步的理解:

$$p(y=c|x,\theta) = \frac{\pi_c |2\pi\Sigma_c|^{-1/2} \exp\left[-1/2(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)\right]}{\Sigma_{c'} \pi_{c'} |2\pi\Sigma_{c'}|^{-1/2} \exp\left[-1/2(x-\mu_{c'})^T \Sigma_{c'}^{-1}(x-\mu_{c'})\right]} \quad (4.33)$$

对此进行阈值处理(thresholding)就得到了一个x的二次函数(quadratic function).这个结果也叫做二次判别分析(quadratic discriminant analysis,缩写为QDA).图4.3所示的是二维平面中决策界线的范例.

此处参考原书图4.3

此处参考原书图4.4

4.2.2 线性判别分析(Linear discriminant analysis,LDA)

接下来考虑一种特殊情况,此事协方差矩阵为各类所共享(tied or shared),即$\Sigma_c = \Sigma$.这时候就可以把等式4.33简化成下面这样:

$$p(y=c|x,\theta) \propto \pi_c \exp[\mu_c^T \Sigma^{-1} x - \frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c] \quad (4.34)$$

$$= \exp[\mu_c^T \Sigma^{-1} x - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c] \exp[-\frac{1}{2} x^T \Sigma^{-1} x] \quad (4.35)$$

由于二次项$x^T \Sigma^{-1}$独立于类别c,所以可以抵消掉分子分母.如果定义了:

$$\gamma_c = -\frac{1}{2} \mu - c^T \Sigma^{-1} \mu_c + \log \pi_c \quad (4.36)$$

(4.37)

$$\beta_c = \Sigma^{-1} \mu_c$$

则有:

$$p(y=c|x,\theta) = \frac{e^{\beta_c^T + \gamma_c}}{\Sigma_{c'} e^{\beta_{c'}^T + \gamma_{c'}}} = S(\eta)_c \quad (4.38)$$

当$\eta = [\beta_1^T x + \gamma_1, \ldots, \beta_C^T x + \gamma_C]$的时候,$S$就是Softmax函数(softmax function,注:柔性最大函数,或称归一化指数函数),其定义如下:

$$S(\eta/T) = \frac{e^{\eta_c}}{\sum_{c'=1}^{C} e^{\eta_{c'}}} \quad (4.39)$$

Softmax函数如同其名中的Max所示,有点像最大函数.把每个η_c除以一个常数T,这个常数T叫做温度(temperature).然后让T趋于零,即$T \to 0$,则有:

$$S(\eta/T)_c = \begin{cases} 1.0 & \text{if } c = \arg\max_{c'} \eta_{c'} \\ 0.0 & \text{otherwise} \end{cases}$$

(4.40)

也就是说,在低温情况下,分布总体基本都出现在最高概率的状态下,而在高温下,分布会均匀分布于所有状态.参见图4.4以及其注解.这个概念来自统计物理性,通常称为玻尔兹曼分布(Boltzmann distribution),和Softmax函数的形式一样.

等式4.38的一个有趣性质是,如果取对数,就能得到一个关于x的线性函数,这是因为$x^T\Sigma^{-1}x$从分子分母中约掉了.这样两个类 c 和 c'之间的决策边界就是一条直线了.所以这种方法也叫做线性判别分析(linear discriminant analysis,缩写为LDA).可以按照如下方式来推导出这条直线的形式:

$$p(y = c \mid x, \theta) = p(y = c' \mid x, \theta) \tag{4.41}$$

$$\beta_c^T x + \gamma_c = \beta_{c'}^T x + \gamma_{c'} \tag{4.42}$$

$$x^T(\beta_{c'} - \beta) = \gamma_{c'} - \gamma_c \tag{4.43}$$

样例参考图4.5.

除了拟合一个线性判别分析(LDA)模型然后推导类后验之外,还有一种办法就是对某$C \times D$权重矩阵(weight matrix)W,直接拟合$p(y \mid x, W) = Cat(y \mid Wx)$.这叫做多类逻辑回归(multi-class logistic regression)或者多项逻辑回归(multinomial logistic regression).此类模型的更多细节将在本书8.2中讲解,两种方法的区别在本书8.6中有解释.

此处查看原书图4.5

此处查看原书图4.6

4.2.3 双类线性判别分析(Two-class LDA)

为了更好理解上面那些等式,咱们先考虑二值化分类的情况.这时候后验为:

$$p(y = 1 \mid x, \theta) = \frac{e^{\beta_1^T x + \gamma_1}}{e^{\beta_1^T x + \gamma_1} + e^{\beta_0^T x + \gamma_0}} \tag{4.44}$$

$$= \frac{1}{1 + e^{(\beta_0 - \beta_1))^T x + (\gamma_0 - \gamma_1)}} = sigm((\beta_1 - \beta_0)^T x + (\gamma_1 - \gamma_0)) \tag{4.45}$$

上式中的$sigm(\eta)$就是之前在等式1.10中提到的S型函数(sigmoid function).现在则有:

$$\gamma_1 - \gamma_0 = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log(\pi_1/\pi_0) \quad (4.46)$$

$$= -\frac{1}{2}(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 + \mu_0) + \log(\pi_1/\pi_0) \quad (4.47)$$

所以如果定义:

$$w = \beta_1 - \beta_0 = \Sigma^{-1}(\mu_1 - \mu_0) \quad (4.48)$$

$$x_0 = -\frac{1}{2}(\mu_1 + \mu_0) - (\mu_1 - \mu_0)\frac{\log(\pi_1/\pi_0)}{(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0)} \quad (4.49)$$

然后就有$w^T x_0 = -(\gamma_1 - \gamma_0)$,因此:
$p(y = 1 | x, \theta) = sigm(w^T(x - x_0))$ (4.50)

这个形式和逻辑回归(logistic regression)关系密切,对此将在本书8.2中讨论.所以最终的决策规则为:将x移动x_0,然后投影到线w上,看结果的正负号.
如果$\Sigma = \sigma^2 I$,那么w就是$\mu_1 - \mu_0$的方向.我们对点进行分类就要根据其投影距离μ_1和μ_0哪个更近.如图4.6所示.另外,如果$\pi_1 = \pi_0$,那么$x_0 = \frac{1}{2}(\mu_1 + \mu_0)$,正好在两个均值的中间位置.如果让$\pi_1 > \pi_0$,则$x_0$更接近$\mu_0$,所以图中所示线上更多位置属于类别1.反过来如果$\pi_1 < \pi_0$则边界右移.因此,可以看到类的先验$\pi_c$只是改变了决策阈值,而并没有改变总体的结合形态.类似情况也适用于多类情景.

w的大小决定了对数函数的陡峭程度,取决于均值相对于方差的平均分离程度.在心理学和信号检测理论中,通常定义一个叫做敏感度指数(sensitivity index,也称作 d-prime)的量,表示信号和背景噪声的可区别程度:

$$d' \triangleq \frac{\mu_1 - \mu_0}{\sigma} \quad (4.51)$$

上式中的μ_1是信号均值,μ_0是噪音均值,而σ是噪音的标准差.如果敏感度指数很大,那么就意味着信号更容易从噪音中提取出来.

4.2.4 对于判别分析(discriminant analysis)的最大似然估计(MLE)

现在来说说如何去拟合一个判别分析模型(discriminant analysis model).最简单的方法莫过于最大似然估计(maximum likelihood).对应的对数似然函数(log-likelihood)如下所示:

$$\log p(D | \theta) = [\sum_{i=1}^N \sum_{c=1}^C I(y_i = c)\log\pi_c] + \sum_{c=1}^C [\sum_{i:y_i=c} \log N(x | \mu_c, \Sigma_c)] \quad (4.52)$$

显然这个式子可以因式分解成一个含有π的项,以及对应每个μ_c, Σ_c的C个项.因此可以分开对这些参数进行估计.对于类先验(class prior),有$\hat{\pi}_c = \frac{N_c}{N}$,和朴素贝叶斯分类器里一样.对于类条件密度

(class-conditional densities),可以根据数据的类别标签来分开,对于每个高斯分布进行最大似然估计:

$$\hat{\mu}_c = \frac{1}{N_c}\sum_{i:y_i=c} x_i, \hat{\Sigma}_c = \frac{1}{N_c}\sum_{i:y_i=c}(x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T \quad (4.53)$$

具体实现可以参考本书配套的PMTK3当中的discrimAnalysisFit是MATLAB代码.一旦一个模型拟合出来了,就可以使用discrimAnalysisPredict来进行预测了,具体用到的是插值近似(plug-in approximation).

4.2.5 防止过拟合的策略

最大似然估计(MLE)的最大优势之一就是速度和简洁.然而,在高维度数据的情况下,最大似然估计可能会很悲惨地发生过拟合.尤其是当$N_c < D$,全协方差矩阵(full covariance matrix)是奇异矩阵的时候(singular),MLE方法很容易过拟合.甚至即便$N_c > D$,MLE也可能是病态的(ill-conditioned),意思就是很接近奇异.有以下几种方法来预防或解决这类问题:

- 假设类的特征是有条件独立的(conditionally independent),对这些类使用对角协方差矩阵(diagonal covariance matrix);这就等价于使用朴素贝叶斯分类器了,参考本书3.5.
- 使用一个全协方差矩阵,但强制使其对于所有的类都相同,即$\Sigma_c = \Sigma$.这称为参数绑定(parameter tying)或者参数共享(parameter sharing),等价于线性判别分析(LDA),参见本书4.2.2.
- 使用一个对角协方差矩阵,强迫共享.这叫做对角协方差线性判别分析,参考本书4.2.7.
- 使用全协方差矩阵,但倒入一个先验,然后整合.如果使用共轭先验(conjugate prior)就能以闭合形式(closed form)完成这个过程,利用了本书4.6.3当中的结果;这类似于本书3.5.1.2当中提到的使用贝叶斯方法的朴素贝叶斯分类器(Bayesian naive Bayes),更多细节参考 (Minka 2000f).
- 拟合一个完整的或者对角协方差矩阵,使用最大后验估计(MAP estimate),接下来会讨论两种不同类型的实现.
- 将数据投影到更低维度的子空间,然后在子空间中拟合其高斯分布.更多细节在本书8.6.3.3,其中讲了寻找最佳线性投影(即最有区分作用)的方法.

接下来说一些可选类型.

4.2.6 正交线性判别分析(Regularized LDA)*

假如我们在线性判别分析中绑定了协方差矩阵,即$\Sigma_c = \Sigma$,接下来就要对Σ进行最大后验估计了,使用一个逆向Wishart先验,形式为$IW(diag(\hat{\Sigma}_{mle}), v_0)$,更多内容参考本书4.5.1.然后就有了:

$$\hat{\Sigma} = \lambda diag(\hat{\Sigma}_{mle}) + (1-\lambda)\hat{\Sigma}_{mle} \quad (4.54)$$

上式中的λ控制的是正则化规模(amount of regularization),这和先验强度(strength of the prior),v_0

有关,更多信息参考本书4.6.2.1.这个技巧就叫做正则化线性判别分析(regularized discriminant analysis,缩写为 RDA,出自Hastie et al. 2009, p656).

当对类条件密度进行评估的时候,需要计算$\hat{\Sigma}^{-1}$,也就要计算$\hat{\Sigma}_{mle}^{-1}$,如果$D > N$那就没办法计算了.不过可以利用对矩阵X的奇异值分解（Singular Value Decomposition,缩写为SVD,参考本书12.2.3)来解决这个问题,如下面所述.(注意这个方法不能用于二次判别分析QDA,因为QDA不是关于x 的线性函数,是非线性函数了.)

