机器学习笔记（11）朴素贝叶斯法（Naïve Bayes）

作者：玖五乾元

朴素贝叶斯分类算法是统计学的一种分类方法，它是一类利用概率统计知识进行分类的算法。在许多场合，朴素贝叶斯(Naïve Bayes，NB)分类算法可以与决策树和神经网络分类算法相媲美，该算法能运用到大型数据库中，而且方法简单、分类准确率高、速度快。

在前面学习过支持向量机，决策树等分类模型，其中决策树通过寻找最佳划分特征进而学习样本路径实现分类，支持向量机通过寻找分类超平面进而最大化类别间隔实现分类。相比之下，朴素贝叶斯独辟蹊径，通过考虑特征概率来预测分类。

(一)数学依据

为了便于理解朴素贝叶斯法，有必要对用到的概率统计的几个重要概念和公式进行介绍如下

（1）、先验概率

执因求果，事件发生前的预判概率。可以是基于历史数据的统计，可以由背景常识得出，也可以是人的主观观点给出。

（2）、后验概率

知果求因，事件发生后求的反向条件概率；或者说，基于先验概率求得的反向条件概率。概率形式与条件概率相同。

（3）、联合概率

在概率论中，联合概率表示两个事件共同发生的概率。A 与 B 的联合概率表示为 P(AB) 或者 P(A,B),或者 P（A∩B）。

（4）、条件概率公式

设 A,B 是两个事件，且 P(B)>0,则在事件 B 发生的条件下，事件 A 发生的条件概率（conditional probability)为：

$$\mathrm{P(A|B)} = \frac{P(AB)}{P(B)}$$

（5）、全概率公式

若事件 B1，B2，…构成一个完备事件组且都有正概率，则对任意一个事件 A，有如下公式成立：

$$P(A) = P(AB_1) + P(AB_2) + \cdots + P(AB_n)$$
$$= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$
$$= \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

（6）、贝叶斯公式

与全概率公式解决的问题相反，贝叶斯公式是建立在条件概率的基础上寻找事件发生的原因（即大事件 A 已经发生的条件下，分割中的小事件 Bi 的概率），设 B1,B2,...是样本空间 Ω 的一个划分，则对任一事件 A（P(A)>0),有

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

上式即为贝叶斯公式（Bayes formula)，Bi 常被视为导致试验结果 A 发生的“原因”，P(Bi)(i=1,2,...)表示各种原因发生的可能性大小，故称先验概率；P(Bi|A)(i=1,2...)则反映当试验产生了结果 A 之后，再对各种原因概率的新认识，故称后验概率。

(二)朴素贝叶斯模型

朴素贝叶斯法的目标是：针对上述训练数据集学习得到输入 X 和输出 Y 的联合分布概率，然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。

将上述概率统计学知识应用到机器学习的分类问题处理，假设有训练数据样本集是：

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

其中：

$$x_i = \left\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(k)}\right\}, i = 1,2, \ldots, N$$

$$y_i \in Y = \{c_1, c_2, \ldots, c_t\}$$

即有 N 个训练数据样本，每个样本有 k 个特征，特征输出有 t 个类别。

那么，朴素贝叶斯法的具体学习过程描述为：

第一步、对于上述训练数据样本集，结合上节描述的条件概率公式，可列出联合分布概率的推导公式如下:

$$P(X, Y = c_i) = P(X = x|Y = c_i)P(Y = c_i)$$

第二步、通过求解$P(Y = c_i)$ 和 $P(X = x|Y = c_i) = P(X^1 = x^1, \ldots X^k = x^k|Y = c_i)$，得出联合分布概率

$P(X, Y = c_i)$

第三步、用于分类，对新输入的数据（X,Y），我们如何判断它属于哪个类型？利用后验概率最大化来判断分类。我们只要计算出所有的 t 个后验条件概率,然后找出最大的条件概率对应的类别，这就是朴素贝叶斯的预测了。后验概率计算根据贝叶斯定理要得出

$$P(Y = c_i|X = x) = \frac{P(X, Y = c_i)}{\sum_{i=1}^{t} P(X = x|Y = c_i)P(Y = c_i)} = \frac{P(X = x|Y = c_i)P(Y = c_i)}{\sum_{i=1}^{t} P(X = x|Y = c_i)P(Y = c_i)}$$

