多项式朴素贝叶斯算法案例--平滑下的计算方式

L先生AI课堂

回顾多项式朴素贝叶斯

- Multinomial Naive Bayes是指当特征属性服从多项分布(**特征是离散的形式的时候**)，直接计算类别数目的占比作为先验概率和条件概率。

$$p(y_k)=\frac{N_{y_k}}{N} \qquad\qquad p(x_i|y_k)=\frac{N_{y_k,x_i}}{N_{y_k}}$$

- N是总样本个数，k是总的类别个数，N_{yk}是类别为yk的样本个数。
- N_{yk}是类别为yk的样本个数，$N_{yk,xi}$为类别yk中第i维特征的值为xi的样本个数，

多项式朴素贝叶斯案例理解

- 对于下列训练数据，使用多项式朴素贝叶斯方式对测试样本(2,M,L)做一个预测判断。

	1	2	3	4	5	6	7	8	9	10
x1	1	1	1	2	2	2	2	3	3	4
x2	S	M	S	L	S	S	L	L	L	S
x3	L	H	L	H	L	M	H	M	H	M
y	-1	1	1	-1	-1	-1	1	1	1	1

	x1=1	x1=2	x1=3	x1=4	
y=1	2	1	2	1	6
y=-1	1	3	0	0	4
	3	4	2	1	10

	x2=S	x2=M	x2=L	
y=1	2	1	3	6
y=-1	3	0	1	4
	5	1	4	10

	x3=L	x3=M	x3=H	
y=1	1	2	3	6
y=-1	2	1	1	4
	3	3	4	10

训练阶段:

- 先验概率:

$$p(y=1)=6/10=0.6 \qquad p(y=-1)=4/10=0.4$$

- 条件概率:

$$p(x_1=1|y=1)=\frac{2}{6} \quad p(x_1=1|y=-1)=\frac{1}{4}$$

$$p(x_1=2|y=1)=\frac{1}{6} \quad p(x_1=2|y=-1)=\frac{3}{4}$$

$$p(x_1=3|y=1)=\frac{2}{6} \quad p(x_1=3|y=-1)=0$$

$$p(x_1=4|y=1)=\frac{1}{6} \quad p(x_1=4|y=-1)=0$$

$$p(x_2=S|y=1)=\frac{2}{6} \quad p(x_2=S|y=-1)=\frac{3}{4}$$

$$p(x_2=M|y=1)=\frac{1}{6} \quad p(x_2=M|y=-1)=0$$

$$p(x_2=L|y=1)=\frac{3}{6} \quad p(x_2=L|y=-1)=\frac{1}{4}$$

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{m} P(x_i \mid y)$$

预测阶段:

样本(2,M,L)的预测概率:

$$p(y=1|x) \propto p(y=1)p(x_1=2|y=1)p(x_2=M|y=1)p(x_3=L|y=1) = \frac{6}{10} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} = \frac{1}{360}$$

$$p(y=-1|x) \propto p(y=-1)p(x_1=2|y=-1)p(x_2=M|y=-1)p(x_3=L|y=-1) = \frac{4}{10} * \frac{3}{4} * 0 * \frac{2}{4} = 0$$

$$\hat{y} = \arg\max_{y} \{p(y=1|x)\ p(y=-1|x)\} = 1$$

多项式朴素贝叶斯—平滑

- Multinomial Naive Bayes是指当特征属性服从多项分布(特征是离散的形式的时候)，直接计算类别数目的占比作为先验概率和条件概率。

$$p(y_k)=\frac{N_{y_k}+\alpha}{N+k*\alpha} \qquad p(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n_{x_i}*\alpha}$$

- N是总样本个数，k是总的类别个数，N_{yk}是类别为yk的样本个数，α为平滑值。
- N_{yk}是类别为yk的样本个数，nx_i为特征属性xi的不同取值数目，$N_{yk,xi}$为类别yk中第i维特征的值为xi的样本个数，α为平滑值。
- 当α=1时，称为Laplace平滑，当0<α<1时，称为Lidstone平滑，α=0时不做平滑；平滑的主要作用是可以克服条件概率为0的问题。