设$X = UDV^T$是对设计矩阵(design matrix)的SVD分解,其中的V/U分别是$D \times N$和$N \times N$的正交矩阵(orthogonal matrix),而D是规模为N的对角矩阵(diagonal matrix).定义一个$N \times N$的矩阵$Z = UD$;这就像是一个在更低维度空间上的设计矩阵,因为我们假设了$N < D$.另外定义$\mu_z = V^T\mu$作为降维空间中的数据均值;可以通过$mu = V\mu_z$来恢复到原始均值,因为$V^TV = VV^T = I$.有了这些定义之后,就可以把最大似然估计(MLE)改写成下面的形式了:

$$\hat{\Sigma}_{mle} = \frac{1}{N}X^TX - \mu\mu^T \tag{4.55}$$

$$= \frac{1}{N}(ZV^T)^T(ZV^T) - (V\mu - z)(V\mu_z)^T \tag{4.56}$$

$$= \frac{1}{N}VZ^TZV^T - V\mu_z\mu_z^TV^T \tag{4.57}$$

$$= V(\frac{1}{N}Z^TZ - \mu_z\mu_z^T)V^T \tag{4.58}$$

$$= V\hat{\Sigma}_zV^T \tag{4.59}$$

上式中的$\hat{\Sigma}_z$是**Z**的经验协方差(empirical covariance).因此要重新写成最大后验估计(MAP)为:

$$\hat{\Sigma}_{map} = V\tilde{\Sigma}_zV^T \tag{4.60}$$

$$\tilde{\Sigma}_z = \lambda diag(\hat{\Sigma}_z) + (1-\lambda)\hat{\Sigma}_z \tag{4.61}$$

注意,我们并不需要真正去计算出来这个$D \times D$矩阵$\hat{\Sigma}_{map}$.这是因为等式4.38告诉我们,要使用线性判别分析(LDA)进行分类,唯一需要计算的也就是$p(y = c \mid x, \theta) \propto \exp(\delta_c)$,其中:

$$\delta_c = -x^T\beta_c + \gamma_c, \beta_c = \hat{\Sigma}^{-1}\mu_c, \gamma_c = -\frac{1}{2}\mu_c^T\beta_c + \log\pi_c \tag{4.62}$$

然后可以并不需要求逆$D \times D$矩阵就能计算正交线性判别分析(RDA)的关键项β_c.

$$\beta_c = \hat{\Sigma}_{map}^{-1}\mu_c = (V\tilde{\Sigma}V^Ts)^{-1}\mu_c = V\tilde{\Sigma}^{-1}V^T\mu_c = V\tilde{\Sigma}^{-1}\mu_{z,c} \tag{4.63}$$

4.2.7 对角线性判别分析(Diagonal LDA)

上文所述的是正交线性判别分析(RDA),有一种简单的替代方法,就是绑定协方差矩阵(covariance matrice),即线性判别分析(LDA)中$\Sigma_c = \Sigma$,然后对于每个类都是用一个对角协方差矩阵.这个模型就

叫做对角线性判别分析模型(diagonal LDA model),等价于$\lambda = 1$时候的正交线性判别分析(RDA).对应的判别函数如下所示(和等式4.33相对比一下):

$$\delta_c(x) = \log p(x, y = c \mid \theta) = -\sum_{j=1}^{D} \frac{(x_j - \mu_{cj})^2}{2\sigma_j^2} + \log \pi_c \quad (4.64)$$

通常设置$\hat{\mu}_{cj} = \bar{x}_{cj}, \hat{\sigma}_j^2 = s_j^2$,这个$s_j^2$是特征j(跨类汇集)的汇集经验方差(pooled empirical variance).

$$s_j^2 = \frac{\sum_{c=1}^{C} \sum_{i:y_i=c} (x_{ij} - \bar{x}_{cj})^2}{N - C} \quad (4.65)$$

对于高维度数据,这个模型比LDA和RDA效果更好(Bickel and Levina 2004).

此处查看原书图4.7

4.2.8 最近收缩质心分类器(Nearest shrunken centroids classifier)*

对角线性判别分析(diagonal LDA)有一个弱点,就是要依赖所有特征.在高维度情况下,可能更需要一个只依赖部分子集特征的方法,可以提高准确性或者利于解释.比如可以使用筛选方法(screening method),基于互信息量(mutual information),如本书3.5.4所述.本节要说另外一种方法,即最近收缩质心分类器(nearest shrunken centroids classifier, Hastie et al. 2009, p652).

基本思想是在稀疏先验(sparsity-promoting/Laplace prior)情况下对对角线性判别分析模型进行最大后验估计(MAP),参考本书13.3.更确切来说,用类独立特征均值(class-independent feature mean)m_j和类依赖偏移量(class-specific offset)Δ_{cj} 来定义类依赖特征均值(class-specific feature mean)μ_{cj}。 则有:

$$\mu_{cj} = m_j + \Delta_{cj} \quad (4.66)$$

接下来对Δ_{cj}这一项设一个先验,使其为零,然后计算最大后验估计(MAP).对特征j,若有对于所有类别c都有$\Delta_{cj} = 0$,则该特征在分类决策中则毫无作用,因为μ_{cj}是与c独立的.这样这些不具有判别作用的特征就会被自动忽略掉.这个过程的细节可以参考 (Hastie et al. 2009, p652)和(Greenshtein and Park 2009).代码可以参考本书配套的PMTK3程序中的 shrunkenCentroidsFit.

基于(Hastie et al. 2009, p652)的内容举个例子.设要对一个基因表达数据集进行分类,其中有2308个基因,4各类别,63个训练样本,20个测试样本.使用对角LDA分类器在测试集当中有五次错误.而是用最近收缩质心分类器对一系列不同的λ值,在测试集中都没有错误,如图4.7所示.更重要的是这个模型是稀疏的,所以更容易解读.图4.8所示的非惩罚估计(unpenalized estimate),灰色对应差值(difference)d_{cj},蓝色的是收缩估计(shrunken estimates)Δ_{cj}.(这些估计的计算利用了通过交叉验证估计得到的λ值.)在原始的2308个基因中,只有39个用在了分类当中.

接下来考虑个更难的问题,有16,603个基因,来自144个病人的训练集,54个病人的测试集,有14种不同类型的癌症(Ramaswamy et al. 2001).Hastie 等(Hastie et al. 2009, p656) 称最近收缩质心分类器用了6520个基因,在测试集上有17次错误,而正交判别分析(RDA,本书4.3.6)用了全部的16,603个基因,在测试集上有12次错误.本书配套的PMTK3程序当中的函数cancerHighDimClassifDemo可以再现这些数字.

此处查看原书图4.8

4.3 联合正态分布的推论(Inference in jointly Gaussian distributions)

给定联合分布$p(x_1, x_2)$,边界分布(marginal)$p(x_1)$和条件分布$p(x_1 | x_2)$是有用的.下面就说一下如何去计算,并且给出一些应用举例.这些运算在最不理想的情况下大概需要$O(D^3)$的时间.本书的20.4.3会给出一些更快的方法.

4.3.1 结果声明

定理 4.3.1

多元正态分布(MVN)的边界和条件分布.设$x = (x_1, x_2)$是联合正态分布,其参数如下:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \quad (4.67)$$

则边缘分布为:

$$\begin{aligned} p(x_1) &= N(x_1 | \mu_1, \Sigma_{11}) \\ p(x_2) &= N(x_2 | \mu_2, \Sigma_{22}) \end{aligned}$$

(4.68)

后验条件分布则为(重要公式):

$$\begin{aligned} p(x_1 | x_2) &= N(x_1 | \mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{1|2}^{-1}(x_2 - \mu_2) \\ &= \mu_1 - \Lambda_{12}\Lambda_{1|2}^{-1}(x_2 - \mu_2) \\ &= \Sigma_{1|2}(\Lambda_{11}\mu_1 - \Lambda_{12}(x_2 - \mu_2)) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1} \end{aligned}$$

(4.69)

上面这个公式很重要,证明过程参考本书4.3.4.

可见边缘和条件分布本身也都是正态分布.对于边缘分布,只需要提取出与x_1或者x_2对应的行和列.条件分布就要复杂点了.不过也不是特别复杂,条件均值(conditional mean)正好是x_2的一个线性函数,而条件协方差(conditional covariance)则是一个独立于x_2的常数矩阵(constant matrix).给出了后验均值(posterior mean)的三种不同的等价表达形式,后验协方差(posterior covariance)的两种不同的等价表达方式,每个表达式都在不同情境下有各自的作用.

4.3.2 举例

接下来就在实际应用中进行举例,可以让上面的方程更直观也好理解.

4.3.2.1 二维正态分布的边缘和条件分布

假设以一个二维正态分布为例,其协方差矩阵为:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

(4.70)

边缘分布$p(x_1)$则是一个一维正态分布,将联合分布投影到x_1这条线上即可得到:

$p(x_1) = N(x_1 \mid \mu_1, \sigma_1^2)$(4.71)

此处查看原书图4.9

假如观测$X_2 = x_2$,则可以通过使用$X_2 = x_2$这条线来对联合分布进行"切片(slicing)"得到条件分布$p(x_1 \mid x_2)$,如图4.9所示:

$p(x_1 \mid x_2) = N(x_1 \mid \mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2})$(4.72)

若$\sigma_1 = \sigma_2 = \sigma$,则有:

$p(x_1 \mid x_2) = N(x_1 \mid \mu_1 + \rho(x_2 - \mu_2), \sigma^2(1 - \rho^2))$(4.73)

图4.9所示的为$\rho = 0.8, \sigma_1 = \sigma_2 = 1, \mu = 0, x_2 = 1$.可见$E[x_1 \mid x_2 = 1] = 0.8$,这很好理解,因为$\rho = 0.8$就意味着如果$x_2$在其均值基础上加1,那么$x_2$则增加0.8.另外还可以发现
$var[x_1 \mid x_2 = 1] = 1 - 0.8^2 = 0.36$.这也好理解,由于通过观测$x_2$而对$x_1$有了非直接的了解,所以对$x_1$的不确定性就降低了.如果$\rho = 1$,就得到了$p(x_1 \mid x_2) = N(x_1 \mid \mu_1, \sigma_1^2)$,因为如果二者不相关也就是互相独

立的话,x_2就不可能承载关于x_1的任何信息了.

4.3.2.2 无噪音数据插值(Interpolating noise-free data)

若我们要估计一个一维函数,定义在闭区间[0,T]上,对于N次观测的点t_i即为$y_i = f(t_i)$.暂时先假设数据没有噪音(noise-free),对其进行插值(interpolate),即拟合一个完全通过数据的函数.(对于有噪音的数据,参考本书4.4.2.3.)那么问题来了:在观测数据点之间间隔的地方,这个函数该是什么样的呢?通常都假设这个函数是光滑的.在本书第15章,会讲如何对函数进行先验编码,以及如何使用观测值对先验进行更新来得到对函数的后验估计.不过本章的内容要简单很多,直接对一维输入上定义的函数进行最大后验估计(MAP),参考了 (Calvetti and Somersalo 2007, p135)的介绍.

