朴素贝叶斯分类器

贝叶斯学派的思想可以概括为先验概率p(x|c)+数据=后验概率p(c|x)。也就是说我们在实际问题中需要得到的后验概率，可以通过先验概率p(x|c)和数据一起综合得到。

- 数据大家好理解，被频率学派攻击的是先验概率，一般来说先验概率就是我们对于数据所在领域的历史经验，但是这个经验常常难以量化或者模型化，于是贝叶斯学派大胆的假设先验分布的模型，比如**正态分布，beta分布**等，
- 每次实验数据独立同分布。

以上假设一般没有特定的依据，因此一直被频率学派认为很荒谬。

贝叶斯理论

贝叶斯理论是概率框架下实施决策的基本方法。对分类任务来说，在所有相关概率都已知的**理想情形**下，贝叶斯决策论考虑如何基于这些概率和误判损失来选择最优的类别标记。

具体来说，若我们决策的目标是最小化分类错误率，贝叶斯最优分类器要对每个样本 x，选择能使后验概率 P(c | x)最大的类别 c 标记。在现实任务中后验概率通常难以直接获得。从这个角度来说，机器学习所要实现的是基于有限的训练样本集尽可能准确地估计出后验概率 P(c | x)。大体来说，主要有两种策略：给定x，可通过直接建模P(c | x)来预测c，这样得到的是“判别式模型”，例如，决策树、BP神经网络、支持向量机等等；也可先对联合概率分布P(x,c)建模，然后在由此获得P(c | x)，这样得到的是“生成式模型”。对于生成式模型来说，必然考虑

$$P(c|x) = \frac{P(c) * P(x|c)}{P(x)}$$

其中，类先验概率 P(c) 表达了样本空间中各类样本所占的比例，根据大数定律，当训练集包含充足的独立同分布样本时，P(c) 可通过各类样本出现的频率来进行估计，而P(x | c)代表先验概率，对于 P(x | c) 来说，由于它涉及关于 x **所有属性**的联合概率，直接根据样本出现的频率来估计将会遇到严重的困难。假如样本的d个属性都是二值的，则样本空间将有2的d次方种可能取值。在现实中，这个种类数往往大于训练样本，也就是说，很多样本取值在**训练集**中根本**没有出现**，直接使用频率来估计 P(x | c)显然不可行，因为“**未被观测到**”与“**出现概率为零**”通常是不同的。这可以通过**极大似然估计**来解决。

贝叶斯假设

基于贝叶斯公式来估计后验概率P(c | x)的主要困难在于：类条件概率P(x | c)是所有属性上的联合概率，难以从有限的训练样本直接估计而得。因此朴素贝叶斯分类器采用了“属性条件独立性假设”：对已知类别，假设所有属性相互独立。也就是说，假设每个属性独立的对分类结果发生影响。

基于属性独立性假设，后验概率P(c | x)可写为

$$P(c|x) = \frac{P(c)}{P(x)} \prod_{i=1}^{d} P(x_i|c)$$

所以朴素贝叶斯计算过程存在两点：

- 训练过程就是基于训练集D来估计类先验概率P(c)，
- 为每个属性估计条件概率P(x|c)。

由于对所有类别来说是等概率的，所以P(x)相同，例如二分类，P(x=1) = 0.5, P(x=0) = 0.5.因此上述后验概率估计可类别极大似然估计

$$L(x_i;c) = argmax_{c\in y}P(c) * \prod_{i=1}^{d} P(x_i|c)$$

计算P(c)

$$P(c) = \frac{|D_c|}{|D|}$$

其中D_c表示训练集合中第c类样本组成集合的数量，而D表示总样本数量。

对于离散型，计算P(x_i|c)

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

其中D_c,x_i代表第c类样本组合集合中取值为i的属性上属于x_i样本数量，例如二分类，就是第c类样本组合总属于0或1的数量；

对于连续性型，计算P(x_i|c)

$$p(x_i|c) = \frac{1}{\sqrt{2\pi\delta_{c,i}}}exp(-\frac{(x-\mu_{c,i})^2}{2\delta^2})$$

拉普拉斯修正

为了避免因训练集不充分而导致概率估值为零的问题，需要对概率估计进行修正，即P(c)和P(x_i|c)

$$P(c) = \frac{|D_c|+1}{|D|+N}; P(x_i|c) = \frac{|D_{c,x_i}|+1}{|D_c|+N_i}$$

并且在训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值渐趋向于实际概率值。