四、4、极大似然估计

2016年3月5日 星期六　　下午12:54

概念

极大/最大似然估计/法，MLE，Maxinum Likelihood Estimation

参数估计

我的理解是，根据观测数据，找到一组合适的参数去刻画观测的数据。

参数刻画数据的方式，往往是根据一些特定的先验。比如，机器学习问题中，我们选定的机器学习模型是先验，告诉我们参数刻画数据的方式，针对不同数据，模型的参数是不一样的，那么对于当前的数据，我们要找出最适合的那一组参数。这个过程叫做参数估计。

参数估计的方法，在统计学中有很多景点的解法，这里我们介绍极大似然估计和贝叶斯估计。

特点

与贝叶斯估计区别

概念

统计学界有两大学派，1）频率主义学派（Frequentist）2）贝叶斯学派（Bayesian）。

对于参数估计，频率主义认为，对于已知的训练集，待估计的参数虽然是未知的，但是是固定的，我们要做的就是根据训练数据，找到能让优化目标取最大值的那组参数，这组参数是让就是最大似然的参数。这种方法叫做极大似然估计。

贝叶斯主义认为，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。

相同点

A 从结果看，得到的结果通常是很接近的。

B 这两中方法，是参数估计最常用的方法。

C 如果是放到贝叶斯分类器中，我们都是使用后验概率作为分类准则。

不同点

A 最大似然估计的结果是一组确定的参数，贝叶斯估计的结果是参数的概率分布，只不过通常贝叶斯估计得到新的观测样本之后，概率密度函数变得更加尖锐，待估计参数在真实值（最优值）附近形成最大的尖峰。这个过程我们称作贝叶斯学习过程。尖峰处对应的参数，就是我们待求的参数。

B 从概念中我们可以看出，贝叶斯估计需要额外的先验信息。

一方面，这使得贝叶斯估计普适性、鲁棒性变差，如果实际的分布和我的假设有出入，或者训练数据不完美，不是很好的服从分布。贝叶斯估计会有偏差。而最大似然估计不存在这样的问题。

另一方面，由于贝叶斯估计使用了更多的先验信息，如果先验信息正确，贝叶斯估计的效果会更好。（记得牛博说过，但是我不确定，资料中没有说）

C 贝叶斯估计要比最大似然估计难理解

D 我看机器学习算法中，使用最大似然估计的比使用贝叶斯估计的多

E 最大似然估计对样本也有要求：独立同分布？？？？？

待补充！！！模式分类P81

（这一区别很重要，在模式分类P67，机器学习-周志华P149都有不错的描述）

基本原理

我们以分类问题为例，介绍最大似然估计。

假设

假设

我们假设样本都是独立同分布（i.i.d）的，而且每一个类条件概率密度P（x | wj）都是确定的，其形式也是已知的，只不过其参数是未知的，现在我们来估计类条件概率密度P（x | wj）。

由于形式是已知的，我们假设服从高斯分布（这样看来和贝叶斯估计很接近了）。我们假设各个类之间是c个独立的问题（c表示类数）。参数是未知的，为了强调这一点，我们把P（x | wj）写成类条件概率密度P（x | wj，θj）。

优化目标

对于整个训练集D中n个样本

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

P（D | θ）表示似然函数，他是一个以θ为自变量的函数。仅仅表示参数θ对训练集的估计，是我们最大似然估计的最大化的目标，而不是概率密度。和之前提到的类条件概率密度P（x | wj）完全不一样，是两个概念这里要区分开。

我们在学习的过程中，如果把P（D | θ）画出来，可以看到不断收敛（模式分类P69）。

优化算法

既然我们找到了最大化的目标，而且是连续可导的，就求导。如果是凸函数，令导数=0即找到闭式解；如果是非凸的，可以对多个导数=0的点逐个判断，或者用常规的数值优化方法，如sgd，牛顿法等。

解的形式

最大似然估计对P（x | wj，θj）的估计（其实也就是对P（xi | θ）的估计），高斯分布的均值和协方差是：

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

$$\hat{\quad} \quad \frac{1}{\quad} \sum^{n}$$

$$\Sigma = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

注意，这里的均值、方差指的是在某一个类别ci上的情况。这样看着就好理解了，每个类对P（xi|θ）估计，均值是类中心，方差是所有样本点协方差矩阵元素对位累加的和。

解的过程模式分类P71只推导了单变量的情况。

估计上的偏差

在训练集规模小的情况下，对协方差的估计有点偏差，需要乘一个系数。详见模式分类P72