先将这个问题离散化(discretizing).首先咱们将这个函数的定义域(支撑,support,我就是要说定义域这种很平民化的词你能怎样?)分割成D个等大子区间(equal subintervals).然后定义:

$$x_j = f(s_j), s_j = jh, h = \frac{T}{D}, 1 \le j \le D \quad (4.74)$$

此处查看原书图4.10

光滑先验的编码可以通过下面的方式实现:假设x_j是邻近两项x_{j-1}, x_{j+1}的均值,再加上正态分布的噪音项:

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + \epsilon_j, 2 \le j \le D-2 \quad (4.75)$$

上式中的$\epsilon \sim N(0, (1/\lambda)I)$.精度项(precision term)$\lambda$控制了函数波动幅度:大的$\lambda$表示我们认为函数非常光滑,而小的$\lambda$则表示这个函数可能"拐来拐去的(wiggly)".用向量形式可以将上面的等式写成如下所示:

$$Lx = \epsilon \quad (4.76)$$

上式中的L是一个$(D-2) \times D$的二阶有限差分矩阵(second order finite difference matrix):

$$L = \frac{1}{2}\begin{pmatrix} -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \dots & & \\ & & & -1 & 2 & -1 \end{pmatrix} \quad (4.77)$$

对应的先验形式如下:

$$p(x) = N(x|0, (\lambda^2 L^T L)^{-1}) \propto \exp(-\frac{\lambda^2}{2}||Lx||_2^2) \quad (4.78)$$

以后就假设已经用λ对L进行过缩放了,所以就会忽略掉λ项,就只将精度矩阵(precision matrix)写成$\Lambda = L^T L$.

这里要注意,虽然x是D维的,但是精度矩阵Λ实际上的秩只是D-2.所以这是一个不适用先验(improper prior),也称作内在高斯随机场(intrinsic Gaussian random field)(更多相关信息参考本书19.4.4).
不过只要观测超过2个数据点,即$N \geq 2$,这个先验就适用了.

接下来设x_2是N个对函数的无噪音观测，而x_1是$D-N$个函数值.不考虑泛化损失,先假设未知变量和已知变量分别被排序.然后就可以对L矩阵进行如下所示的分割:
$L = [L_1, L_2], L_1 \in R^{(D-2)\times(D-N)}, L_2 \in R^{(D-2)\times(N)}$(4.79)
然后也可以对联合分布的精度矩阵进行分割:

$$\Lambda = L^T L = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \begin{pmatrix} L_1^T L_1 & L_1^T L_2 \\ L_2^T L_1 & L_2^T L_2 \end{pmatrix}$$

(4.80)

利用等式4.69,就可以写出条件分布:

$$p(x_1 | x_2) = N(\mu_{1|2}, \Sigma_{1|2}) \tag{4.81}$$
$$\mu_{1|2} = -\Lambda_{11}^{-1}\Lambda_{12}x_2 = -L_1^T L_2 x_2 \tag{4.82}$$
$$\Sigma_{1|2} = \Lambda_{11}^{-1} \tag{4.83}$$

解下面的线性方程组就可以计算出均值:
$L_1\mu_{1|2} = -L_2 x_2$(4.84)

L_1是三角形矩阵,所以这解起来很容易.图4.10所示为这些等式的图像.从图中可见后验均值$\mu_{1|2}$等于特定各点的观测数据,而中间位置也都进行了光滑插值.
图中灰色的部分是95%的逐点边界置信区间(pointwise marginal credibility intervals),$\mu_j \pm 2\sqrt{\Sigma_{1|2,jj}}$.观察这部分会发现,远离数据点则方差增大,降低先验精度λ也会导致方差的增大.不过有趣的是λ并不会影响后验均值,因为在乘以Λ_{11}和Λ_{12}的时候消掉了.与之形成对比的是本书4.4.2.3的有噪音数据,那时候咱们就能发现先验精度会影响后验均值估计的光滑程度了.
边界置信区间并没有捕获到邻近位置上的相关性.对此可以从后验中推出完整的函数,也就是向量x,然后绘制函数图像.如图4.10中的细线所示.这就不像后验均值本身那么光滑了.这是因为先验只是处理了一阶差分(prior only penalizes first-order differences).更多相关细节参考本书4.4.2.3.

4.3.2.3 数据插补(Data imputation)

假设一个设计矩阵(design matrix)中缺失了某些值(entries).如果各列(columns)相关,可以用观察到的值对缺失值进行预测.如图4.11所示.从一个20维的正态分布中取样若干数据,然后故意在每行(row)隐藏掉一般的数据.接下来使用存留数据对缺失数据进行推测,使用真实(生成)模型(true

(generating) model).具体来说,对于每一行i,都计算$p(x_{h_i}|x_{v_i},\theta)$,其中的$h_i$和$v_i$分别是 i条件下隐藏和可见值的索引值(indices).从这里就能计算出每个缺失值的边缘分布$p(x_{h_{ij}}|x_{v_i},\theta)$.然后对这个分布的均值$\hat{x}_{ij}=E[x_j|x_{v_i},\theta]$进行投图;这就代表着对缺失值位置真实值的"最佳猜测",因为这使得期望平方误差(expected squared error)最小,更多细节参考本书5.7.如图4.11所示,这些估计和真实值还是很接近的.(当然了,如果$j\in v_i$则期望值就等于观测值,即$\hat{x}_{ij}=x_{ij}$.)

然后可以用$var[x_{h_{ij}}|x_{v_i},\theta]$来衡量对这个猜测的信心,不过图中没有展示出来.或者可以从$p(x_{h_{ij}}|x_{v_i},\theta)$中进行多次取样,这样就叫做多重插补(multiple imputation).

除了推算缺失值之外,我们可能还要计算表格中每个特定观测到的行(row)的似然率(likelihood),$p(x_{v_i}|\theta)$,这可以用等式4.68进行计算.这可以用于检测异常值(outliers)(也就是不太正常的观测结果,atypical observations).

此处参考原书图4.11

4.3.3 信息形式(Information form)

设$x\sim N(\mu,\Sigma)$.很明显$E[x]=\mu$就是均指向量,而$cov[x]=\Sigma$就是协方差矩阵(covariance matrix).这些都叫做分布的矩参数(moment parameters).不过有时候可能使用规范参数(canonical parameters)或者自然参数(natural parameters)更有用,具体定义如下所示:

$\overset{*}{\Lambda}=\Sigma^{-1},\overset{*}{\xi}=\Sigma^{-1}\mu$(4.85)

还可以转换回矩参数:

$\mu=\Lambda^{-1}\xi,\Sigma=\Lambda^{-1}$(4.86)

使用规范参数,可以将多元正态分布(MVN)写成信息形式(information form)(也就是写成指数组分布的形式(exponential family form),具体定义在本书9.2):

$N_c(x|\xi,\Lambda)=(2\pi)^{-D/2}|\Lambda|^{\frac{1}{2}}\exp[-\frac{1}{2}(x^T\Lambda x+\xi^T\Lambda^{-1}\xi-2x^T\xi)]$(4.87)

上式中使用了$N_c()$是为了和矩参数表达形式$N()$相区分.

边缘分布和条件分布公式也都可以推导出信息形式.为:

$p(x_2)=N_c(x_2|\xi_2-\Lambda_{21}\Lambda_{11}^{-1}\xi_1,\Lambda_{22}-\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12})$(4.88)

$p(x_1|x_2)=N_c(x_1|\xi_1-\Lambda_{12}x_2,\Lambda_{11})$(4.89)

通过上式可见比矩参数形式求边缘分布更容易,而信息形势下求条件分布更容易.

这种信息形式记法的另外一个好处是将两个正态分布相乘更简单了.如下所示:

$N_c(\xi_f,\lambda_f)N_c(\xi_g,\lambda_g)=N_c(\xi_f+\xi_g,\lambda_f+\lambda_g)$(4.90)

而在矩参数形式下,这相乘起来可就麻烦了:

$N(\mu_f,\sigma_f^2)N(\mu_g,\sigma_g^2)=N(\frac{\mu_f\sigma_g^2+\mu_g\sigma_f^2}{\sigma_f^2+\sigma_g^2},\frac{\sigma_f^2\sigma_g^2}{\sigma_f^2+\sigma_g^2})$(4.91)

4.3.4 结论证明*

这一节是要证明定理4.3.1.害怕矩阵代数计算的读者可以直接跳过这部分内容.本节开始要先推到一些有用的结果,这些结论不仅在这节要用到,在本书其他地方也有用.然后在结尾部分就给出证明.

4.3.4.1 使用Schur补(Schur complements)得到分区矩阵的逆矩阵(Inverse)

要想个办法对一个分区矩阵求逆矩阵.可以使用下面的结论.

定理4.3.2

分区矩阵的逆矩阵.设有一个常规分区矩阵(general partitioned matrix):

$$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

(4.92)

假设其中的E和H都是可逆的,则有:

$$M^{-1} = \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix} \quad (4.93)$$

$$= \begin{pmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{pmatrix} \quad (4.94)$$

其中:

$\overset{*}{M/H} = E - FH^{-1}G$(4.95)

$\overset{*}{M/E} = H - GE^{-1}F$(4.96)

我们就说M/H是$MwrtH$的Schur补(Schur complement).等式4.93就叫做分区求逆公式(partitioned inverse formula).

证明

如果把矩阵M的对角(diagonalize)去掉(block),就更好求逆矩阵了.可以用如下方式预处理,矩阵M左侧乘以一个三角矩阵来得使矩阵M的右上角部分为零:

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} E-FH^{-1} & 0 \\ G & H \end{pmatrix}$$

(4.97)

用如下方式预处理,矩阵M右侧乘以一个三角矩阵来得使矩阵左下角部分为零:

$$\begin{pmatrix} E-FH^{-1} & 0 \\ G & H \end{pmatrix}\begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} = \begin{pmatrix} E-FH^{-1} & 0 \\ 0 & H \end{pmatrix}$$

(4.98)

把上面两步结合起来就得到了:

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} E & F \\ G & H \end{pmatrix}\begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} = \begin{pmatrix} E-FH^{-1} & 0 \\ 0 & H \end{pmatrix}$$

(4.99)
上面的四个矩阵从左到右分别为X,M,Z,W,对这几个矩阵同时求逆矩阵就得到了:
$Z^{-1}M^{-1}X^{-1} = W^{-1}$(4.100)

然后就能推出:
$M^{-1} = ZW^{-1}X$(4.101)

用定义拆解出来就得到了:

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix}\begin{pmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix}\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \tag{4.102}$$

$$= \begin{pmatrix} (M/H)^{-1} & 0 \\ -H^{-1}G(M/H)^{-1} & H^{-1} \end{pmatrix}\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \tag{4.103}$$

$$= \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1}+H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix} \tag{4.104}$$

或者也可以把矩阵M分解成用E来表示,这样就有$M/E = (H - GE^{-1}F)$,就得到了:

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} E^{-1}+E^{-1}F(</E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -H^{-1}-(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{pmatrix}$$

(4.105)

证明完毕

4.3.4.2 矩阵求逆引理(the matrix inversion lemma)

接下来要利用上面的结果推出一些有用的推论.

推论4.3.1 矩阵求逆引理(matrix inversion lemma)

设有一个常规分区矩阵(general partitioned matrix)$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$,假设E和H都可逆.则有:

$(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1}$(4.106)

$(E - FH^{-1}G)^{-1}FH^{-1} = E^{-1}F(H - GE^{-1}F)^{-1}$(4.107)

$|E - FH^{-1}G| = |H - GE^{-1}F||H^{-1}||E|$(4.108)

上式中前两个方程就叫做矩阵求逆引理(matrix inversion lemma)或者叫做Sherman Morrison-Woodbury 公式(Sherman Morrison-Woodbury formula).第三个等式叫做矩阵行列式引理(matrix determinant lemma).在机器学习和统计学中上面这些公式的典型用法如下所示.设$E = \Sigma$是一个$N \times N$的对角矩阵,设$F = G^T = X$规模为$N \times D$,其中的N远大于D,即$N >> D$,设$H^{-1} = -I$.则有:

$(\Sigma + XX^T)^{-1} = \Sigma^{-1} - \Sigma^{-1}X(I + X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$(4.109)

等号左侧的计算需要$O(N^3)$时间,等号右侧的计算需要$O(D^3)$时间.

