朴素贝叶斯法的学习和分类

2024年4月23日　　15:13

朴素贝叶斯假设：

假设特征x=（x1，x2，x3，x4.... xn）,相对于标签值y条件独立

对于特征$X=(x_1,x_2,\dots,x_n)$，**满足** $x_i \perp x_j \mid y \ (i \neq j)$

$$p(X\,|y)=p(x_1,x_2,\dots,x_n|y)=\prod_{j=1}^{n} p(x_j|y)$$

朴素贝叶斯法：

根据贝叶斯法则，先验概率算后验概率。

$$p(y|x)=\frac{p(x,y)}{p(x)}=\frac{p(x|y)p(y)}{p(x)}=\frac{p(x|y)p(y)}{\sum_i p(x|y_i)p(y_i)}\propto p(x|y)p(y)$$

目标是：根据最大似然法，让得出来的最后分法是所有分法中概率最大的那个

$$y=\arg\max_y p(y|X)=\arg\max_y\frac{p(X,\ y)}{p(X)}=\arg\max_y p(X|y)p(y)$$
$$=\arg\max_y \prod\nolimits_i p(x_i|y)\,p(y)$$

例子：

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 $x=(2,S)^{\mathrm{T}}$ 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

	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

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

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

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

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

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

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 $y=-1$。　■

贝叶斯估计：

测试样例的特征没有在训练集中出现，即训练集的数据不完整，有特征值缺失，比

如头发特征有黑头发，白头发，红头发。但是训练集里面只有黑头发和白头发，会影响后验概率

解决：加个调节因子κ（xk）xk特征有多少种取值和 λ （取值1，拉普拉斯平滑）

思考：在前面的分类算法中，如果测试样例中的特征没有在训练集中出现会造成什么结果？

会影响到后验概率的计算结果，使分类产生偏差。解决这一问题的方法是采用**贝叶斯估计**。具体地，估计特征x_k的条件概率为：

$$p(x_k|y_i)=\frac{C(x_k,y_i)+\lambda}{C(y_i)+K(x_k)\lambda}$$

估计y_i的概率计算为：

$$p(y_i)=\frac{C(y_i)+\lambda}{N+K(y_i)\lambda}$$

式中C表示符合条件的样本个数，$K(x)$为特征x的取值种类数，$\lambda\geq 0$。等价于在随机变量各个取值的频数上赋予一个正数$\lambda\geq 0$。当$\lambda=0$时就是极大似然估计。一般取$\lambda=1$，这时称为**拉普拉斯平滑** (Laplacian smoothing)。

例子：

	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

$$P(Y=1)=\frac{10}{17},\quad P(Y=-1)=\frac{7}{17}$$

$$P(X^{(1)}=1|Y=1)=\frac{3}{12},\quad P(X^{(1)}=2|Y=1)=\frac{4}{12},\quad P(X^{(1)}=3|Y=1)=\frac{5}{12}$$

$$P(X^{(2)}=S|Y=1)=\frac{2}{12},\quad P(X^{(2)}=M|Y=1)=\frac{5}{12},\quad P(X^{(2)}=L|Y=1)=\frac{5}{12}$$

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

$$P(X^{(2)}=S|Y=-1)=\frac{4}{9},\quad P(X^{(2)}=M|Y=-1)=\frac{3}{9},\quad P(X^{(2)}=L|Y=-1)=\frac{2}{9}$$

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12}=\frac{5}{153}=0.0327$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9}=\frac{28}{459}=0.0610$$

由于 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$。 ■

KNN算法：

想法：取离q最近的k个样本的标签众数为q的标签

例子：

- **K-近邻（KNN）算法—KNN处理分类问题：步骤**

Document number	I	buy	an	apple	...	friend	has	emotion
train 1	1	1	1	1	...	0	0	happy
train 2	1	0	0	1	...	0	0	happy
train 3	0	0	0	1	...	0	0	sadness
test 1	0	0	1	1	...	1	1	?

