四、5、贝叶斯估计

2016年3月5日 星期六　　下午12:53

简介

贝叶斯估计认为，即使训练集已知，待估计的参数仍然是未知的，但是待估计的参数服从某种特定的分布，因此可以假设待估计参数服从一个先验的分布，训练数据作为观测数据，见过观测数据之后，对待估计参数有一个更准确的估计（把先验概率转化成后验概率密度），这个估计就是我们待求的参数。这个方法我们称作贝叶斯估计。

我的理解，之所以叫做贝叶斯估计，是因为1、参数估计的思想借鉴贝叶斯学派2、参数估计的过程是使用贝叶斯公式，通过逐步增加观测数据，把参数θ的先验概率变成后验概率，最终产出一个更准确的参数θ的概率分布。

算法原理

整体思路

第一步、贝叶斯估计：

使用贝叶斯公式，通过逐步增加观测数据$x \in D$，把参数θ的先验概率$P(\theta)$变成后验概率$P(\theta|D)$，从而产出一个更准确的参数θ的概率分布。这个更准确的概率分布其实就是后验概率$P(\theta|D)$，得到参数θ的概率分布，贝叶斯估计就已经完成了。

第二步、计算类条件概率：

在得到参数θ更准确的概率分布基础上，往往要结合实际问题，用到实际问题中。这里以贝叶斯分类器为例，我们要做一个贝叶斯分类器，最终目标是要通过参数θ估计的是类条件概率：对于第i类：$P(x|w_i,\theta_i)$，即$P(x|\theta_i)$，一

般省去类下标写成$P(x|\theta)$。

这里可以回顾一下，类条件概率的作用是带入贝叶斯分类器的贝叶斯公式中，计算出后验概率。后验概率是贝叶斯分类器分类的依据

由于这里我们得到的是θ的概率分布，（而不仅仅是最大似然估计中，得到参数的一个固定值），所以我们计算类条件概率时可以充分利用“概率分布”的强大之处。我们对θ在各种取值情况下产出的类条件概率做积分。

$$P(x|D)=\int P(x|\theta)P(\theta|D)d\_\theta。$$

注意这里类条件概率写成$P(x|D)$而非$P(x|\theta)$是因为θ是一个概率分布，不是固定值，之前的“在参数θ情况下，x的概率”写成“在数据集D情况下，θ在各种取值下的x的概率的积分”。其中$P(\theta|D)$已经估计出，$P(x|\theta)$后面会简化，至此可得出类条件概率$P(x|D)$

求解目标

第一步：$P(\theta)$的后验概率$P(\theta|D)$

第二步：类条件概率$P(x|D)$（通常表示为$P(x|\theta)$，只不过贝叶斯估计比较特殊）

已知与假设

$P(\theta)$

我们要估计$P(\theta)$的概率（其实是概率分布），需要先假定$P(\theta)$服从概率分布的形式，我们假定他服从高斯分布。

$P(\theta)\sim N(\mu\_0|\sigma\_0\,\hat{}2)$

$P(x|\theta)$

假设θ中协方差$\sum$是已知的，只有μ是未知的。

而且$P(x|\mu)\sim N(\mu|\sum)$服从正态分布

这里两个μ是一样的，是正好一致吗，可能是，没太理解，详

见模式分类P74

单变量情况

我们先讨论单变量情况，（变量x等是一维的，而不是多维特征），多变量情况其实是单变量基础上使用向量、矩阵乘法得到的详见模式分类P77。

求解过程（尤其在特殊情况下）

第一步：P（θ | D）

贝叶斯公式求解过程

由于θ中只有μ是未知的，表示为P（μ | D），根据贝叶斯公式

P（μ | D）= P（D | μ ）* P （μ）/ P（ D）

其中P（D | μ ）= ∏ P（x_k | μ ）　对于D中的n个样本x_k，∏是连乘符号

P（x_i | μ ）服从高斯分布

P （μ）服从高斯分布

P（ D）看做常量，记作1/a

进而

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)} \\
&= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \\
&= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]
\end{aligned}
$$

两个高斯分布经过化简成为了一个高斯分布，P（μ | D）也服从高斯分布。P（μ | D）的均值μ_n和方差σ_n可以用之前那两个高斯分布的均值方差表示为

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

其中 μ_n^ 是数据集样本均值

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

分析：

从结果中，我们能看出来对于待估计参数θ先验知识和样本观测是如何结合的。

μ_n是使得θ概率最大的参数估计，σ_n是经过多次观测之后，仍然留有的对估计的不确定程度。可以看出

A σ_n随样本n增加是单减的。

B μ_n随样本增加趋近样本均值

C 随着样本增加P（μ | D）越来越尖，这个过程称为贝叶斯学习过程。

D 特殊情况：σ_0 = 0，说明先验十分置信，任何样本都无法改变后验概率。

E 特殊情况：σ_0 >> σ，说明先验十分不置信，直接把样本μ当做参数的后验的均值即可。

第二步：P（x | D）

得到参数估计P（μ | D）之后，我们要求类条件概率P（x | D）（其他算法中表示为P（μ | θ））

利用贝叶斯估计的优势：求得的参数θ是一个概率分布，而非特定值。求解P（x | D）时要结合概率分布，所以：

P（x | D）= ∫ P（x | θ）P（θ | D）d_θ

由于参数θ只有μ未知

P（x | D）= ∫ P（x | μ）P（μ | D）d_μ

P（x | μ）服从高斯分布

P（μ | D）我们已经求得，而且也服从高斯分布

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\,d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n)
\end{aligned}
$$

其中

$$
f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu
$$

所以P（x | D）也服从高斯分布

$$
p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2)
$$

至此得到P（x | D）

最终结果上，贝叶斯估计得到的是类条件概率的概率分布，最大似然估计只得到均值和方差的值