另外一种应用涉及到了对逆矩阵的一阶更新(rank one update)进行计算.设$H = -1$是一个标量(scalar),$F = u$是一个列向量(column vector),而$G = v^T$是一个行向量(row vector).然后则有:

$$(E + uv^T)^{-1} = E^{-1} + E^{-1}u(-1 - v^TE^{-1}u)v^TE^{-1} \quad (4.110)$$

$$= E^{-1} - \frac{E^{-1}uv^TE^{-1}}{1 + v^TE^{-1}u} \quad (4.111)$$

在对设计矩阵逐渐添加数据向量和对充分统计量进行更新的时候,可以用上上面的式子.(移除一个数据向量的方程与之类似,大家自己推导一下.)

证明

要证明等式4.106,只需要把等式4.93的左上部分和4.94等同起来(equate).为了证明等式4.107,则将等式4.93的右上部分和4.94等同起来(equate).等式4.108的证明留作练习.

4.3.4.3 高斯条件公式(Gaussian conditioning formulas)的证明

接下来回到主线,也就是推导等式4.69.首先把联合概率分布$p(x_1, x_2)$因式分解成$p(x_2)p(x_1|x_2)$:

$$E = \exp\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\}$$

(4.112)

利用等式4.102,则有:

$$E = \exp\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix}^{-1} \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \tag{4.113}$$

$$\times \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\} \tag{4.114}$$

$$= \exp\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}) \tag{4.115}$$

$$(x_1-\mu_1-\Sigma/\Sigma_{22})^{-1}(x_2-\mu_2))\} \times \exp\{-\frac{1}{2}(x_2-\mu_2)^T\Sigma_{-1}(x_2-\mu_2)\} \tag{4.116}$$

这就成了下面这种形式:

$\exp(x_1, x_2$的二次型(quadratic form)) × $\exp(x_2$的二次型)(4.117)

因此就可以成功地将联合分布拆解开:

$$p(x_1, x_2) = p(x_1|x_2)p(x_2) \tag{4.118}$$

$$= N(x_1|\mu_{1|2}, \Sigma_{1|2})N(x_2|\mu_2, \Sigma_{22}) \tag{4.119}$$

通过上面的等式也可以得到条件概率分布的参数:

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \tag{4.120}$$

$$\Sigma_{1|2} = \Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{4.121}$$

还可以根据$|M| = |M/H||H|$来检验归一化常数(normalization constants)是否正确:

$$(2\pi)^{(d_1+d_2)/2}|\Sigma|^{\frac{1}{2}} = (2\pi)^{(d_1+d_2)/2}(|\Sigma/\Sigma_{22}||\Sigma_{22}|)^{\frac{1}{2}} \tag{4.122}$$

$$= (2\pi)^{d_1/2}|\Sigma/\Sigma_{22}|^{\frac{1}{2}}(2\pi)^{d_2/2}|\Sigma_{22}|^{\frac{1}{2}} \tag{4.123}$$

上式中的$d_1 = \dim(x_1), d_2 = \dim(x_2)$.

对等式4.69的其他形式证明留作练习了.

4.4 线性高斯系统(Linear Gaussian systems)

加入我们有两个变量x和y.然后设$x \in R^{D_x}$是隐藏变量(hidden variable),而$y \in R^{D_y}$是对x的有噪音观察(noisy observation).假设有下面的先验和似然率(重要公式):

$$p(x) = N(x \mid \mu_x, \Sigma_x)$$
$$p(y \mid x) = N(y \mid Ax + b, \Sigma_y)$$

(4.124)
上式中的A是一个$D_y \times D_x$的矩阵.这就是一个线性高斯系统(linear Gaussian system).可以表示为$x \rightarrow y$,意思也就是x生成(generates)了y.本节会讲如何去"逆转箭头方向(invert the arrow)",也就是根据y来推测x.首先是给出结论,然后举几个例子,最后给出推导结论的过程.后面的章节还能看到对这些结论的更多应用.

4.4.1 结论表述

定理4.4.1

线性高斯系统的贝叶斯规则.
给定一个线性高斯系统,如等式4.124所述,则后验$p(x \mid y)$为(重要公式):

$$p(x \mid y) = N(x \mid \mu_{x \mid y}, \Sigma_{x \mid y})$$
$$\Sigma_{x \mid y}^{-1} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A$$
$$\mu_{x \mid y} = \Sigma_{x \mid y}[A^T \Sigma_y^{-1}(y - b) + \Sigma_x^{-1} \mu_x]$$

(4.125)

另外,归一化常数(normalization constant)$p(y)$为(重要公式):
$p(y) = N(y \mid A\mu_x + b, \Sigma_y + A\Sigma_x A^T)$(4.126)
证明过程参考本章4.4.3

4.4.2 样例

本节举几个对上述结论进行应用的例子.

4.4.2.1 从有噪音测量(noisy measurements)中推测未知标量(unknown scalar)

假如我们对某个隐藏量(underlying quantity)x进行N此有噪音测量 y_i;然后假设测量噪音有固定的

精确度(precision)$\lambda_y = \frac{1}{\sigma^2}$,则似然率(likelihood)为:

$$p(y_i|x) = N(y_i|x, \lambda_y^{-1}) \quad (4.127)$$

然后对未知源(unknown source)的值使用一个高斯先验:

$$p(x) = N(x|\mu_0, \lambda_0^{-1}) \quad (4.128)$$

我们需要计算的是$p(x|y_1, \ldots, y_N, \sigma^2)$.可以把这个改写成一种形式,以便于使用对高斯分布的贝叶斯规则,可以通过定义$y = (y_1, \ldots, y_N), A = a_N^T, \Sigma_y^{-1} = diag(\lambda_y I)$,其中的A的意思是一个$1 \times N$的由1构成的行向量(row vector).则有:

$$p(x|y) = N(x|\mu_N, \lambda_N^{-1}) \quad (4.129)$$

$$\lambda_N = \lambda_0 + N\lambda_y \quad (4.130)$$

$$\mu_N = \frac{N\lambda_y \bar{y} + \lambda_0 \mu_0}{\lambda_N} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\bar{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0 \quad (4.131)$$

这几个等式很直观了:后验精度(posterior precision)λ_N就正好是先验精度(prior precision)λ_0和N个单位的测量精度(measurement precision)λ_y的和.另外后验均值(posterior mean)μ_N也就是最大似然估计(MLE)$\bar{y}$和先验均值(prior mean)μ_0的凸组合(convex combination).很明显,这就表明了后验均值是在最大似然估计(MLE)和先验(prior)之间的妥协折中(compromise).如果先验相对于信号强度来说比较弱(即λ_0相对于λ_y来说较小),就赋予最大似然估计(MLE)更多权重.如果先验相对信号强度更强(即λ_0相对于λ_y来说更大),就给先验(prior)更高权重.这如图4.12所示,这和图3.6当中的β二项模型(beta-binomial model)的模拟结果(analogous results)很相似.

这里要注意后验均值写成了$N\lambda_y\bar{y}$的形式,因此具有测量N次,每次精度λ_y就相当于进行一次测量得到值$\bar{y}$而精度为$N\lambda_y$.

此处查看原书图4.12

我们可以把上面的结果写成后验方差(posterior variance)的形式,而不用后验精度(posterior precision),如下所示:

$$p(x|D, \sigma^2) = N(x|\mu_N, \Gamma_N^2 \quad (4.132)$$

$$\Gamma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\Gamma_0^2}} = \frac{\sigma^2\Gamma_0^2}{N\Gamma_0^2 + \sigma^2} \quad (4.133)$$

$$\mu_N = \Gamma_N^2\left(\frac{\mu_0}{\Gamma_0^2} + \frac{N\bar{y}}{\sigma^2}\right) = \frac{\sigma^2}{N\Gamma_0^2 + \sigma^2}\mu_0 + \frac{N\Gamma_0^2}{N\Gamma_0^2 + \sigma^2}\bar{y} \quad (4.134)$$

上式中的$\Gamma_0^2 = 1/\lambda_0$是先验方差(prior variance),而$\Gamma_N^2 = 1/\lambda_N$是后验方差(posterior variance).

我们也可以通过每次观测后更新来逐渐计算后验.如果$N = 1$,在进行一次单独观测后就可以重写后验,如下所示(下面定义了$\Sigma_y = \sigma^2, \Sigma_0 = \Gamma_0^2, \Sigma_1 = \Gamma_1^2$分别是似然函数/先验/后验的方差):

$$p(x|y) = N(x|\mu_1, \Sigma_1) \tag{4.135}$$

$$\Sigma_1 = (\frac{1}{\Sigma_0} + \frac{1}{\Sigma_y})^{-1} = \frac{\Sigma_y \Sigma_0}{\Sigma_0 + \Sigma_y} \tag{4.136}$$

$$\mu_1 = \Sigma_1(\frac{\mu_0}{\Sigma_0} + \frac{y}{\Sigma_y}) \tag{4.137}$$

可以以下面三种不同形式来写出后验均值(posterior mean):

$$\mu_1 = \frac{\Sigma_y}{\Sigma_y + \Sigma_0}\mu_0 + \frac{\Sigma_0}{\Sigma_y + \Sigma_0}y \tag{4.138}$$

$$= \mu_0 + (y - \mu_0)\frac{\Sigma_0}{\Sigma_y + \Sigma_0} \tag{4.139}$$

$$= y - (y - \mu_0)\frac{\Sigma_y}{\Sigma_y + \Sigma_0} \tag{4.140}$$

上面的三个等式中,第一个式子就是对先验和数据的凸组合(convex combination).第二个是将先验均值朝向数据进行调整.第三个是将数据朝向先验均值调整,这也叫做收缩过程(shrinkage).这三者都是等价的,都表达了在似然率和先验之间的权衡妥协.如果Σ_0相对于Σ_Y较小,对应的就是强先验(strong prior),收缩规模(amount of shrinkate)就很大,参考图4.12(a),而如果反过来Σ_0相对于Σ_Y更大,对应的就是弱先验(weak prior)收缩规模就小了,参考图4.12(b).

另外一种对收缩规模定量的方法是用信噪比(signal-to-noise ratio,缩写为SNR),定义如下:

$$SNR \stackrel{*}{=} \frac{E[X^2]}{E[\epsilon^2]} = \frac{\Sigma_0 + \mu_0^2}{\Sigma_y} \tag{4.141}$$

上式中的$x \sim N(\mu_0, \Sigma_0)$是真是信号(true signal),而$y = x + \epsilon$是观测信号,而$\epsilon \sim N(0, \Sigma_y)$就是噪音项.

4.4.2.2 从有噪音测量(noisy measurements)中推测未知矢量(unknown vector)

接下来考虑一个N次向量值的观测$y_i \sim N(x, \Sigma_y)$,有高斯先验$x \sim N(\mu_0, \Sigma_0)$.设$A = I, b = 0$,精度为$N\Sigma_y^{-1}$的有效观测设为$\bar{y}$,则有:

$$p(x|y_1, \ldots, y_N) = N(x|\mu_N, \Sigma_N) \tag{4.142}$$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + N\Sigma_y^{-1} \tag{4.143}$$

$$\mu_N = \Sigma_N(\Sigma_y^{-1}(N\bar{y}) + \Sigma_0^{-1}\mu_0) \tag{4.144}$$

图4.13所示是一个二维情况下的样例.可以把x理解为一个物体在二维空间内的真实位置,但这个位置是未知的,可以想象成导弹或者飞机,然后y_i就是带噪音的观测,也就类似雷达上面的信号点.随着收到的信号点越来越多了,就更好去对信号源的位置进行定位了.具体参考本书18.31,其中讲述了对这个例子进行扩展,去追踪运动物体,使用著名的卡尔曼滤波算法(Kalman filter algorithm).

