朴素贝叶斯分类器 (Naive Bayes Classifier)

朴素贝叶斯算法

判断股票涨跌

财务质量	市值	周期性	性质	表现
差	大市值	周期	国有	上涨
中	小市值	非周期	民营	上涨
中	大市值	非周期	国有	上涨
好	大市值	周期	国有	上涨
好	小市值	周期	国有	下跌
差	中市值	周期	国有	下跌
差	大市值	非周期	民营	上涨
好	小市值	周期	民营	下跌
差	中市值	非周期	民营	上涨
中	中市值	周期	国有	下跌
中	中市值	非周期	民营	上涨
好	小市值	非周期	民营	下跌
中	大市值	非周期	国有	上涨
差	小市值	周期	民营	下跌

财务质量	市值	周期性	性质	表现
中	中市值	非周期	国营	?

贝叶斯公式

$$P(y=1|X)=\frac{P(X|y=1)P(y=1)}{P(X)}$$

$$P(y=0|X)=\frac{P(X|y=0)P(y=0)}{P(X)}$$

比较大小

估计概率

$$P(y=1)=\frac{N(y=1)}{N}$$

$P(X \mid y=1)$ 一般难以估计

$$\prod P(x_i \mid y=1)$$

$$P(x_i \mid y=1)=\frac{N(y=1,x_i)}{N(y=1)}$$

财务质量	市值	周期性	性质	表现
差	大市值	周期	国有	上涨
中	小市值	非周期	民营	上涨
中	大市值	非周期	国有	上涨
好	大市值	周期	国有	上涨
好	小市值	周期	国有	下跌
差	中市值	周期	国有	下跌
差	大市值	非周期	民营	上涨
好	小市值	周期	民营	下跌
差	中市值	非周期	民营	上涨
中	中市值	周期	国有	下跌
中	中市值	非周期	民营	上涨
好	小市值	非周期	民营	下跌
中	大市值	非周期	国有	上涨
差	小市值	周期	民营	下跌

财务质量	市值	周期性	性质	表现
中	中市值	非周期	国有	?

N = 14

N(y=1) = 8

P(y=1) = 8/14 = 0.5714

N(y=1, 财务质量 = 中) = 4
N(y=1, 市值 = 中市值) = 2
N(y=1, 周期性 = 非周期) = 6
N(y=1, 性质 = 国有) = 4

P(X | y=1)
= (4/8)(2/8)(6/8)(4/8)
= 0.0469

P(X | y=1) * P (y=1)
= 0.0268

N(y=0) = 6

P(y=0) = 6/14 = 0.4286

N(y=0, 财务质量 = 中) = 1
N(y=0, 市值 = 中市值) = 2
N(y=0, 周期性 = 非周期) = 1
N(y=0, 性质 = 国有) = 3

P(X | y=0)
= (1/6)(2/6)(1/6)(3/6)
= 0.0046

P(X | y=0) * P (y=0)
= 0.0020

>

上涨

Laplace Smoothing

零概率问题

财务质量	市值	周期性	性质	表现
差	大市值	周期	国有	上涨
中	小市值	非周期	民营	上涨
中	大市值	非周期	国有	上涨
好	大市值	周期	国有	上涨
好	小市值	周期	国有	下跌
差	中市值	周期	国有	下跌
差	大市值	非周期	民营	上涨
好	小市值	周期	民营	下跌
差	中市值	非周期	民营	上涨
中	中市值	周期	国有	下跌
中	中市值	非周期	民营	上涨
好	小市值	非周期	民营	下跌
中	大市值	非周期	国有	上涨
差	小市值	周期	民营	下跌

财务质量	市值	周期性	性质	表现
中	中市值	非周期	国营	上涨
好	大市值	周期	国营	?

N=14

N(y=0) = 6

P(y=0) = 6/14 = 0.4286

N(y=0, 财务质量 = 好) = 3
N(y=0, 市值 = 大市值) = 0
N(y=0, 周期性 = 周期) = 5
N(y=0, 性质 = 国有) = 3

P(X | y=0)
= (3/6)(0/6)(5/6)(3/6)
= 0

P(X | y=0) * P (y=0)
= 0
= P(y=0 | X)

Laplace Smoothing

$$P(x_i \mid y) = \frac{N(y, x_i) + 1}{N(y) + possible\_values(x_i)}$$

第 i 个特征的可能取值个数

- 假定特征的取值服从均匀分布
- 1 / possible_values(xi) 是一个先验概率
- Laplace smoothing：将观察到的频率与先验概率**“勾兑”**，得到对实际概率的估计
- 当样本量足够大，先验概率的“勾兑”影响比较小，估计的概率趋向于实际概率

使用 Laplace Smoothing

财务质量	市值	周期性	性质	表现
差	大市值	周期	国有	上涨
中	小市值	非周期	民营	上涨
中	大市值	非周期	国有	上涨
好	大市值	周期	国有	上涨
好	小市值	周期	国有	下跌
差	中市值	周期	国有	下跌
差	大市值	非周期	民营	上涨
好	小市值	周期	民营	下跌
差	中市值	非周期	民营	上涨
中	中市值	周期	国有	下跌
中	中市值	非周期	民营	上涨
好	小市值	非周期	民营	下跌
中	大市值	非周期	国有	上涨
差	小市值	周期	民营	下跌

财务质量	市值	周期性	性质	表现
中	中市值	非周期	国营	上涨
好	大市值	周期	国营	?

