贝叶斯

1.基本思想：

对给定的x，我们想要得到的是x的类别。假设类别共有ck种，贝叶斯分类的思想是分别计算x属于每一种类别的概率P(ck|x)，然后根据后验概率最大化原则选择概率最大的类别作为最终输出。

后验概率 = 先验概率*似然

意思是说我们先预估计一个概率P(A),然后进行实验调整，重新得到后验概率P(A|B),似然P(B|A)/P(B)就是个调整因子

2.后验概率最大化的含义：

对一个样本点xi来说，它可能被分为ck中的任一类cj，对应的会有它被分为每一类的损失λij，对样本点xi来说，总的损失是一个期望Loss(xi)=E(ci|xi)=∑λijP(cj|x)，即对j从1到k求和。对所有的样本点来说，总的损失Loss也是一个期望E。根据机器学习的一般思想(期望风险最小化准则)，我们要最小化这个总的损失。根据贝叶斯判定准则(为最小化总体损失，只需在每个样本上选择那个能使该样本的损失Loss(xi)最小的类别标记)，我们只需要对每个xi，选择损失最小的cj，也就是argminLoss(xi),当损失函数选择0-1函数时(λij等于0或者1)，就有下式：

$$\begin{aligned}\text{argminLoss}(x_i)&=\text{argmin}\sum\lambda_{ij}P(c_j|x)\\&=\text{argmin}\sum P(y\neq c_j|x)\\&=\text{argmax}\sum P(y=c_j|x)\end{aligned}$$

由此就得到了后验概率最大化准则。

3.整体流程

对于一个样本点x，我们要计算它属于每个类的概率P(ck|x)，选择概率最大的那个类别输出，但是这个概率没有办法直接求出。但是根据贝叶斯公式，我们可以通过求出联合概率P(x,ck)来求得P(ck|x)。对于联合概率，我们又可以利用贝叶斯公式通过求P(x|ck)和P(ck)两个概率求出。但在计算P(x|ck)时有一个问题，x是一个向量，有很多分量，会产生维数灾难，这就需要独立性假设的帮忙：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$

$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

就是说i每个维上的特征相互独立。于是，我们有了一下推导：

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)}$$

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}$$

这个公式用来计算样本点x属于类别ck的概率，我们最终要的是概率最大的那个类别，就是说计算x属于每个类别ck的概率，选择概率最大的那个类别ck作为输出：

$$y=\arg\max_{c_k}P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)$$

训练过程其实就是在计算先验概率和似然。连乘可能导致溢出，因此工程中要取对数，这样就变成了连加，把这些概率

和似然存进一张表中，在预测时直接查表。

4. 极大似然估计

极大似然估计是贝叶斯参数估计背后的思想,在这里看着不明显,其实已经用上了.极大似然的意思是参数可能服从任意一种分布,我们应该选择其出现概率最大的那个,哪个出现概率最大呢?就是我们手里已经有的样本服从的概率分布.用手中现有的样本来估计后验概率,这背后就是极大似然估计.

5. 三种模型

多项式模型：在训练时，重复的特征也进行计算

伯努利模型：不计算重复的特征。一个特征出现多次的话，只计算一次

高斯模型(混合模型)：在训练时考虑重复， 在预测是不考虑重复。

6. 平滑

7. 补充

贝叶斯算一般不适合用来进行集成学习。因为集成学习的一个要求就是构造差异的基分类器而贝叶斯非常稳定，原因在于特征条件独立性假设，这个假设太强，大大简化了模型，所以导致了高偏差低方差，低方差的话泛化能力强，改变数据并不能带来什么不一样。

8. 工程trick

8.1 取对数，把相应的值存到一个表里（每个词的概率log('发票'|S)）

朴素贝叶斯是通过先验概率来修正正负样例数据偏斜，SVM是通过支持向量来修正（如果数据偏斜，多的那类的支持向量个数会更多，最终的损失也会不一样）

9. 优缺点

优点：

速度贼快

多分类不用做啥改进，照样适用

需要的样本量少

对于类别类的变量效果很好，对于连续值变量，默认服从正态分布

泛化能力强

缺点：

要平滑

输出的概率可以比较大小，但是物理意义不明显