然后设想我们有多个测量设备,然后想要将他们结合起来;这也就是传感器融合(sensor fusion).如果我们进行了多次观测,每次都有不同的协方差(covariances)(对应的就是不同可靠程度的传感器),后验分布就应当是适当地对数据的加权平均.如图4.14所示.采用的是对x的无信息先验(uninformative prior),名为$p(x) = N(\mu_0, \Sigma_0) = N(0, 10^{10} I_2)$.进行了两次有噪音的观测,分别为$y_1 \sim N(x, \Sigma_{y,1})$和$y_2 \sim N(x, \Sigma_{y,2})$.然后就可以计算$p(x|y_1, y_2)$.

在图4.14(a)中,设置了$\Sigma_{y,1} = \Sigma_{y,2} = 0.01 I_2$,所以两个传感器就都是可靠程度相同.这时候后验均值就是两次观测y_1, y_2的均值.在图4.14(b)中,设置的是$\Sigma_{y,1} = 0.05 I_2, \Sigma_{y,2} = 0.01 I_2$,这也就意味着第二个传感器比第一个更可靠.,这时候后验均值就距离y_2更近了.在图4.14(c)中,设置有:

$$\Sigma_{y,1} = 0.01 \begin{pmatrix} 10 & 1 \\ 1 & 1 \end{pmatrix}, \Sigma_{y,2} = 0.01 \begin{pmatrix} 1 & 1 \\ 1 & 10 \end{pmatrix} \quad (4.145)$$

所以也就是说第一个传感器对于y_2成分(component)(竖着的方向)更可靠,而第二个传感器对于y_1成分(横着的方向)更可靠.这时候后验均值就使用了y_1的竖直元素和y_2的水平元素.

此处查看原书图4.13

此处查看原书图4.14

要注意,这个方法关键在于对每个传感器的不确定性的建模;没有考虑权重就计算均值会得到错误结果.不过这是已经假设了每个传感器的精度都已知了.如果不知道每个传感器的精确度,也还是要对Σ_1, Σ_2的精确度进行建模.更多细节参考本书4.6.4.

4.4.2.3 插入噪音数据

再回头看看本书4.3.2.2当中的例子.这次咱们不再假设观测是无噪音的.而是假设进行了N次的有噪音观测y_i,为了通用,就假设对应了$x_1, \ldots, x_N$.可以用一个线性高斯系统来对此进行建模:
$y = Ax + \epsilon$(4.146)
上式中的$\epsilon \sim N(0, \Sigma_y), \Sigma_y = \sigma^2 I$,$\sigma^2$就是观测噪音,而A是一个$N \times D$的投影矩阵(projection matrix),对观测到的元素进行了筛选.例如,如果N =2,D =4,则有:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad (4.147)$$

还是用同样的不适用先验(improper prior)$\Sigma_x = (L^T L)^{-1}$,可以很容易计算得出后验均值和方差.如图4.15所示,对后验均值/后验方差以及一些后验样本进行投图.然后可以看出先验精确度λ同时影响着

后验的均值和方差.对于一个强先验(大的λ),这时候的估计就很光滑,而不确定性就很低.但对于弱先验(小的λ),估计结果就扭来扭曲,远离数据部分的估计结果的不确定性就高了.

解下面的优化问题就能计算出后验均值:

$$\min_x \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{D}[(x_j - x_{j-1})^2 + (x_j - x_{j+1})^2] \quad (4.148)$$

上式中定义了$x_0 = x_1, x_{D+1} = x_D$来简化记号.这实际上是对下面问题的离散近似:

$$\min_f \frac{1}{2\sigma^2} \in (f(t) - y(t))^2 dt + \frac{\lambda}{2}\int [f'(t)]^d t \quad (4.149)$$

其中的$f'(t)$是f的一阶导数(first derivative).第一项用于拟合数据,第二项用于抑制函数避免过于扭曲.这是Tikhonov正则化(Tikhonov regularization)的一个例子,这是一种很流行的函数数据分析方法.参见本书第十五章可以看到更高级的方法,保证了更高阶的光滑性(也就是得到的结果不会看上去有很多锯齿).

此处查看原书图4.15

4.4.3 结论证明*

接下来推导一下等式4.125.基本思想是推导联合分布$p(x, y) = p(x)p(y \mid x)$,然后使用本书4.3.1的结论来计算$p(x \mid y)$.

更详细来说,按照下面的步骤进行.首先是得到联合分布函数的对数形式,如下所示(取对数是去除了不相关的常数项):

$$\log p(x, y) = -\frac{1}{2}(x - \mu_x)^T \Sigma_x^{-1}(x - \mu_x) - \frac{1}{2}(y - Ax - b)^T \Sigma_y^{-1}(y - Ax - b) \quad (4.150)$$

很明显这就是一个联合高斯分布,因为是一个二次型的指数.

扩展有x和y的二次项,然后线性项和常数项全部忽略掉,就得到了:

$$Q = -\frac{1}{2}x^T \Sigma_x^{-1} x - \frac{1}{2}y^T \Sigma_y^{-1} y - \frac{1}{2}(Ax)^T \Sigma_t^{-1}(Ax) + y^T \Sigma_y^{-1} Ax \quad (4.151)$$

$$= \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Sigma_x^{-1} + A^T \Sigma_y^{-1} A & -A^T \Sigma_y^{-1} \\ -\Sigma_y^{-1} A & \Sigma_y^{-1} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (4.152)$$

$$= \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \quad (4.153)$$

联合分布的精度矩阵则定义为:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_x^{-1} + A^T\Sigma_y^{-1}A & -A^T\Sigma_y^{-1} \\ -\Sigma_y^{-1}A & \Sigma_y^{-1} \end{pmatrix}^* = \Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix} \quad (4.154)$$

$$p(x|y) = N(\mu_{x|y}, \Sigma_{x|y}) \quad (4.155)$$

$$\Sigma_{x|y} = \Lambda_{xx}^{-1} = (\Sigma_x^{-1} + A^T\Sigma_y^{-1}A)^{-1} \quad (4.156)$$

$$\mu_{x|y} = \Sigma_{x|y}(\Lambda_{xx}\mu_x - \Lambda_{xy}(y - \mu_y)) \quad (4.157)$$

$$= \Sigma_{x|y}(\Sigma_x^{-1}\mu + A^T\Sigma_y^{-1}(y - b)) \quad (4.158)$$

4.5 题外话(Digression):威沙特分布(Wishart distribution)

威沙特分布(Wishart distribution)是将γ分布(Gamma distrustion)对正定矩阵(positive definite matrices)的推广.(Press 2005, p107) 称:按照重要性和有用性的顺序来排列,在多元统计中,威沙特分布仅次于正态分布.通常用这个模型来对协方差矩阵Σ或者逆矩阵$\Lambda = \Sigma^{-1}$的不确定性来进行建模.

Wishart 分布的概率密度函数定义如下:

$$Wi(\Lambda|S, v) = \frac{1}{Z_{Wi}}|\Lambda|^{(v-D-1)/2}\exp(-\frac{1}{2}tr(\Lambda S^{-1}) \quad (4.159)$$

上式中的v也叫做自由度(degrees of freedom),S就是缩放矩阵(scale matrix).稍后会对这些参数的含义给出更多讲解.

这个分布的归一化常数(normalization constant)(需要在整个对称的概率密度矩阵上进行积分)为下面的表达式:

$$Z_{Wi} = 2^{vD/2}\Gamma_D(v/2)|S|^{v/2} \quad (4.160)$$

上式中的Γ_D是多元γ函数(multivariate gamma function):

$$\Gamma_D(x) = \pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(x + (1-i)/2) \quad (4.161)$$

因此$\Gamma_1(a) = \Gamma(a)$,以及:

$$\Gamma_D(v_0/2) = \prod_{i=1}^{D}\Gamma(\frac{v_0 + 1 - i}{2}) \quad (4.162)$$

只有当$v > D - 1$的时候才存在归一化常数,因此概率密度函数也仅在此时有意义.

Wishart分布和正态分布之间有一定联系.具体来说就是,设$x_i \sim N(0, \Sigma)$为正态分布,那么散点矩阵(scatter matrix)$S = \sum_{i=1}^{N}x_i x_i^T$就有一个Wishart分布:$S \sim Wi(\Sigma, 1)$.因此$E[S] = N\Sigma$.另外可以得出分布$Wi(S, v)$的均值(mean)和众数(mode)为:

$$mean = vS, mode = (v - D - 1)S \quad (4.163)$$

其中众数(mode)仅当$v > D + 1$的时候才存在.
如果D=1,那么Wishart就降回到了γ分布(Gamma distribution):
$Wi(\lambda | s^{-1}, v) = Ga(\lambda | \frac{v}{2}, \frac{s}{2})$(4.164)

4.5.1 逆威沙特分布(Inverse Wishart distribution)

在练习2.10中,如果$\lambda \sim Ga(a, b)$则有$\frac{1}{\lambda} \sim IG(a, b)$.类似地,如果有$\Sigma^{-1} \sim Wi(S, v)$,则有$\Sigma \sim IW(S^{-1}, v + D + 1)$,IW就是逆威沙特分布(inverse Wishart),是对逆γ分布(inverse Gamma)的多维推广.定义方式为:对于$v > D - 1, S \succ 0$:

$$IW(\Sigma | S, v) = \frac{1}{Z_{IW}} |\Sigma|^{-(v+D+1)/2} \exp(-\frac{1}{2} tr(S^{-1}\Sigma^{-1})) \quad (4.165)$$

$$Z_{IW} = |S|^{-v/2} 2^{vD/2} \Gamma_D(v/2) \quad (4.166)$$

很显然,这个分布有如下的性质:

$mean = \frac{S^{-1}}{v-D-1}, mode = \frac{S^{-1}}{v+D+1}$(4.167)

如果D=1,这个分布就降到了拟γ分布了:

$IW(\sigma^2 | S^{-1}, v) = IG(\sigma^2 | |v/2, S/2)$(4.168)

此处查看原书图4.16

4.5.2 威沙特分布可视化*

威沙特分布(Wishart)是矩阵的分布,所以很难画出密度函数.不过在二维情况下,可以对其进行取样,使用取样结果矩阵的特征向量来定义一个椭圆,具体如本书4.1.2所述.图4.16是一些样例.

对更高维度的矩阵,就可以投影威沙特分布的边缘分布(marginals).威沙特分布的矩阵的对角元素服从γ分布,所以也容易投影出来.非对角元素的分布通常就比较难以解出来了,不过可以从分钟抽样矩阵,然后根据经验计算抽样得到的矩阵的分布.可以把抽样得到的矩阵转各自转换成一个相关矩阵(correlation matrix)然后进行蒙特卡洛估计(参考本书2.7),来得到相关系数期望:

$E[R_{ij}] \approx \frac{1}{S} \sum_{s=1}^{S} R(\Sigma^s)_{ij}$(4.169)

其中的$\Sigma^{(s)} \sim Wi(\Sigma, v)$和$R(\Sigma)$就把矩阵$\Sigma$转换成了一个相关矩阵:
$R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$(4.170)

可以用核密度估计(kernel density estimation,参考本书14.7.2)来对单变量密度$E[R_{ij}]$生成一个光滑

近似来投图.图4.16是一些例子.