多项式朴素贝叶斯案例理解

- 对于下列训练数据，使用多项式朴素贝叶斯方式对测试样本(2,M,L)做一个预测判断。

	1	2	3	4	5	6	7	8	9	10
x1	1	1	1	2	2	2	2	3	3	4
x2	S	M	S	L	S	S	L	L	L	S
x3	L	H	L	H	L	M	H	M	H	M
y	-1	1	1	-1	-1	-1	1	1	1	1

	x1=1	x1=2	x1=3	x1=4	
y=1	2	1	2	1	6
y=-1	1	3	0	0	4
	3	4	2	1	10

	x2=S	x2=M	x2=L	
y=1	2	1	3	6
y=-1	3	0	1	4
	5	1	4	10

	x3=L	x3=M	x3=H	
y=1	1	2	3	6
y=-1	2	1	1	4
	3	3	4	10

$$N = 10$$

$$k = 2 \quad n_1 = 4$$

$$n_2 = 3 \quad n_3 = 3$$

训练阶段: $\alpha = 1$ $\quad p(y_k) = \dfrac{N_{y_k} + \alpha}{N + k * \alpha}$ $\quad p(x_i | y_k) = \dfrac{N_{y_k, x_i} + \alpha}{N_{y_k} + n_i * \alpha}$

- 先验概率：

$p(y=1) = (6+1)/(10+2*1) = 7/12 \quad p(y=-1) = 5/12$

- 条件概率：

$$p(x_1 = 1 | y = 1) = \frac{2+1}{6+4*1} = \frac{3}{10} \quad p(x_1 = 1 | y = -1) = \frac{2}{8} \quad p(x_2 = S | y = 1) = \frac{3}{9} \quad p(x_2 = S | y = -1) = \frac{4}{7}$$

$$p(x_1 = 2 | y = 1) = \frac{2}{10} \quad p(x_1 = 2 | y = -1) = \frac{4}{8} \quad p(x_2 = M | y = 1) = \frac{2}{9} \quad p(x_2 = M | y = -1) = \frac{1}{7}$$

$$p(x_1 = 3 | y = 1) = \frac{3}{10} \quad p(x_1 = 3 | y = -1) = \frac{1}{8} \quad p(x_2 = L | y = 1) = \frac{4}{9} \quad p(x_2 = L | y = -1) = \frac{2}{7}$$

$$p(x_1 = 4 | y = 1) = \frac{2}{10} \quad p(x_1 = 4 | y = -1) = \frac{1}{8}$$

训练阶段:　$\alpha = 1$

- 条件概率:

$$p(x_3 = L \mid y = 1) = \frac{2}{9} \qquad p(x_3 = L \mid y = -1) = \frac{3}{7}$$

$$p(x_3 = M \mid y = 1) = \frac{3}{9} \qquad p(x_3 = M \mid y = -1) = \frac{2}{7}$$

$$p(x_3 = H \mid y = 1) = \frac{4}{9} \qquad p(x_3 = H \mid y = -1) = \frac{2}{7}$$

预测阶段:

样本(2,M,L)的预测概率概率: $\alpha = 1$

$$p(y=1|x) \propto p(y=1)p(x_1=2|y=1)p(x_2=M|y=1)p(x_3=L|y=1) = \frac{7}{12}*\frac{2}{10}*\frac{2}{9}*\frac{2}{9} = \frac{7}{1215}$$

$$p(y=-1|x) \propto p(y=-1)p(x_1=2|y=-1)p(x_2=M|y=-1)p(x_3=L|y=-1) = \frac{5}{12}*\frac{4}{8}*\frac{1}{7}*\frac{3}{7} = \frac{5}{392}$$

$$\hat{y} = \arg\max_{y}\{p(y=1|x), p(y=-1|x)\} = -1$$