N=14

N(y=0) = 6

P(y=0) = 6/14 = 0.4286

N(y=0, 财务质量 = 好) = 3
N(y=0, 市值 = 大市值) = 0
N(y=0, 周期性 = 周期) = 5
N(y=0, 性质 = 国有) = 3

P(X | y=0)
= [(3+1) / (6+3)] * [(0+1) / (6+3)] *
[(5+1) / (6+2)] * [(3+1) / (6+2))
= 0.0185

P(X | y=0) * P (y=0)
= 0.0079

√

Smoothing

$$P(x_i \mid y) = \frac{N(y, x_i) + 1}{N(y) + \alpha \times possible\_values(x_i)} \qquad (\alpha \geq 0)$$

$\alpha = 0$ No smoothing

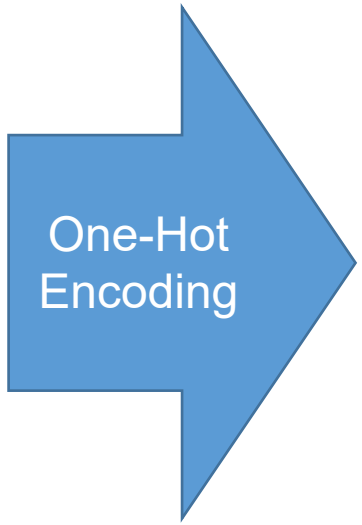
$\alpha = 1$ Laplace smoothing

$\alpha < 1$ Lidstone smoothing

BernoulliNB

BernoulliNB

财务质量	市值	周期性	性质
差	大市值	周期	国有
中	小市值	非周期	民营
中	大市值	非周期	国有
好	大市值	周期	国有
好	小市值	周期	国有
差	中市值	周期	国有
差	大市值	非周期	民营
好	小市值	周期	民营
差	中市值	非周期	民营
中	中市值	周期	国有
中	中市值	非周期	民营
好	小市值	非周期	民营
中	大市值	非周期	国有
差	小市值	周期	民营



财务质量_中	财务质量_好	财务质量_差	市值_中市值	市值_大市值	市值_小市值	周期性_周期	周期性_非周期	性质_国有	性质_民营
0	0	1	0	1	0	1	0	1	0
1	0	0	0	0	1	0	1	0	1
1	0	0	0	1	0	0	1	1	0
0	1	0	0	1	0	1	0	1	0
0	1	0	0	0	1	1	0	1	0
0	0	1	1	0	0	1	0	1	0
0	0	1	0	1	0	0	1	0	1
0	1	0	0	0	1	1	0	0	1
0	0	1	1	0	0	0	1	0	1
1	0	0	1	0	0	1	0	1	0
1	0	0	1	0	0	0	1	0	1
0	1	0	0	0	1	0	1	0	1
1	0	0	0	1	0	0	1	1	0
0	0	1	0	0	1	1	0	0	1

model = BernoulliNB().fit(X,y)

默认 alpha = 1.0， 即 Laplace smoothing

处理连续数值型的特征

判断股票涨跌

净资产收益率	市值	表现
-10.1	100	上涨
6.4	10	上涨
5.2	98	上涨
15	110	上涨
20	9	下跌
0	45	下跌
1.1	83	上涨
13.9	14	下跌
2.3	56	上涨
5	43	下跌
7.4	55	上涨
13	12	下跌
3.2	87	上涨
-2	10	下跌

ROE	市值	表现
30	100	?

$$P(X \mid y=1)$$

$$= P(ROE=30 \mid y=1) \times P(MV=100 \mid y=1)$$

$$= p_{ROE}(30 \mid y=1) \times p_{MV}(100 \mid y=1)$$

假设服从正态分布

计算概率

$$y = 1$$

$$ROE \sim N\left(3.8125, 7.0734^2\right)$$

$$p_{ROE}(30 \mid y = 1) = \frac{1}{\sqrt{2\pi}\,7.0734} \exp\left(-\frac{(30 - 3.8125)^2}{2 \times 7.0734^2}\right) = 5.9\text{e-}05$$

$$MV \sim N(74.8750, 32.8913^2)$$

$$p_{MV}(100 \mid y = 1) = \frac{1}{\sqrt{2\pi}\,32.8913} \exp\left(-\frac{(100 - 74.8750)^2}{2 \times 32.8913^2}\right) = 0.0091$$

GaussianNB

model = GaussianNB().fit(X,y)

GaussianBN 没有 Laplace Smoothing 参数

Hybrid Data

处理 Hybrid Data

sklearn 没有直接处理 hybird data 的方法

对于 hybrid data，考虑一下两种处理方法：

1. 将数值型数据进行分组，转换为标签行数据

2. 将特征按照类型分成两组 $X^{numeric}$ X^{label}
分别进行建模，并使用 model.predict_proba 得到属于正类的概率 $p^{numeric}$ p^{label}
使用以下公式计算对于合并数据下正类的概率：

$$p^{all} = \frac{1}{1 + \frac{p(y=1)}{p(y=0)}\left(1 - \frac{1}{p^{numeric}}\right)\left(1 - \frac{1}{p^{label}}\right)}$$

朴素贝叶斯算法总结

优点	缺点
算法简单 计算速度快，适用于大数据 可用于多分类问题 在满足独立性假设的前提下，效果出众 在文本分析领域有广泛的应用	独立性的假设不满足时，计算出来的概率不可靠（分类依然相对可靠）