四、贝叶斯

2016年2月14日 星期日　　下午3:57

一、简介

二、贝叶斯决策论

概念

贝叶斯决策论是在概率框架写实施决策的基本方法。（机器学习-周志华）

利用改了的不同分类决策，与相应的决策代价之间的定量折中，决策问题以概率的形式来描述（模式分类）

解释

比如，我们要做分类任务“判断下一条捕捞到的是鲈鱼还是桂鱼”。在实施观察前，不知道任何关于下一条的信息。我们现在只能根据“先验概率”，即“河里鲈鱼和桂鱼的比例是多少”，进行判断。

但是如果我们观测到了待预测数据的特征，比如“这条鱼的光泽度怎么样”，我们预测起来就更有把握。这样，加入数据的特征之后，我们预测结果更加置信，预测出的概率也会被更新。这个加入观测数据信息的概率叫做“后验概率”。

对于下一条鱼这个样本，我们用x表示，鲈鱼/桂鱼这两类用wj表示，我们用概率论中的贝叶斯公式可以表示为

$$P(\omega_j|x)=\frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

即

$$posterior=\frac{likelihood \times prior}{evidence}$$

P（wj）是我们的类先验概率，河里两种鱼的比例是多少。

P（wj | x）后验概率，这是我们待求解的概率。先验概率加入观测数据之后，对事件更准确的概率估计。在具体问题中，常常是“假

数据之后，对事件更准确的概率估计。在具体问题中，常常是"假定有一个样本，这个样本是什么类的"，这是我们需要的。

P（x）是证据因子，因为后验概率是由先验概率和似然函数决定的，P（x）这里可以仅仅看做是一个标量因子。如果一定要深究其含义的话，这里"捕捞到鱼、而且是x鱼"事件对应的概率形式应该是联合概率P（x，wj），但是一般我们希望分类器做的是，"假定有一个样本，这个样本是什么类的"，是条件概率，也就是后验概率的形式P（wj | x）。

P（x | wj）似然函数，似然。这是最关键也最难理解的一个参数。他一般被称作类条件概率密度（class-conditional probability density）、状态条件概率密度。代表类别取wj时，数据（或者说是特征）x的概率分布。这个信息帮助我们算出后验概率，这个信息是我们从训练数据学习出来的。

误差、风险

分类效果的好坏怎么衡量，效果怎么优化，我们需要对分类有个评估指标。

单次判断的误差

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{如果判定 } \omega_2 \\ P(\omega_2|x) & \text{如果判定 } \omega_1 \end{cases}$$

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)]$$

对所有样本

$$P(error) = \int_{-\infty}^{\infty} P(error, x)\, dx = \int_{-\infty}^{\infty} P(error|x)p(x)\, dx$$

更严格、更一般的我们定义

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

R（αi | x）称为条件风险，代表使用分类器预测样本x带来的风险。这里λ（αi | wj）表示在已知wj类下，做出ai行为可能带来的损失。我们可以最小化条件风险，贝叶斯决策过程就是提供一个总风险优化的过程

风险优化的过程。

引入λ（αi | wj）是一种严格的表示，很多问题中预测出是wj类之后，行为ai是固定的（只有一种），各种行为的损失也是相等的，所有很多问题中λ（αi | wj）恒定。

在全体数据集上总风险表示为

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$$

最小化后的总风险值，称为贝叶斯风险，他是可获得的最优结果。

极小化极大准则

这里多说一句，有时候我们追求在先验概率取任何值的情况下，都让分类器有风险尽量小。这里有一种方法叫极小化极大准则，就是我们找到让风险取最大值的先验概率，我们让这种情况下（先验概率取这个值，使得风险最大）的风险最小。

不过这并不是主流做法，详见模式分类P21

三、朴素贝叶斯分类器

naive Bayes，我们要根据贝叶斯决策论的思想，构造一个基本的分类器。

优化目标的形态：

从贝叶斯决策论中，很自然看出我们的目标是，让条件风险最小化R（αi | x）最小，就是最小误差概率P（error | x），也就是最大后验概率P（wj | x）。

参数估计的形式：

从贝叶斯决策论中，我们知道为了找到最大后验概率，我们需要知道1）先验概率2）类条件概率P（x | wj）（似然函数）。其中，类条件概率更难求解，是我们的核心目标。

最简单的求解概率的想法是基于频率统计，统计出每个类别在每个特征下的频次，算出概率，预测的时候根据样本的特征对类别进行

估算，这样的话由于样本空间有2^d种取值的可能（d是特征维度）。往往过大，训练样本覆盖不到，所以不可用。

我们回想概率论中，如何对一个概率做成估计，首先概率是服从一定概率分布的，我们可以先确定（或者假定）概率分布，然后用现有的训练数据拟合概率分布，估计出概率分布中的未知参数，从而得到P（x | wj）概率密度。

P（x | wj）基于高斯分布的假设，得出判别函数的形式

朴素贝叶斯分类器中，我们假设P（x | wj）服从高斯分布。由于一般输入x是由多个特征构成的一个向量，我们得假设各个特征都符合高斯分布，用多元密度函数（详见模式分类P26）的形式进行表示。

为了计算方便，我们根据对数函数单调且连续可导的特性，使用对数似然代替之前的似然函数。

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

带入高斯分布的概率密度

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

化简为关于x的二次形式

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

其中二次、一次、常数项分别是

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

对应的判别面是超二次曲面（超平面/超平面对/超球体/超脱球体/超双曲面），图见模式分类 P33

对于一些特殊情况，可以得出一些有意思的结论，详见模式分类

对于一些特殊情况，可以得出一些有意思的结论，详见模式分类P28

比如，对于各个特征统计独立且方差相同的情况，（协方差矩阵是对角阵，且是方差σ^2与单位阵I的乘积）.判别函数是一次形式

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

且超平面发向量是w

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$$

两个类均值的连线的中点过分类超平面。

四、极大似然估计

详见四、4、极大似然估计

五、贝叶斯估计

详见四、5、贝叶斯估计

六、期望最大化算法

详见四、6、期望最大化算法

七、半朴素贝叶斯分类器（低优）

八、贝叶斯网（中优）

九、隐马尔可夫模型（中优）

十、条件随机场