

# CH04 朴素贝叶斯法

---

## CH04 朴素贝叶斯法

前言

章节目录

导读

朴素贝叶斯法

参数数量

算法推导

条件独立假设

参数估计

极大似然估计

贝叶斯估计

例子

例4.1

例子4.2

扩展

树增强朴素贝叶斯

贝叶斯网

LR与NB

参考

## 前言

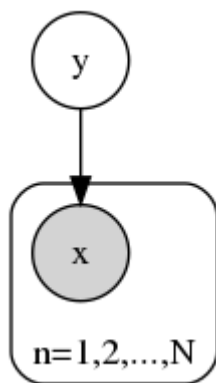
## 章节目录

1. 朴素贝叶斯法的学习与分类
  - a. 基本方法
  - b. 后验概率最大化的含义
2. 朴素贝叶斯法的参数估计
  - a. 极大似然估计
  - b. 学习与分类算法
  - c. 贝叶斯估计

## 导读

- 书中讲解的生成模型不多，生成模型学习联合概率分布，NB是典型的生成模型，与之对应的判别模型，是逻辑回归可以参考[CH06](#)
- **0-1损失函数**时的期望风险最小化
- 给定数据集的时候，我们能知道不同类别的分布，这个很好统计的量叫先验 $P(Y = c_k)$ 
  - 垃圾邮件识别可以用朴素贝叶斯的方法

- NB是常见的概率图模型之一<sup>[4]</sup>。最简单的概率图模型，有向图。概率图模型是执行贝叶斯推理的主流工具。



这个图中， $y$ 是输出变量， $x$ 是输入变量，是观测变量。生成模型直接描述了输出以多大的概率生成输入。在图模型里，已知观测是灰色的，未知变量是白色的。

- 第一个参考文献<sup>1</sup>，这里推荐阅读，链接中的手稿在2017年还有更新，开头部分介绍了很多符号的定义，可以配合CH01的内容来理解。讲述了生成模型和判别模型，朴素贝叶斯和逻辑回归
- 另外在讲到CRF的时候，也有对比贝叶斯和逻辑回归，可以在CH11的时候再回顾这里。

## 朴素贝叶斯法

### 参数数量

书中有这样一段内容

条件概率分布 $P(X = x|Y = c_k)$ 有指数级数量的参数，其实际估计是不可行的。

上面这段有两点：

#### 1. 指数级数量的参数

这点书后面解释了， $K \prod_{j=1}^n S_j$ ，这个为什么是指数级数量，假设 $S_j = S$ ，表达式变为 $KS^n$ ，实际上在参考文献<sup>2</sup>中有这个问题的讨论，里面用的是二值函数，可以参考理解

#### 2. 实际估计是不可行的

估计这么多参数需要更多的样本来刷参数，实际上获取这么多样本是不可行的。

### 算法推导

朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。

- 贝叶斯定理
- 特征条件独立假设

## 条件独立假设

### independent and identically distributed (i.i.d. or iid or IID)

求 $P(Y|X)$ ，其中 $X \in \{X_1, X_2, \dots, X_n\}$ ，条件独立假设这里给定 $Y$ 的情况下：

1. 每一个 $X_i$ 和其他的每个 $X_k$ 是条件独立的
2. 每一个 $X_i$ 和其他的每个 $X_k$ 的子集是条件独立的

条件独立性假设是：

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

上面这个公式可能看起来不是太容易理解独立在哪里，这里引用一下文献<sup>2</sup>中关于贝叶斯算法推导中的一部分

$$\begin{aligned} P(X|Y) &= P(X_1, X_2|Y) \\ &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

红色部分从上到下基于IID

条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的。

## 参数估计

### 极大似然估计

为了估计状态变量的条件分布，利用贝叶斯法则，有

$$\underbrace{P(X|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|X)P(X)}^{\text{likelihood prior}}}{\underbrace{P(Y)}_{\text{evidence}}} = \frac{\overbrace{P(Y|X)P(X)}^{\text{likelihood prior}}}{\underbrace{\sum_x P(Y|X)P(X)}_{\text{evidence}}}$$

其中 $P(X|Y)$ 为给定 $Y$ 下 $X$ 的后验概率(Posterior)， $P(Y|X)$ 称为似然(Likelihood)， $P(X)$ 称为先验(Prior)<sup>1</sup>。

