朴素贝叶斯

李 波

朴素贝叶斯分类算法（Naive Bayesian Classifier）

- 在有监督学习任务中，样本输入与样本输出存在一个关系

$$y = f(\boldsymbol{x})$$

- 有监督学习算法从众多（可能无数多个）假设中（hypothesis）找出一个假设，逼近真实关系$y = f(\boldsymbol{x})$.
- 从统计学角度讲，一个很好的逼近关系为

$$p(y|\boldsymbol{x})$$

- 在逻辑回归中，

$$p(y|\boldsymbol{x}) = \begin{cases} \sigma(\boldsymbol{\omega}^T\boldsymbol{x} + b) & \text{如果 } y = 1 \\ 1 - \sigma(\boldsymbol{\omega}^T\boldsymbol{x} + b) & \text{如果 } y = 0 \end{cases}$$

- 可否直接估计或计算$p(y|\boldsymbol{x})$?

朴素贝叶斯分类算法（Naive Bayesian Classifier）

贝叶斯公式

$$p(y|\boldsymbol{x}) = \frac{p(y,\boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$$

朴素贝叶斯分类算法（Naive Bayesian Classifier）

贝叶斯公式

$$p(y|\boldsymbol{x}) = \frac{p(y,\boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$$

类别y出现的概率

在类别y给定条件下，样本特征$\boldsymbol{x}$的概率

给定一个样本输入特征，样本类别为y的概率

样本特征$\boldsymbol{x}$的概率

朴素贝叶斯分类算法（Naive Bayesian Classifier）

贝叶斯公式

先验概率
类别y出现的概率

$$p(y|\boldsymbol{x}) \propto p(y)\, p(\boldsymbol{x}|y)$$

后验概率
给定一个样本输入特征，样本类别为y的概率

似然概率
在类别y给定条件下，样本特征$\boldsymbol{x}$的概率

- **先验概率** (prior)：类别y出现的概率.
- **似然概率** (likelihood)：在类别y给定条件下，样本特征$\boldsymbol{x}$的概率.
- **后验概率** (posterior)：给定一个样本输入特征，样本类别为y的概率.

朴素贝叶斯分类算法（Naive Bayesian Classifier）

估计先验概率和似然概率

- 训练数据为 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ ，其中 $\boldsymbol{x}_i \in \boldsymbol{R}^{m\times 1}$, $y_i \in \{0,1\}$.
- 标签为0的训练样本有n_0个，标签为1的训练样本有n_1个.
- 先验概率的估计为

$$p(y=0)=\frac{n_0}{n} \qquad p(y=1)=\frac{n_1}{n}$$

- 如果特征都是连续变量
  - ✓ $f(\boldsymbol{x}|y)$是什么分布?

朴素贝叶斯分类算法（Naive Bayesian Classifier）

估计先验概率和似然概率

- 训练数据为 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ ，其中 $\boldsymbol{x}_i \in \boldsymbol{R}^{m\times 1}$, $y_i \in \{0,1\}$.
- 标签为1的训练样本有n_0个，标签为0的训练样本有n_1个.
- 先验概率的估计为

$$p(y=0)=\frac{n_0}{n} \qquad p(y=1)=\frac{n_1}{n}$$

- 如果特征都是连续变量
  - ✓ $f(\boldsymbol{x}|y)$是什么分布?

    多维高斯分布.
  - ✓ 高斯分布的参数都是什么？有多少个?

朴素贝叶斯分类算法（Naive Bayesian Classifier）

估计先验概率和似然概率

- 训练数据为 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ ，其中 $\boldsymbol{x}_i \in \boldsymbol{R}^{m\times 1}$, $y_i \in \{0,1\}$.
- 标签为1的训练样本有n_0个，标签为0的训练样本有n_1个.
- 先验概率的估计为

$$p(\mathrm{y}=0)=\frac{n_0}{n} \qquad p(y=1)=\frac{n_1}{n}$$

- 如果特征都是连续变量
  - ✓ $f(\boldsymbol{x}|y)$是什么分布?

    多维高斯分布.
  - ✓ $f(\boldsymbol{x}|y)$是高斯分布，参数都是什么？有多少个?

    高斯分布有两个参数，均值和协方差矩阵。均值有m个参数；协方差矩阵有$\frac{m^2+m}{2}$个参数.

朴素贝叶斯分类算法（Naive Bayesian Classifier）

估计先验概率和似然概率

- 训练数据为 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ ，其中 $\boldsymbol{x}_i \in \boldsymbol{R}^{m\times 1}$, $y_i \in \{0,1\}$.
- 标签为1的训练样本有n_0个，标签为0的训练样本有n_1个.
- 先验概率的估计为

$$p(y=0)=\frac{n_0}{n} \qquad p(y=1)=\frac{n_1}{n}$$

- 如果特征都是离散变量
  - ✓ 计算$p(\boldsymbol{x}|y)$需要每个类别条件下m个特征的联合概率，会出现**特征组合爆炸情况**.

