统计学习方法——朴素贝叶斯

create 2021/11/9 广州

引言

朴素贝叶斯多用于分类问题，实现简单，并且学习和预测的效率都很高。而贝叶斯作为朴素贝叶斯的基础这里做一个简单的介绍。

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

朴素贝叶斯

给定训练集$T = (x_1, y_1), (x_2, y_2), .., (x_N, y_N)$，设类别可选数目为K，即$c_1, c_2, \ldots, c_K$，特征维度为m，即$x_i = (x_i^1, x_i^2, \ldots x_i^m)$，第j维的特征可取值数目为$S_j$，分别为$a_j^1, a_j^2, \ldots, a_j^{S_j}$。

这里的描述比较抽象，我用一个简单的例子来表示：

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x = (2, S)^T$ 的类标记 y. 表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1 = \{1, 2, 3\}$，$A_2 = \{S, M, L\}$，Y 为类标记，$Y \in C = \{1, -1\}$.

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	−1	−1	1	1	−1	−1	−1	1	1	1	1	1	1	1	−1

这里是统计学习方法p50页的例题，这里我不解答，而是针对里面的内容来简单的表示上面的描述。

- **可选数目K**就是题目中的Y的类别数也就是2，$c_1, c_2, \ldots, c_K$可以具体表示为-1和1。
- **特征维度m**对应题目就是$x = (2, S)^T$，此处m=2有两个维度，对应了表格中的$X^{(1)}$和$X^{(2)}$。
- **特征可取值数目**S_j就是具体每个特征可能的取值，题目中$X^{(1)}$就三个取值1，2，3，因此S_1就为3。

$a_j^1, a_j^2, \ldots, a_j^{S_j}$对应了每个取值，此处$a_1^1 = 1, a_1^2 = 2, a_1^3 = 3$。对于$X^{(2)}$来说$a_1^1 = S, a_1^2 = M, a_1^3 = L$。

而我们的目标就是**知道x的数据，将x分到正确的类别中。**

在有了上面的描述后，我们可以得到以下的先验概率和条件概率：

先验概率为：

$$P(Y = c_k), k = 1, 2, \ldots, K$$

条件概率为：

$$P(X = x|Y = c_k) = P(X^1 = x^1, X^2 = x^2, \ldots, X^m = x^m|Y = c_k), k = 1, 2, \ldots, K$$

因此也就得到了联合概率：

$$p(X = x, Y = c_k) = P(Y = c_k)P(X = x|Y = c_k)$$

为了降低模型的复杂度，朴素贝叶斯作了条件独立性的假设：

$$\begin{aligned} P(X = x|Y = c_k) &= P(X^1 = x^1, X^2 = x^2, \ldots, X^m = x^m|Y = c_k) \\ &= \prod_{j=1}^{m} P(X^j = x^j|Y = c_k) \end{aligned}$$

因为这是一个强假设，朴素贝叶斯由此得名

对于后验概率$P(Y = c_k|X = x)$，由贝叶斯公式有：

$$\begin{aligned} P(Y = c_k|X = x) &= \frac{p(X = x, Y = c_k)}{P(X = x)} \\ &= \frac{P(Y = c_k)P(X = x|Y = c_k)}{P(X = x)} \\ &= \frac{P(Y = c_k)\prod_{j=1}^{m} P(X^j = x^j|Y = c_k)}{P(X = x)} \end{aligned}$$

而我们的目标就是选取使得$P(Y = c_k|X = x)$概率最大的c_k类别，因此分母$P(X = x)$并没有太多用处，不影响c_k的取值。

因为我的目标就变成了如下的式子：

$$y = \arg\max_{c_k} P(Y = c_k)\prod_{j=1}^{m} P(X^j = x^j|Y = c_k)$$

找到一个合适的c_k使$P(Y = c_k|X = x)$概率最大。

朴素贝叶斯的参数估计

极大似然估计

对于目标公式中的先验概率$P(Y = c_k)$：

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, \ldots, K$$

其中$I(y_i = c_k)$为信号函数，成立的时候返回1，不成立返回0

对与目标公式中的条件概率$P(X^j=x^j|Y=c_k)$，设第j 个特征x^j可能的取值集合为 $a_j^1,a_j^2,a_j^3 ...,a_j^{S_j}$，可以得到：

$$P(X^j=a_j^l|Y=c_k)=\frac{\sum_{i=1}^N I(x_i^j=a_j^l,y_i=c_k)}{\sum_{i=1}^N I(y_i=c_k)},l=1,2,\ldots,S_j$$

这个举个例子来简单说明下，假设有两个类别(1，2)，3个特征(a, b , c)，每个特征都有4个可能的取值，上面那个公式说明的就是在给定具体类别的前提下(1或者2)，每个特征(a\b\c)中每一个可能取值的概率(4个可能取值)，如$P(a=a_1|Y=1)$表示的就是在给定类别1的前提下，特征a的第一个可能取值的概率。因为我们这里用了信号函数，所以可以通过统计数据集直接得到概率。

在得到先验概率和条件概率后，对于给定的数据$x=(x_i^1,x_i^2,\ldots x_i^m)$就可以得到：

$$P(Y=c_k)\prod_{j=1}^m P(X^j=x^j|Y=c_k),k=1,2,3,\ldots.K$$

最后在找到使上面式子最大的c_k，就是最后的结果：

$$y=\arg\max_{c_k} P(Y=c_k)\prod_{j=1}^m P(X^j=x^j|Y=c_k)$$

配合例子肯定更好理解，例子在统计学习方法P50页有，也就是我上面提到的例子，这里附上完整版的。

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^T$ 的类标记 y. 表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，Y 为类标记，$Y\in C=\{1,-1\}$.

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	−1	−1	1	1	−1	−1	−1	1	1	1	1	1	1	1	−1

解 根据算法 4.1，由表 4.1，容易计算下列概率：

$P(Y=1)=\frac{9}{15}$， $P(Y=-1)=\frac{6}{15}$

$P(X^{(1)}=1|Y=1)=\frac{2}{9}$， $P(X^{(1)}=2|Y=1)=\frac{3}{9}$， $P(X^{(1)}=3|Y=1)=\frac{4}{9}$

$P(X^{(2)}=S|Y=1)=\frac{1}{9}$， $P(X^{(2)}=M|Y=1)=\frac{4}{9}$， $P(X^{(2)}=L|Y=1)=\frac{4}{9}$

$P(X^{(1)}=1|Y=-1)=\frac{3}{6}$， $P(X^{(1)}=2|Y=-1)=\frac{2}{6}$， $P(X^{(1)}=3|Y=-1)=\frac{1}{6}$

$P(X^{(2)}=S|Y=-1)=\frac{3}{6}$， $P(X^{(2)}=M|Y=-1)=\frac{2}{6}$，. $P(X^{(2)}=L|Y=-1)=\frac{1}{6}$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$.