2. 相似度计算：计算test1与每个train的距离

欧氏距离：

$$d(train1,test1)=\sqrt{(1-0)^2+(1-0)^2+\cdots+(0-1)^2}=\sqrt{6};$$

$$d(train2,test1)=\sqrt{(1-0)^2+(1-0)^2+\cdots+(0-1)^2}=\sqrt{8};$$

$$d(train3,test1)=\sqrt{(0-0)^2+(0-0)^2+\cdots+(0-1)^2}=\sqrt{9};$$

（也可以使用其他距离度量方式）

3. 类别计算：最相似的k个样本之标签的众数

若k=1，test1的标签即为train1的标签happy；

若k=3，test1的标签为train1,train2,train3的标签中数量较多的，即为happy。

参数设置：

通过验证集对参数k进行调优：

- **通过验证集对参数（k值）进行调优**
  - **如果k值取的过大，学习的参考样本更多，会引入更多的噪音，所以可能存在欠拟合的情况；**
  - **如果k值取的过小，参考样本少，容易出现过拟合的情况**
  - **关于k的经验公式：一般取k= $\sqrt{N}$，N为训练集实例个数，大家可以尝试一下**

权重归一化：

权重归一化

Name	Formula	Explain
Standard score	$X' = \frac{X-\mu}{\sigma}$	μ is the mean and σ is the standard deviation
Feature scaling	$X' = \frac{X-X_{min}}{X_{max}-X_{min}}$	X_{min} is the min value and X_{max} is the max value

不同权重的距离度量公式：

曼哈顿距离和欧式距离：

距离公式：

$\mathrm{L_p}$**距离（所有距离的总公式）：**

$$\mathrm{L_p}(x_i, x_j) = \left\{\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right\}^{\frac{1}{p}}$$

$p = 1$：**曼哈顿距离；**

$p = 2$：**欧氏距离，最常见。**

例 3.1 已知二维空间的 3 个点 $x_1 = (1,1)^{\mathrm{T}}$，$x_2 = (5,1)^{\mathrm{T}}$，$x_3 = (4,4)^{\mathrm{T}}$，试求在 p 取不同值时，L_p 距离下 x_1 的最近邻点。

解 因为 x_1 和 x_2 只有第一维的值不同，所以 p 为任何值时，$L_p(x_1, x_2) = 4$，而

$$L_1(x_1,x_3) = 6,\quad L_2(x_1,x_3) = 4.24,\quad L_3(x_1,x_3) = 3.78,\quad L_4(x_1,x_3) = 3.57$$

于是得到：p 等于 1 或 2 时，x_2 是 x_1 的最近邻点；p 大于等于 3 时，x_3 是 x_1 的最近邻点。■

余弦相似度：

于是得到：p 等于 1 或 2 时，x_2 是 x_1 的最近邻点；p 大于等于 3 时，x_3 是 x_1
邻点。

余弦相似度：

$\cos\left(\underset{A}{\rightarrow}, \underset{B}{\rightarrow}\right) = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}||\vec{B}|}$，**其中**$\underset{A}{\rightarrow}$和$\underset{B}{\rightarrow}$表示**两个文本特征向量；**

余弦值作为衡量两个个体间差异的大小的度量

为正且值越大，表示两个文本差距越小，为负代表差距越大，请大家自行脑补两个向量余弦值

6.1 K-近邻（KNN）算法—KNN算法效率

假设训练集有N个样本，测试集有M个样本，每个样本是一个V维的向量。

如果使用线性搜索的话，那么*k*-NN的时间花销就是O(N*M*V)。

改善：KD树(感兴趣可自行尝试)

#将所有数据的后验概率算出，分两个list【dict】 第一个list是y为2是特征值的情况，dict的key为特征值，count为计数， sum局部计数，y也是找个sum计数

#还要算Y的概率

#乘起来，找概率最大的

朴素贝叶斯分类完成