对于上述联合分布概率，其中$P(Y = c_i)$比较好求解，理解一下就是在样训练样本集中对每个输入出类别c_i出现的概率。$P(X = x|Y = c_i) = P(X^1 = x^1, \dots X^k = x^k|Y = c_i)$，k 为输入 X 的特征值个数,可以采用极大似然估计，但实际上其估计量是指数级的，所以不可行。

为此，朴素贝叶斯法对条件概率分布$P(X^1 = x^1, \dots X^k = x^k|Y = c_i)$作了条件独立性的假设，可以理解为输入 X 的各特征值分布条件独立没有相互依赖和影响。正是由于这个较强的假设朴素贝叶斯由此得名。这一假设使朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率。

基于上条件独立假设， 可得出

$$P(X = x|Y = c_i) = P(X^1 = x^1, \dots X^k = x^k|Y = c_i) = \prod_{j=1}^{k} P(X^j = x^j|Y = c_i)$$

因此，利用贝叶斯定理通过计算后验概率进行分类的朴素贝叶斯法的模型为：

$$y = f(x) = \underset{i=1..t}{\operatorname{argmax}} P(Y = c_i|X = x) = \underset{i=1..t}{\operatorname{argmax}} \frac{P(X = x|Y = c_i)P(Y=c_i)}{\sum_{i=1}^{t} P(X=x|Y=c_i)P(Y=c_i)} = \underset{i=1..t}{\operatorname{argmax}} \frac{\prod_{j=1}^{k} P(X^j = x^j|Y = c_i)\, P(Y = c_i)}{\sum_{i=1}^{t}[\prod_{j=1}^{k} P(X^j = x^j|Y = c_i)\, P(Y = c_i)]}$$

因此分母相同，等价于：

$$\underset{i=1..t}{\operatorname{argmax}} \prod_{j=1}^{k} P(X^j = x^j|Y = c_i)\, P(Y = c_i)$$

注[1]：argmax 是一种函数，函数 y=f(x)，x0= argmax(f(x)) 的意思就是参数 x0 满足 f(x0)为 f(x)的最大值；换句话说就是 argmax(f(x))是使得 f(x)取得最大值所对应的变量 x。

(三)朴素贝叶斯法的学习策略

在朴素贝叶斯法中，学习意味着估计$P(Y = c_i)$ 和 $P(X = x|Y = c_i) = P(X^1 = x^1, \dots X^k = x^k|Y = c_i)$，可以应用极大似然估计法估计相应的概率。

极大似然估计，只是一种概率论在统计学的应用，它是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。极大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。极大似然估计只是一种粗略的数学期望。求极大似然函数估计值的一般步骤：

（1） 写出似然函数；

（2） 对似然函数取对数，并整理；

（3） 求导数；

（4） 解似然方程 。

第一部分：对先验概率$P(Y = c_i)$ 的估计：

$$P(Y = c_i) = \frac{\sum_j^N I(y_j = c_i)}{N}, i = 1,2, \dots k$$

可以理解为 N 个样本中每个样本出现的次数/样本总数

第二部分：对条件概率 $P(X = x|Y = c_i) = P(X^1 = x^1, \dots X^k = x^k|Y = c_i)$，的估计：

设第 j 个特征值x^j可能取值的集合为$\{a_{j1}, a_{j2}, \dots, a_{js}\}$，那么条件概率$P(X^j = a_{jl}|Y = c_k)$的极大似然估计是

$$P(X^j = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^N I(x_i^j = a_{jl}, y_i = c_k)}{\sum_{i=1}^N I(y_i = c_k)}$$
$$j = 1,2, \dots n; l = 1,2, \dots, S_j; k = 1,2, \dots, K$$

式中，x_i^j是第 i 个样本的第 j 个特征；a_{jl}是第 j 个特征可能取的第 l 个值；I 为指示函数。

注[2]：后验概率最大实际上等价于 0-1 损失函数的期望风险最小化

(四)朴素贝叶斯法的算法

朴素贝叶斯法的算法即是计算先验概率 $P(X = x|Y = c_i) = P(X^1 = x^1, \dots X^k = x^k|Y = c_i)$，的数学过程，属于单于的概率公式求解过程，没有复杂的求导和矩阵运算，也正是因为如朴素贝叶斯法的计算效率很高。

乾元小站