§3.5 Bayes学习

递推的贝叶斯估计

考虑贝叶斯估计的收敛性：记学习样本个数为N，样本集 $\mathcal{X}^N=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$，$N>1$ 时有：

$$p(\mathcal{X}^N\mid\theta)=p(\mathbf{x}_N\mid\theta)p(\mathcal{X}^{N-1}\mid\theta)$$

而贝叶斯估计关于待估计参数的后验概率公式为

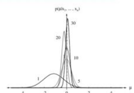
$$p(\theta\mid\mathcal{X})=\frac{p(\mathcal{X}\mid\theta)p(\theta)}{\int_\Theta p(\mathcal{X}\mid\theta)p(\theta)d\theta} \Rightarrow p(\theta\mid\mathcal{X}^N)=\frac{p(\chi^N\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_\Theta p(\chi^N\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta}=\frac{p(x_N\mid\theta)p(\chi^{N-1}\mid\theta)p(\theta)}{\int_\Theta p(x_N\mid\theta)p(\chi^{N-1}\mid\theta)p(\theta)d\theta}$$

$$p(\boldsymbol{\theta}\mid\chi^{N-1})=\frac{p(\chi^{N-1}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_\Theta p(\chi^{N-1}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta} \qquad =\frac{p(x_N\mid\boldsymbol{\theta})p(\theta\mid\chi^{N-1})}{\int p(x_N\mid\boldsymbol{\theta})p(\theta\mid\chi^{N-1})d\boldsymbol{\theta}}$$

及 $p(\theta\mid\mathcal{X}^0)=p(\theta)$，则随着样本数增多，可得后验概率密度函数序列：

$p(\theta)$，$p(\theta\mid\mathbf{x}_1)$，$p(\theta\mid\mathbf{x}_1,\mathbf{x}_2)$，$\cdots$ —— 参数估计的递推贝叶斯方法

贝叶斯学习

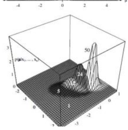


贝叶斯估计关于待估计参数的后验概率公式

$$p(\theta\mid\mathcal{X}^N)=\frac{p(x_N\mid\boldsymbol{\theta})p(\theta\mid\chi^{N-1})}{\int p(x_N\mid\boldsymbol{\theta})p(\theta\mid\chi^{N-1})d\boldsymbol{\theta}}$$

贝叶斯估计关于待估计参数的后验概率序列

$p(\theta)$，$p(\theta\mid\mathbf{x}_1)$，$p(\theta\mid\mathbf{x}_1,\mathbf{x}_2)$，$\cdots$



如果随着样本数的增加，待估计参数后验概率序列逐渐尖锐，趋向于以参数真实值为中心的一个尖峰，当样本无穷多时收敛于在参数真实值上的脉冲函数，则称这一过程为**贝叶斯学习**。

§3.6 正态分布下的贝叶斯估计

单变量正态分布$p(x\mid\mu)\sim N(\mu,\sigma^2)$， 已知 σ^2，估计 μ，假设先验分布$p(\mu)\sim N(\mu_0,\sigma_0^2)$。

平方误差损失函数条件下求解 Bayes 估计步骤：

（1）确定 θ 的先验分布 $p(\theta)$；

（2）由样本集 $\chi=\{x_1,x_2,\ldots,x_N\}$ 求样本联合分布 $p(\chi|\theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left[-\frac{1}{2}\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]$

（3）求 θ 的后验分布 $p(\boldsymbol{\theta}|\chi)=\dfrac{p(\chi|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_\Theta p(\chi|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$

（4）$\hat{\boldsymbol{\theta}}=\int_\Theta \boldsymbol{\theta}p(\boldsymbol{\theta}|\chi)d\boldsymbol{\theta}$

现（1）已知。 $\theta=\mu$

(3) 计算后验分布 $p(\mathbf{x}_k|\mu)\sim N(\mu,\sigma^2)$ $p(\mu)\sim N(\mu_0,\sigma_0^2)$

$$p(\mu\mid\chi)=\frac{p(\chi|\mu)p(\mu)}{\int_\Theta p(\chi|\mu)p(\mu)d\mu}=\alpha p(\chi\mid\mu)p(\mu)=\alpha\prod_{i=1}^{n}p(x_i\mid\mu)p(\mu)$$

$$=\alpha\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

here, $\alpha=1\Big/\int_\Theta p(\chi|\mu)p(\mu)d\mu$

(3) 计算后验分布

$$p(\mu\mid\chi)=\alpha\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}$$

$$=\alpha'\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]\right\}$$

$$=\alpha'\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(x_k^2+\mu^2-2\mu x_k\right)-\frac{1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu\mu_0)\right\}$$

$$=\alpha'\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n}x_k^2+n\mu^2-2\mu\sum_{k=1}^{n}x_k\right)-\frac{1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu\mu_0)\right.$$