- 后验概率最大化的含义

朴素贝叶斯法将实例分到后验概率最大的类中，这等价于期望风险最小化。

- 后验，观察到 $Y$ 之后，对 $X$ 的信念

## 贝叶斯估计

对于 $x$ 的某个特征的取值没有在先验中出现的情况，如果用极大似然估计，这种情况的可能性就是0。

但是出现这种情况的原因通常是因为数据集不能全覆盖样本空间，出现未知情况处理的策略就是做平滑。

公式(4.10)对应了出现未知样本的情况下，该给出一个什么样的值才合理的方案。

$$P_{\lambda}(X^{(j)} = a_{jl} | Y = c_k) = \frac{\sum_{i=1}^N I(x_i^j = a_{jl}, y_j = c_k) + \lambda}{\sum_{i=1}^N I(y_j = c_k) + S_j \lambda}$$

其中 $\lambda \geq 0$

当 $\lambda = 0$ 的时候，就是极大似然估计。

当 $\lambda = 1$ 的时候，这个平滑方案叫做Laplace Smoothing。拉普拉斯平滑相当于给未知变量给定了先验概率。

## 例子

### 例4.1

这个例子，有两个特征，一个标签，我们看看通过已知的数据表4.1能拿到哪些计算结果

1. 先验Prior，通过统计 $Y$ 的数据分布可以知道
2. 不同 $X$ 和 $Y$ 的组合会产生多少参数， $X^{(1)}$ 可能的取值集合 $A_1 = \{1, 2, 3\}$ 大小为 $S_1 = 3$ ， $X^{(2)}$ 可能的取值集合 $A_2 = \{S, M, L\}$ 大小为 $S_2 = 3$ ， $Y \in C = \{1, -1\}$ 大小为 $K = 2$   
参数的数量为 $K S_1 S_2 = 18$ ，具体的空间的分布是一个 $3 \times 3 \times 2$ 的三维矩阵
3. 每个特征的增加，本来应该在原来的 $Y$ ， $X$ 的基础上增加 $S_i$ 倍的维度，但因为做了特征条件独立假设，增加的可能性，是base在给定的标签 $Y$ 上的，也就是说实际上增加了 $S_i$ 个取值
4. 这个题的解题过程可以考虑为：Groupby  $Y$ ；Groupby  $Y$ ,  $X_1$ ；Groupby  $Y$ ,  $X_2$ ；对于新样本查表连乘。

### 例子4.2

## 扩展

### 树增强朴素贝叶斯

IID强限制放宽，就是TAN(Tree Augmented Naive Bayes)。可以看做是NB到贝叶斯网的过渡。

## 贝叶斯网

朴素贝叶斯法中假设输入变量都是条件独立的，如果假设他们之间存在概率依存关系，模型就变成了贝叶斯网络。这个书中简单提了一句作为扩展。

在SLAM的后端优化部分，因子图优化可以从概率的角度分析这个问题。从贝叶斯网络的角度，SLAM可以自然的表达成一个动态贝叶斯网络。在SLAM中有观测方程和运动方程，对应了贝叶斯网络中的条件概率关系。

## LR与NB

逻辑斯谛回归与朴素贝叶斯的关系

这个在参考文献<sup>2</sup>中有描述，另外在CH06中也有说明如下：

逻辑斯谛回归与朴素贝叶斯模型的关系参见文献[4]

这里文献[4]和本章的参考文献<sup>2</sup>同源，都是Mitchell TM的Machine Learning，现在依然是第一版，第二版的更新章节Tom有挂在网上<sup>3</sup>

这里插一句Mitchell教授的介绍，CMU计算机学院院长，1997年创建自动化学习和探索中心，该中心是全球的高校中首个机器学习系，也是首个开设机器学习博士课程的机构。1997-2016，Mitchell是该中心负责人。

这些都还OK，主要是Mitchell教授加入了松鼠AI。

## 参考

- 1.
- 2.
- 3.
- 4.

[↑ top](#)

---

1. 视觉SLAM十四讲, 高翔↩

2. Generative and discriminative classifiers: Naive Bayes and logistic regression↩↩↩

3. Machine Learning New Chapter↩