4.6 多元正态分布(MVN)的参数推测

之前已经讲的是在已知参数$\theta = (\mu, \Sigma)$的时候对一个高斯分布(正态分布)的推测.现在来讨论对这些参数本身的推测.假设数据形式为$x_i \sim N(\mu, \Sigma), i = 1:N$的全部范围都得到了观测,所以就没有缺失数据(本书11.6.1是讨论在有缺失数据的情况下对多元正态分布(MVN)进行参数估计).简单来说,就是把后验推断分成三部分,首先是计算$p(\mu | D, \Sigma)$,然后计算$p(\Sigma | D, \mu)$,最后计算联合分布$p(\mu, \Sigma | D)$.

4.6.1 μ的后验分布

之前说过如何对μ进行最大似然估计(MLE)了,现在说下如何计算其后验,这对于对其本身值的不确定性进行建模很有用.

似然函数形式为:

$$p(D | \mu) = N(\bar{x} | \mu, \frac{1}{N}\Sigma) \quad (4.171)$$

为了简化,使用共轭先验(conjugate prior),这里用的是一个高斯分布.具体来说就是如果$p(\mu) = N(\mu | m_0, V_0)$,然后就可以推出一个对$\mu$的高斯后验分布,这要基于本书4.4.2.2的结论.这样得到了:

$$p(\mu | D, \Sigma) = N(\mu | m_N, V_N) \quad (4.172)$$

$$V_N^{-1} = V_0^{-1} + N\Sigma^{-1} \quad (4.173)$$

$$m_N = V_N(\Sigma^{-1}(N\bar{x}) + V_0^{-1}m_0) \quad (4.174)$$

这就跟基于有噪音的雷达光电来推测目标位置是一模一样的过程,只不过这时候在推测的是一个分布的均值,而不是有噪音的样本.(对于一个贝叶斯方法来说,参数的不确定性和其他任何事情的不确定性没有区别.)

可以设置$V_0 = \infty I$来建立一个无信息先验.这样则有$p(\mu | D, \Sigma) = N(\bar{x}\frac{1}{N}\Sigma)$,所以后验均值就等于最大似然估计(MLE).另外我们还能发现后验方差降低到了$\frac{1}{N}$,这是频率视角概率统计(frequentist statistics)的标准结果.

4.6.2 Σ的后验分布*

然后说如何计算$p(\Sigma | D, \mu)$.似然函数形式如下:

$$p(D | \mu, \Sigma) \propto |\Sigma|^{-\frac{N}{2}} \exp(-\frac{1}{@} tr(S_\mu \Sigma^{-1})) \quad (4.175)$$

对应的共轭先验正好是逆威沙特分布,参考4.5.1.还记得这就有下面的概率密度函数(pdf):

$$IW(\Sigma|S_0^{-1}, v_0) \propto |\Sigma|^{-(v_0+D+1)/2}\exp(-\frac{1}{2}tr(S_0\Sigma^{-1}))(4.176)$$

上式中$v_0 > D-1$就是自由度(degrees of freedom,缩写为dof),而S_0是对称的概率密度矩阵(symmetric pd matrix).S_0^{-1}就是先验散布矩阵(prior scatter matrix),而$N_0 \overset{*}{=} v_0 + D + 1$控制了先验强度,所以扮演的角色也就类似于取样规模N.

此处查看原书图4.17

把似然函数和先验乘在一起,就可以发现后验也是一个逆威沙特分布(inverse Wishart):

$$p(\Sigma|D,\mu) \propto |\Sigma|^{\frac{N}{2}}\exp(-\frac{1}{2}tr(\Sigma^{-1}S_\mu)|\Sigma|^{-(v_0+D+1)/2}))\exp(-\frac{1}{2}tr(\Sigma^{-1}S_0)) \quad (4.177)$$

$$= |\Sigma|^{-\frac{N+(v_0+D+1)}{2}}\exp(-\frac{1}{2}tr[\Sigma^{-1}(S_\mu+S_0)]) \quad (4.178)$$

$$= IW(\Sigma|S_N, v_N) \quad (4.179)$$

$$v_N = v_0 + N \quad (4.180)$$

$$S_N^{-1} = S_0 + S_\mu \quad (4.181)$$

用文字来表述,就是说后验强度(posterior strength)v_N就是先验强度(prior strength) $v_)$加上观测次数N,而后验散布矩阵(posterior scatter matrix)S_N也就是先验散布矩阵(prior scatter matrix)S_0加上数据散布矩阵(data scatter matrix)S_μ.

4.6.2.1 最大后验估计(MAP estimation)

通过等式4.7可知$\hat{\Sigma}_{mle}$是一个秩(rank)为 $\min(N, D)$的矩阵.如果$N < D$,就是一个非满秩的(not full rank),因此就不可逆(uninvertible).而如果$N > D$,也可能$\hat{\Sigma}$是病态的(ill-conditioned)(意思就是近乎奇异矩阵).

要解决这些问题,可以用后验模(posterior mode)或者均值(mean).使用和最大似然估计(MLE)推导类似的技巧,就可以推出最大后验估计(MAP):

$$\hat{\Sigma}_{map} = \frac{S_N}{v_N+D+1} = \frac{S_0+S_\mu}{N_0+N}(4.182)$$

如果用一个不适用均匀先验(improper uniform prior),对应的就是$N_0 = 0, S_0 = 0$,也就恢复到了最大似然估计(MLE).

如果使用一个适当的含信息先验(proper informative prior),只要D/N比较大,比如超过0.1的时候,就很被咬了.设$\mu = \bar{x}$,则$S_\mu = S_{\bar{x}}$.然后就可以吧最大后验估计(MAP)写成一个先验模(prior mode)和最大

似然估计(MLE)的凸组合(convex combination).设$\Sigma_0 \stackrel{*}{=} \frac{S_0}{N_0}$为先验模(prior mode).然后可以把后验模(posterior mode)写成如下形式:

$$\hat{\Sigma}_{map} = \frac{S_0 + S_{\tilde{x}}}{N_0 + N} = \frac{N_0}{N_0 + N}\frac{S_0}{N_0} + \frac{N_0}{N_0 + N}\frac{S}{N} = \lambda\Sigma_0 + (1-\lambda)\hat{\Sigma}_{mle} \quad (4.183)$$

其中的$\lambda = \frac{N_0}{N_0 + N}$,控制的是朝向先验收缩(shrinkage)的规模(amount).

这就引出了另外一个问题:先验的那些参数都是哪来的?通常可以通过交叉验证来设置λ.或者可以使用闭合形式公式(closed-form formula),出自(Ledoit and Wolf 2004b,a; Schaefer and Strimmer 2005),是在使用平方损失(squared loss)的情况下的频率论角度的最优估计(optimal frequentist estimate).关于这是不是对协方差矩阵(covariance matrices)最自然的损失函数(loss function)还有争议,因为忽略了正定约束(positive definite constraint),不过确实能得到一个简单的估计器(estimator),本书配套的PMTK软件中的shrinkcov函数是一个实现.稍后再讨论贝叶斯角度对λ的估计.

至于先验协方差矩阵(prior covariance matrix)S_0,可以用下面的(依赖数据的)先验:$S_0 = diag(\hat{\Sigma}_{mle})$.这时候最大后验估计为:

$$\hat{\Sigma}_{map}(i,j) = \begin{cases} \hat{\Sigma}_{mle}(i,j) & \text{if } i = j \\ (1-\lambda)\hat{\Sigma}_{mle}(i,j) & \text{otherwise} \end{cases} \quad (4.184)$$

这样就能发现对角项目等于他们的最大似然估计(MLE),而非对角元素就朝着0收缩了.这也叫收缩估计(shrinkage estimation)或者正则化估计(regularized estimation).

图4.17中就展示了最大后验估计(MAP)的好处.设对一个50维的正态分布进行拟合,分别使用$N = 100, N = 50, N = 25$个数据点.很明显最大后验分布总是良好状态的(well-conditioned),而不像最大似然估计(MLE)会有病态的情况出现.特别是最大后验估计(MAP)的特征谱(eigenvalue spectrum)会比最大似然估计(MLE)的更接近真是矩阵.不过特征向量(eigenvectors)不受影响.

在后面的章节中,当我们要对高维度数据的协方差矩阵进行拟合的时候,对Σ的正则估计的重要性就很明显了.

4.6.2.2 单变量后验(Univariate posterior)

在一维情况下,似然函数(likelihood)形式如下所示:

$$p(D \mid \sigma^2) \propto (\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right) \quad (4.185)$$

标准共轭先验(standard conjugate prior)正好就是一个逆γ分布(inverse Gamma distribution),也就

是标量版本的逆威沙特分布(inverse Wishart):

$$IG(\sigma^2 \mid a_0, b_0) \propto (\sigma^2)^{1(a_0+1)} \exp(-\frac{b_0}{\sigma^2}) \quad (4.186)$$

此处参考原书图4.18

把似然函数(likelihood)和先验(prior)乘起来就会发现后验(posterior)也是IG:

$$p(\sigma^2 \mid D) = IG(\sigma^2 \mid a_N, b_N) \quad (4.187)$$

$$a_N = a_0 + N/2 \quad (4.188)$$

$$b_N = b_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (4.189)$$

图4.18为图示.

后验的形式不像多元情况下的那样好看,因为有了因子$\frac{1}{2}$.这是因为$IW(\sigma^2 \mid s_0, v_0) = IG(\sigma^2 \mid \frac{s_0}{2}, \frac{v_0}{2})$.使用逆正态分布$IG(a_0, b_0)$的另一个问题是先验同时对$a_0, b_0$进行编码(encoded).要避免这些问题,通常从统计学角度来说,都是使用对逆向高斯分布(IG distribution)的替代参数化,也就是(缩放)逆卡方分布((scaled) inverse chi-squared distribution),定义如下所示:

$$\chi^{-2}(\sigma^2 \mid v_0, \sigma_0^2) = IG(\sigma^2 \mid \frac{v_0}{2}, \frac{v_0\sigma_0^2}{2}) \propto (\sigma^2)^{-v_0/2-1} \exp(-\frac{v_0\sigma_0^2}{2\sigma^2}) \quad (4.190)$$

上式中的v_0控制了先验的强度,而σ^2对先验的值进行了编码.这样后验则成了:

$$p(\sigma^2 \mid D, \mu) = \chi^{-2}(\sigma^2 \mid v_N, \sigma_N^2) \quad (4.191)$$

$$v_N = v_0 + N \quad (4.192)$$

$$\sigma_N^2 = \frac{v_0\sigma_0^2 + \sum_{i=1}^{N}(x_i - \mu)^2}{v_N} \quad (4.193)$$

可见后验的自由度(dof)v_N是先验自由度(dof)v_0加上N,而后验平方和$v_n\sigma_N^2$就是先验平方和$v_0\sigma_0^2$加上数据的平方和.

可以设$v_0 = 0$来模拟一个无信息先验(uninformative prior)$p(\sigma^2) \propto \sigma^{-2}$,也很好直观理解,就是对应着零虚拟样本规模(zero virtual sample size).

4.6.3 μ和Σ的后验分布*

现在来讨论一下如何计算$p(\mu, \Sigma \mid D)$.这些结论有点复杂,不过在本书后面的章节会很有用.对于第一次阅读的读者来说,可以先跳过.