其中：$\alpha'=a\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\frac{1}{\sqrt{2\pi}\sigma_0}$

$$p(\mu\mid\chi)=\alpha'\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n}x_k^2+n\mu^2-2\mu\sum_{k=1}^{n}x_k\right)-\frac{1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu\mu_0)\right\}$$

$$=\alpha'\exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]-\frac{1}{2}\underbrace{\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k^2+\frac{1}{\sigma_0^2}\mu_0^2\right)}_{\text{无关项}}\right\}$$

$$=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

由此可见 $p(\mu|\chi)$ 是 μ 的二次函数的指数函数，因此 $p(\mu|\chi)$ 满足正态分布

(3) 计算后验分布

$$p(\mu|\chi)=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}x_k+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

$\because p(\mu|\chi)\sim N(\mu_N,\ \sigma_N^2)$

$$p(\mu|\chi)=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right]\right)$$

$$=\alpha''\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_N^2}\mu^2-2\frac{\mu_N}{\sigma_N^2}\mu\right]\right\}$$

$$\begin{cases}\dfrac{1}{\sigma_N^2}=\dfrac{N}{\sigma^2}+\dfrac{1}{\sigma_0^2}\\[2mm]\dfrac{\mu_N}{\sigma_N^2}=\dfrac{1}{\sigma^2}\sum_{k=1}^{N}x_k+\dfrac{\mu_0}{\sigma_0^2}=\dfrac{N}{\sigma^2}m_N+\dfrac{\mu_0}{\sigma_0^2}\end{cases}$$

$$\begin{cases}\mu_N=\dfrac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\dfrac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0\\[2mm]\sigma_N^2=\dfrac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\end{cases}$$

$m_N=\frac{1}{N}\sum_{k=1}^{N}x_k$

(4) $\hat{\boldsymbol{\theta}}=\int_\Theta\boldsymbol{\theta}p(\boldsymbol{\theta}|\chi)d\boldsymbol{\theta}$

$$\begin{cases}\mu_N=\dfrac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\dfrac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 & \boxed{N\to\infty\quad\mu_N\to m_N}\\[2mm]\sigma_N^2=\dfrac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2} & \boxed{N\to\infty\quad\sigma_N^2\to0}\end{cases}$$

$$\hat{\mu}=\int_\Theta\mu p(\mu|\chi)d\mu=\int_\Theta\mu\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right]d\mu$$

$$=\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

$m_N=\frac{1}{N}\sum_{k=1}^{N}x_k$

样本信息与先验知识的线性组合

$$\begin{cases}N=0,\ \hat{\mu}=\mu_0;\ N\to\infty,\ \hat{\mu}\to m_N\\ \sigma_0^2=0\text{时}，\hat{\mu}=\mu_0\ \text{先验可靠}\\ \sigma_0\gg\sigma\text{时}，\hat{\mu}=m_N\ \text{先验不确定}\end{cases}$$

σ_0^2↑先验不定

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\boxed{m_N}+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

样本均

$$\sigma_N^2=\frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\sigma^2$$

样本方差

正态分布的贝叶斯学习

单变量正态分布$p(x\mid\mu)\sim N(\mu,\sigma^2)$， 已知 σ^2，估计 μ。假设先验分布$p(\mu)\sim N(\mu_0,\sigma_0^2)$。

$p(\mu|\chi)\sim N(\mu_N,\ \sigma_N^2)$

Bayes 学习是利用 θ 的先验分布及样本提供的信息求出 θ 的后验分布 $p(\theta|\chi)$，然后直接求总体分布。

$$p(\boldsymbol{x}|\chi)=\int_\Theta p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\chi)d\boldsymbol{\theta}=\int_\Theta\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\times\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right]d\mu$$

$$=\frac{1}{2\pi\sigma\sigma_N}f(\sigma,\sigma_N)\exp\left[-\frac{1}{2}\left(\frac{x-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right]\Rightarrow N(\mu_N,\sigma^2+\sigma_N^2)$$

$$f(\sigma,\sigma_N)=\int\exp\left[-\frac{\sigma_N^2+\sigma^2}{2\sigma_N^2\sigma^2}\left(\mu-\frac{\sigma_N^2x-\sigma^2\mu_N}{\sigma_N^2+\sigma^2}\right)^2\right]d\mu$$

正态分布的贝叶斯学习

$$\begin{cases}\mu_N=\dfrac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\dfrac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 & \boxed{N\to\infty\quad\mu_N\to m_N}\\[2mm]\sigma_N^2=\dfrac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2} & \boxed{N\to\infty\quad\sigma_N^2\to0}\end{cases}$$

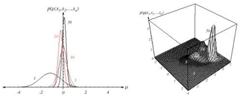


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- 当观察一个样本时，$N=1$ 就会有一个 μ 的估计值的修正值;当观察 $N=4$ 时，对 μ 进行修正，向真正的 μ 靠近；当观察 $N=9$ 时，对 μ 进行修正，向真正的 μ 靠的更近
- 当N↑，μ_N就反映了观察到N个样本后对μ的最好推测，而σ_N^2反映了这种推测的不确定性，N↑，σ_N^2↓，σ_N^2 随观察样本增加而单调减小，且当$N\to\infty$，$\sigma_N^2\to0$
- 当N↑，$P(\mu|\chi)$越来越尖峰突起。$N\to\infty$, $P(\mu|\chi)\to$冲击函数，这个过程称为贝叶斯学习。