朴素贝叶斯分类算法（Naive Bayesian Classifier）

特征组合爆炸

- 以银行贷款为例，特征为 **性别** (男、女), **收入**(中、低、高), **教育程度**(高中,本科,研究生).
- 类别为通过贷款(y=1)，未通过贷款(y=0).
- 共计有36种特征组合，需要估计所有特征可能组合的概率

y = 1

男、低、高中	男、中、高中	男、高、高中	女、低、高中	女、中、高中	女、高、高中
男、低、本科	男、中、本科	男、高、本科	女、低、本科	女、中、本科	女、高、本科
男、低、研	男、中、研	男、高、研	女、低、研	女、中、研	女、高、研

y = 0

男、低、高中	男、中、高中	男、高、高中	女、低、高中	女、中、高中	女、高、高中
男、低、本科	男、中、本科	男、高、本科	女、低、本科	女、中、本科	女、高、本科
男、低、研	男、中、研	男、高、研	女、低、研	女、中、研	女、高、研

朴素贝叶斯分类算法（Naive Bayesian Classifier）

为解决特征组合爆炸问题，假设标签给定条件下m个特征相互独立

$$p(\boldsymbol{x}|y) = p(x_1, x_2, \cdots, x_m|y) = p(x_1|y)p(x_2|y)\cdots p(x_m|y) = \prod_{j=1}^{m} p(x_j|y)$$

y = 1

男性	女性	低收入	中等收入	高收入	高中	本科	研究生

y = 0

男性	女性	低收入	中等收入	高收入	高中	本科	研究生

需要估计或计算概率的个数，从36个降低到16个.

朴素贝叶斯分类算法（Naive Bayesian Classifier）

对于**连续**数值特征，同样假设在标签给定条件下m个特征相互独立

$$f(\boldsymbol{x}|y) = f(x_1, x_2, \cdots, x_m|y) = f(x_1|y)f(x_2|y)\cdots f(x_m|y) = \prod_{j=1}^{m} f(x_j|y)$$

- 仍然假设在类别给定条件下，每个特征高斯分布，具有两个参数，即均值和方差。
- 对一个类别，m个特征共有m个均值和m个方差，共计$2m$个参数。参数数量从$\frac{m^2+m}{2}+m$降低到$2m$。

朴素贝叶斯分类算法（Naive Bayesian Classifier）

问题：如果某个特定特征值在训练数据中出现次数比较少怎么办？比如在银行贷款例子中，如果具有教育程度为“高中”的样本很少，那么估计$p(\text{高中}|Y=1)$和$p(\text{高中}|Y=0)$变得不准确。

男性	女性	低收入	中等收入	高收入	高中	本科	研究生
15	5	8	11	1	9	6	5

解决方法: 拉普拉斯平滑（Laplace smoothing）。假设每个特征都人为增加k

$$p(A_{ji}|Y=0) \approx \frac{n_{0ji}}{n_0} \quad \Rightarrow \quad p(A_{ji}|Y=0) \approx \frac{n_{0ji}+k}{n_0+mk}$$

，其中A_{ji}表示第j个特征的第i个取值。

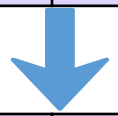
朴素贝叶斯分类算法（Naive Bayesian Classifier）

拉普拉斯平滑（Laplace smoothing）。假设每个特征都人为增加k

$$p(A_{ji}|Y=0) \approx \frac{n_{0ji}}{n_0} \quad \rightarrow \quad p(A_{ji}|Y=0) \approx \frac{n_{0ji}+k}{n_0+mk}$$

比如，贷款例子中，类别为0中，

男性	女性	低收入	中收入	高收入	高中	本科	研究生
15	5	8	11	1	9	6	5



男性	女性	低收入	中收入	高收入	高中	本科	研究生
15 + k	5 + k	8 + k	11 + k	1 + k	9 + k	6 + k	5 + k

$$p(\text{高收入}|Y=0) \approx \frac{1}{20} \quad \rightarrow \quad p(\text{高收入}|Y=0) \approx \frac{1+k}{20+8k}$$

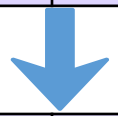
朴素贝叶斯分类算法（Naive Bayesian Classifier）

拉普拉斯平滑（Laplace smoothing）。假设每个特征都人为增加k

$$p(A_{ji}|Y=0) \approx \frac{n_{0ji}}{n_0} \quad \Rightarrow \quad p(A_{ji}|Y=0) \approx \frac{n_{0ji}+k}{n_0+mk}$$

比如，贷款例子中，类别为0中，

男性	女性	低收入	中收入	高收入	高中	本科	研究生
15	5	8	11	1	9	6	5



男性	女性	低收入	中收入	高收入	高中	本科	研究生
15 + k	5 + k	8 + k	11 + k	1 + k	9 + k	6 + k	5 + k

$$p(\text{高收入}|Y=0) \approx \frac{1}{20} \quad \Rightarrow \quad p(\text{高收入}|Y=0) \approx \frac{1+k}{20+8k}$$

如果k值无限大？

朴素贝叶斯分类算法（Naive Bayesian Classifier）

训练过程：

1. 估计$p(y=0)$, $p(y=1)$, $p(x_{.i}|y=0)$, $p(x_{.i}|y=1)$, $i=1,2,\cdots,m$.

测试过程

1. 计算$p(y=0|\boldsymbol{x})$和$p(y=1|\boldsymbol{x})$.
2. 如果$p(y=0|\boldsymbol{x})>p(y=1|\boldsymbol{x})$，预测标签$\hat{y}=0$.
3. 如果$p(y=0|\boldsymbol{x})<p(y=1|\boldsymbol{x})$，预测标签$\hat{y}=1$.

朴素贝叶斯分类算法（Naive Bayesian Classifier）

朴素贝叶斯学习优缺点

优点：

- 易于理解，代码易于实现；
- 训练容易，没有复杂的参数估计问题；
- 预测速度快，可在线运行。

缺点：

- 标签给定条件下特征相互独立假设；
- 小概率特征概率估计不准确。

朴素贝叶斯分类算法（Naive Bayesian Classifier）

典型应用

- 垃圾邮件分类
- 本文情感分类
- 在线行为预测