4.6.3.1 似然函数(likelihood)

似然函数为:

$$p(D\,|\,\mu,\Sigma) = (2\pi)^{-ND/2}\,|\Sigma|^{-\frac{N}{2}}\exp(-\frac{N}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) \quad (4.194)$$

很明显:

$$\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = tr(\Sigma^{-1}S_{\bar{x}}) + N(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \quad (4.195)$$

因此可以把似然函数写成如下的形式:

$$p(D\,|\,\mu,\Sigma) = (2\pi)^{-ND/2}\,|\Sigma|^{-\frac{N}{2}}\exp(-\frac{N}{2}(\mu-\bar{x})^T\Sigma^{-1}(\mu-\bar{x})) \quad (4.196)$$

$$\exp(-\frac{N}{2}tr(\Sigma^{-1}S_{\bar{x}})) \quad (4.197)$$

后面会用到这个形式.

4.6.3.2 先验(Prior)

先验形式为:

$$p(\mu,\Sigma) = N(\mu\,|\,m_0,V_0)IW(\Sigma\,|\,S_0,v_0) \quad (4.198)$$

很不幸,这和似然函数不共轭(not conjugate).为什么呢?注意μ和Σ在似然函数(likelihood)中以非因子形式(non-factorized way)共同出现,因此在后验中也会耦合在一起(coupled together).

上面的先验就也被叫做半共轭(semi-conjugate)或者条件共轭(conditionally conjugate),因为两个条件分布$p(\mu\,|\,\Sigma),p(\Sigma\,|\,\mu)$都是独立共轭(individually conjugate)的.要建立一个完全共轭先验(full conjugate prior),需要让μ,Σ两者相互依赖.所以可以使用下面这样形式的联合分布:

$$p(\mu,\Sigma) = p(\Sigma)p(\mu\,|\,\Sigma) \quad (4.199)$$

参考一下等式4.197中的似然函数等式,就可以发现自然共轭先验(natural conjugate prior)的形式为正态逆威沙特分布(Normal-inverse-wishart,缩写为 NIW),定义形式如下所示:

$$NIW(\mu, \Sigma \mid m_0, k_0, v_0, S_0) \stackrel{*}{=} \tag{4.200}$$

$$N(\mu \mid m_0, \frac{1}{k_0}\Sigma) \times IW(\Sigma \mid S_0, v_0) \tag{4.201}$$

$$= \frac{1}{Z_{NIW}} |\Sigma|^{\frac{1}{2}} \exp(-\frac{k_0}{2}\mu - m_0()^T \Sigma^{-1}(\mu - m_0)) \tag{4.202}$$

$$\times |\Sigma|^{-\frac{v_0+D+1}{2}} \exp(-\frac{1}{2} tr(\Sigma^{-1} S_0)) \tag{4.203}$$

$$= \frac{1}{Z_{NIW}} |\Sigma|^{-\frac{v_0+D+2}{2}} \tag{4.204}$$

$$\times \exp(-\frac{k_0}{2}(\mu - m_0)^T \Sigma^{-1}(\mu - m_0) - \frac{1}{2} tr(\Sigma^{-1} S_0)) \tag{4.205}$$

$$Z_{NIW} = 2^{V_0 D/2} \Gamma_D(v_0/2)(2\pi/k_0)^{D/2} |S_0|^{-v_0/2} \tag{4.206}$$

上式中的$\Gamma_D(a)$是多元γ分布(multivariate Gamma function).
上面这个逆威沙特分布的参数可以通过如下步骤来进行推断:m_0就是μ的先验均值,而k_0就是对这个先验的相信程度,S_0 是正比于Σ的先验均值,而v_0 是对这个先验的相信程度.

参考(Minka 2000f)可以发现,(不适用(improper))无信息先验(uninformative prior)的形式如下所示:

$$\lim_{k \to 0} N(\mu \mid m_0, \Sigma/k) IW(\Sigma \mid S_0, k \propto |2\pi\Sigma|^{\frac{1}{2}} |\Sigma|^{-(D+1)/2} \tag{4.207}$$

$$\propto |\Sigma|^{-(D/2+1)} \propto NIW(\mu, \Sigma \mid 0, 0, 0, 0I) \tag{4.208}$$

在实践中,一般都是使用弱含信息数据依赖先验(weakly informative data-dependent prior)比较好.
常规选择(参考(Chipman et al. 2001, p81), (Fraley and Raftery 2007, p6))是设置
$S_0 = diag(S_{\bar{x}})/N, v_0 = D + 2$来确保$E[\Sigma] = S_0$,然后设$\mu_0 = \bar{x}$以及$k_0$为比较小的数值,比如0.01.

4.6.3.3 后验

如练习4.11所示,后验可以表示成更新过参数的逆威沙特分布(NIW):

$$p(\mu, \Sigma \mid D) = NIW(\mu, \Sigma \mid m_N, k_N, v_N, S_N) \tag{4.209}$$

$$m_N = \frac{k_0 m_0 + N\bar{x}}{k_N} = \frac{k_0}{k_0 + N} m_0 + \frac{N}{k_0 + N}\bar{x} \tag{4.210}$$

$$k_N = k_0 + N \tag{4.211}$$

$$v_N = v_0 + N \tag{4.212}$$

$$S_N = S_0 + S_{\bar{x}} + \frac{k_) N}{k_0 + N}(\bar{x} - m_0)(\bar{x} - m_0)^T \tag{4.213}$$

$$= S_0 + S + k_0 m_0 m_0^T - k_N m_N m_N^T \tag{4.214}$$

上式中我们定义了$S = \sum_{i=1}^{N} x_i x_i^T$,这是一个未中心化的平方和矩阵(uncentered sum-of-squares matrix),相比中心化矩阵这样的更容易进行渐进的增量更新.

结果很直观:后验均值(posterior mean)就是对先验均值(prior mean)和最大似然估计(MLE)的凸组合(convex combination),附带上强度控制项$k_0 + N$.而后验散布矩阵(posterior scatter matrix)S_N就是先验散布矩阵(prior scatter matrix)S_0加上经验散布矩阵(empirical scatter matrix)$S_{\bar{x}}$,再加上由均值不确定性带来的附加项(这也创造了自己的一个虚拟散布矩阵(virtual scatter matrix)).

4.6.3.4 后验模(Posterior mode)

联合分布的众数(mode)如下所示:

$$\arg\max p(\mu, \Sigma | D) = (m_N, \frac{S_N}{v_N + D + 2}) \quad (4.215)$$

如果设置$k_0 = 0$,就降低(reduce)成了:

$$\arg\max p(\mu, \Sigma | D) = (\bar{x}, \frac{S_0 + S_{\bar{x}}}{v_N + N + D + 2}) \quad (4.216)$$

对应的估计$\hat{\Sigma}$几乎和等式4.183所述一样,唯一区别是分母上差了一个1,这是因为这个众数(mode)是联合分布的,而不是边缘分布的.

4.6.3.5 后验边缘分布

Σ的后验边缘分布就很简单了,如下所示:

$$p(\Sigma | D) = \int p(\mu, \Sigma | D) d\mu = IW(\Sigma | S_N, v_N) \quad (4.217)$$

这个边缘分布的众数(mode)和均值(mean)分别为:

$$\hat{\Sigma}_{map} = \frac{S_N}{v_N + D + 1}, E[\Sigma] = \frac{S_N}{v_N - D - 1} \quad (4.218)$$

不难发现对μ的后验边缘分布正好就是一个多元学生 T 分布:

$$p(\mu | D) = \int p(\mu, \Sigma | D) d\Sigma = T(\mu | m_N, \frac{1}{v_N - D + 1} S_N, v_N - D + 1) \quad (4.219)$$

这是由于学生分布可以表示做多个高斯分布（正态分布）的缩放混合，参考本书等式11.61.

此处参考原书图4.19

4.6.3.6 后验预测

后验预测(posterior predictive)如下所示:

$$p(x \mid D) = \frac{p(x, D)}{p(D)} \quad (4.220)$$

所以很容易用一系列边缘似然函数(marginal likelihood)的比值的形式来进行估算.
结果这个比值也是多元学生T分布:

$$p(x \mid D) = \iint N(x \mid \mu, \Sigma) NIW(\mu, \Sigma \mid m_N, k_N, S_N) d\mu d\Sigma \quad (4.221)$$

$$= T(x \mid m_N, \frac{k_N + 1}{k_N(v_N - D + 1)} S_N, v_N - D + 1) \quad (4.222)$$

4.6.3.7 标量数据的后验

现在把上面的结论用到一个特殊情况,即x_i是一维的.这些结果在统计领域中有很广泛的应用.如本书4.6.2.2所示,通常可能不适用正常逆威沙特分布(normal inverse Wishart),而是使用正常逆卡方分布(normal inverse chi-squared,缩写为NIX),定义如下所示:

$$NI\chi^2(\mu, \sigma^2 \mid m_0, k_0, v_0, \sigma_0^2) \stackrel{*}{=} N(\mu \mid m_0, \sigma^2 / k_0) \chi^{-2}(\sigma^2 \mid v_0, \sigma_0^2) \quad (4.223) \propto (\frac{1}{\sigma^2})^{(v_0+3)/2} \exp(-\frac{v_0\sigma_0^2 + k_0(]}{2\sigma}$$

图4.19所示为其图像.沿着μ轴,分布形状类似正态分布,而沿着σ^2轴分布形状就像是逆卡方分布(χ^{-2});整个联合概率密度函数的轮廓形状就像是压扁的蛋.有意思的是我们会发现μ的形状比较小数值的σ^2有更显著的峰值,这也很好理解,因为数据本身方差小(low variance),就能进行更准确的估计了.

后验如下所示

$$p(\mu, \sigma^2 \mid D) = NI\chi^2(\mu, \sigma^2 \mid m_N, k_N, v_N, \sigma_N^2) \quad (4.225) m_N = \frac{k_0 m_0 + N\bar{x}}{k_N} \quad (4.226) k_N = k_0 + N \quad (4.227) v$$

σ^2的后验边缘分布为:

$$p(\sigma^2 \mid D) = \int p(\mu, \sigma^2 \mid D) d\mu = \chi^{-2}(\sigma^2 \mid v_N, \sigma_N^2) \quad (4.230)$$

其后验均值为:$E[\sigma^2 \mid D] = \frac{v_N}{v_N - 2} \sigma_N^2$

μ的后验边缘分布为学生T分布,是学生分布的缩放混合形式,如下所示:

$$p(\mu \mid D) = \int p(\mu, \sigma^2 \mid D) d\sigma^2 = T(\mu \mid m_N, \sigma_N^2 / k_N, v_N) \quad (4.231)$$

其后验均值为:$E[\mu \mid D] = m_N$

如果我们使用下面的无信息先验,结果会是什么样呢?

$$p(\mu, \sigma^2) \propto p(\mu) p(\sigma^2) \propto \sigma^{-2} \propto NI\chi^2(\mu, \sigma^2 \mid \mu_0 = 0, k_0 = 0, v_0 = -1, \sigma_0^2 = 0) \quad (4.232)$$

有了上面的先验,后验形式如下所示:

$$p(\mu, \sigma^2 | D) = NI\chi^2(\mu, \sigma^2 | \mu_N = \bar{x}, k_N = N, v_N = N - 1, \sigma_N^2 = \text{是}^2) \quad (4.233)$$

上式中的:

$$s^2 \stackrel{*}{=} \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{N}{N-1} \sigma_{mle}^2 \quad (4.234)$$

就是标准取样偏差(sample standard deviation).在本书6.4.2中会说明这是一个对方差的无偏见估计(unbiased estimate).这样后验均值的边缘分布为:

$$p(\mu | D) = T(\mu | \bar{x}, \frac{s^2}{N}, N - 1) \quad (4.235)$$

而μ的后验方差为:

$$\text{var}[\mu | D] = \frac{v_N}{v_N - 2} \sigma_N^2 \quad (4.236)$$

上面这个后验方差的平方根就是均值的标准差(standard error of the mean):

$$\sqrt{\text{var}[\mu | D]} \approx \frac{s}{\sqrt{N}} \quad (4.237)$$

然后均值的估计95%后验置信区间(credible interval)为:

$$I_{.95}(\mu | D) = \bar{x} \pm 2 \frac{s}{\sqrt{N}} \quad (4.238)$$

(贝叶斯理论的置信空间在本书的5.2.2有更多讲解,而频率论的置信区间与之对比的内容在本书6.6.1.)

4.6.3.8 贝叶斯T检验

我们要检验一个假设:给定正态分布$x \sim N(\mu, \sigma^2)$,对某个未知值μ_0(通常都是0),$\mu \neq \mu_0$,这叫做双面单样本t检验(two-sided, one-sample t-test).简单方法就是检查$\mu_0 \in I_{0.95}(\mu | D)$是否成立.如果不成立,则有95%的信心认为$\mu \neq \mu_0$.更普遍的做法是检验两对样本是否有同样的均值.更确切来说,设$y_i \sim N(\mu_1, \sigma^2), z_i \sim N(\mu_2, \sigma^2)$.就可以使用$x_i = y_i - z_i$来验证是否有$\mu = \mu_1 - \mu_2 > 0$.可以用下面的形式来对这个量进行估计:

$$p(\mu > \mu_0 | D) = \int_{\mu_0}^{\infty} p(\mu | D) d\mu \quad (4.239)$$

这也叫做单面成对T检验(one sided paired t-text).(对未配对测试(unpaired test)有类似的方法,对比在二项比例(binomial proportions)上有所不同,本书5.2.3会介绍.)

要计算这个后验,必须要指定一个先验.设用一个无信息先验.如上所述,这样μ的后验边缘分布形式

为:

$$p(\mu \mid D) = T(\mu \mid \bar{x}, \frac{s^2}{N}, N-1) \quad (4.240)$$

然后我们定义下面的T统计(t statistic):

$$t^* = \frac{\bar{x}-\mu_0}{s/\sqrt{N}} \quad (4.241)$$

期中的分母是均值标准差.然后有:

$$p(\mu \mid D) = 1 - F_{N-1}(t) \quad (4.242)$$

上式中的$F_\nu(t)$是标准学生 T 分布$T(0, 1, \nu)$的累积密度函数(cdf).

4.6.3.9 和频率论统计学的联系

如果我们使用了无信息先验,就会发现上面的贝叶斯分析给出的结果和使用频率论方法推导的一样.(关于频率论统计学的内容在本书第六章会有详细讲解.)比如从上面的结果中,会看到有:

$$\frac{\mu-\bar{x}}{\sqrt{s/N}} \mid D \sim t_{N-1} \quad (4.243)$$

这和最大似然估计(MLE)的取样分布(sampling distribution)有一样的形式:

$$\frac{\mu-\bar{x}}{\sqrt{s/N}} \mid \mu \sim t_{N-1} \quad (4.244)$$

这是因为学生T分布是关于前两个参数(arguments)对称的(symmetric),所以有$T(\bar{x} \mid \mu, \sigma^2, \nu) = T(\mu \mid \bar{x}, \sigma^2, \nu)$;因此$\mu$的后验和$\bar{x}$的取样分布有一样的形式.结果导致了频率测试(frequentist test)返回的(单向(one sided))p值(在本书6.6.2中有定义)和贝叶斯方法返回的$p(\mu > \mu_0 \mid D)$一样.具体参考本书配套的PMTK3当中的bayesTtestDemo为例.

尽管看着非常相似,这两个结果还是有不同阐述的:在贝叶斯方法中,μ是未知的,而$\bar{x}$是固定的,而在频率论方法中正好相反,$\bar{X}$是未知的,而μ是固定的.使用无信息先验的简单模型时,频率论和贝叶斯方法之间的更多共同点可以参考(Box and Tiao 1973),本书的7.6.3.3也有更多讲解.

4.6.4 未知精度下的传感器融和*

本节会利用4.6.3当中的结论来解决传感器融合的问题,每个测量设备的精确度都不知道.这对本书4.4.2.2的结论进行了泛化,在4.4.2.2里是设测量模型的位置精确度服从正态分布.未知的精确度会导致有定量意义的不同结果,产生一个潜在的多态后验(multi-modal posterior).这里的内容参考了(Minka 2001e).

假如我们想要从多个来源汇集数据,来估计某个量$\mu \in R$,但是信号源的可靠性都不知道.例如有两个不同的测试设备 x 和 y,有不同的精确度:$x_i|\mu \sim N(\mu, \lambda_x^{-1}, y_i|\mu \sim N(\mu, \lambda_y^{-1}$.对两个设备各自进行独立测量,就得到了:

$x_1 = 1.1, x_2 = 1.9, y_1 = 2.9, y_2 = 4.2$(4.245)

对$\mu, p(\mu) \propto 1$使用一个无信息先验(non-imformative prior),使用一个无限宽度的正态分布$p(\mu) = N(\mu|m_0 = 0, \lambda_0^{-1} = \infty)$来模拟.如果$\lambda_x, \lambda_y$都知道了,那么后验就也是正态分布了:

$$p(\mu|D, \lambda_x, \lambda_y) = N(\mu|m_N, \lambda_N^{-1}) \quad (4.246)$$

$$\lambda_N = \lambda_0 + N_x\lambda_x + N_y\lambda_y \quad (4.247)$$

$$m_N = \frac{\lambda_x N_x \bar{x} + \lambda_y N_y \bar{y}}{N_x\lambda_x + N_y\lambda_y} \quad (4.248)$$

上式中的$N_x = 2, N_y = 2$分别是x和y的观测次数,而$\bar{x} = \frac{1}{N_x}\sum_{i=1}^{N} x_i = 1.5, \bar{y} = \frac{1}{N_y}\sum_{i=1}^{N} y_i = 3.5$.这是因为后验精度(posterior precision)是测量精度的综合,而后验均值是先验均值(这里是0)和数据均值的加权和.

不过测试精度还是不知道啊.开始用最大似然估计来估计一下吧.对数似然函数(log-likelihood)为:

$l(\mu, \lambda_x, \lambda_y) = \log\lambda_x - \frac{\lambda_x}{2}\sum_i(x_i - \mu)^2 + \log\lambda_y - \frac{\lambda_y}{2}\sum_i(y_i - \mu)^2$(4.249)

解出下面的联立方程,就能得到最大似然估计(MLE)了:

$$\frac{\partial l}{\partial \mu} = \lambda_x N_x(\bar{x} - \mu) + \lambda_y N_y(\bar{y} - \mu) = 0 \quad (4.250)$$

$$\frac{\partial l}{\partial \lambda_x} = \frac{1}{\lambda_x} - \frac{1}{N_x}\sum_{i=1}^{N_x}(x_i - \mu)^2 = 0 \quad (4.251)$$

$$\frac{\partial l}{\partial \lambda_y} = \frac{1}{\lambda_y} - \frac{1}{N_y}\sum_{i=1}^{N_y}(y_i - \mu)^2 = 0 \quad (4.252)$$

解出来就是:

$$\hat{\mu} = \frac{N_x\hat{\lambda}_x\bar{x} + N_y\hat{\lambda}_y\bar{y}}{N_x\hat{\lambda}_y + N_y\hat{\lambda}_y} \quad (4.253)$$

$$\frac{1}{\hat{\lambda}_x} = \frac{1}{N_x}\sum_i(x_i - \hat{\mu})^2 \quad (4.254)$$

$$\frac{1}{\hat{\lambda}_y} = \frac{1}{N_y}\sum_i(y_i - \hat{\mu})^2 \quad (4.255)$$

很明显,μ的最大似然估计(MLE)与后验均值m_N有同样的形式.

使用固定点迭代(fixed point iteration)就可以解出来了.首先初始化估计$\lambda_x = 1/s_x^2, \lambda_y = 1/s_y^2$,其中的$s_x^2 = \frac{1}{N_x}\sum_{i=1}^{N_x}(x_i - \bar{x})^2 = 0.16, s_y^2 = \frac{1}{N_y}\sum_{i=1}^{N_y}(y_i - \bar{y})^2 = 0.36.$
然后就解出来了$\hat{\mu} = 2.1154$,所以有$p(\mu|D, \hat{\lambda}_x, \hat{\lambda}_y) = N\mu|2.1154, 0.0554)$.如果现在进行迭代,最终会收敛到:$\hat{\lambda}_x = 1/0.1662, \hat{\lambda}_y = 1/4.0509, p(\mu|D, \hat{\lambda}_x, \hat{\lambda}_y) = N(\mu|, 1.5788, 0.0798).$

对这个后验的插值估计如图4.20(a)所示.每个传感器的权重是根据其估计误差赋予的.由于估计误差标明传感器y远不如传感器x可靠,所以就有$E[\mu|D\hat{\lambda}_x, \hat{\lambda}_y] \approx \bar{x}$,实际上就是忽略了传感器y.

接下来我们用贝叶斯方法来来积分求未知精度,而不对其进行估计.也就是要计算:

$$p(\mu|D) \propto p(\mu)[\int p(D_x|\mu, \lambda_x)p(\lambda_x|\mu)d\lambda_x][\int p(D_y|\mu, \lambda_y)p(\lambda_y|\mu)d\lambda_y] \quad (4.256)$$

使用无信息Jeffrey先验(uninformative Jeffrey's priors)$p(\mu) \propto 1, p(\lambda_x|\mu) \propto 1/\lambda_x, p(\lambda_y|mu) \propto 1/\lambda_y$.x和y两项对称,所以只看其中一个就可以了.关键的积分步骤是:

$$I = \int p(D_x|\mu, \lambda_x)p(\lambda_x|\mu)d\lambda_x \propto \int \lambda_x^{-1}(N_x\lambda_x)^{N_x/2} \quad (4.257) \exp(-\frac{N_x}{2}\lambda_x(\bar{x} - \mu)^2 - \frac{N_x}{2}s_x^2\lambda_x)d\lambda_x \quad (4.258)$$

利用$N_x = 2$来简化到:

$$I = \int \lambda_x^{-1}\lambda_x^1 \exp(-\lambda_x[(\bar{x} - \mu)^2 + s_x^2])d\lambda_x \quad (4.259)$$

看出来了吧,这个和一个非正则γ密度函数(unnormalized Gamma density)的积分成正比:

$$Ga(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b} \quad (4.260)$$

其中的$a = 1, b = (\bar{x} - \mu)^2 + s_x^2$.因此这个积分也就和$\gamma$分布的归一化常数(normalizing constant) $\Gamma(a)b^{-a}$成正比,就得到了:

$$I \propto \int p(D_x|\mu, \lambda_x)p(\lambda_x|\mu)d\lambda_x \propto [(\bar{x} - \mu)^2 + s_x^2]^{-1} \quad (4.261)$$

然后后验则成了:

$$p(\mu|D) \propto \frac{1}{(\bar{x}-\mu)^2+s_x^2}\frac{1}{(\bar{y}-\mu)^2+s_y^2} \quad (4.262)$$

具体的后验如图4.20(b)所示.可以看到有两个众数(mode),分别在$\bar{x} = 1.5, \bar{y} = 3.5$这两个位置.对应的就是x传感器比y传感器更精确.第一个众数(mode)的权重更高,因为x传感器给出的数据互相更接近,所以看上去就是这个传感器更可靠.(很明显不可能两个都可靠,因为他们给出的值都不一样的.)不过贝叶斯的方案保持了开放性,就是始终保持了y传感器可能更可靠的概率;从两次测量,其实还不能说就按照差值估计得到的结果一样来选择x传感器,这个结果可能过分有信心了,后验太窄了.

练习略.