lec10.2 Naive Bayes

Naive Bayes Classifier

属于生成模型

独立事件与概率基础

在介绍贝叶斯前，首先介绍独立事件的基础知识

两个独立事件

无论 A、B 是否独立，都有

$$P(A,B)=P(A|B)P(B)$$

如果 A 独立于 B，有

$$P(A|B)=P(A)$$

则当 A、B 独立时，其联合概率为

$$P(A,B)=P(A)P(B)$$

多个独立事件

对于有限事件集 $\{A_1,\dots,A_n\}$，元素相互独立。对于任意 k 个事件的子集 $\{A_{i_1},\dots,A_{i_k}\}\subseteq\{A_1\dots,A_n\}$，我们有

$$P(A_{i_1},\dots,A_{i_k})=P(A_{i_1})\cdots P(A_{i_k})=\prod_{j=1}^{k}P(A_{i_j})$$

其中，$2\leq k\leq n$

这意味着每个子集的联合概率等于这些事件各自概率的乘积。换句话说，任何数量的事件的联合发生概率等于它们各自独立发生概率的乘积

先验与后验

- 后验概率：给定证据后，某一假设的概率
- 先验概率：在考虑证据之前，某一假设的概率
- 似然：在假设成立的条件下，证据出现的概率

Bayes' Rule 贝叶斯法则

$$P(H|E)=\frac{P(E|H)P(H)}{P(E)}$$

其中

- H 和 E 是事件
- $P(E)\neq 0$

- $P(H|E)$ 给定证据 E 后假设 H 的概率（后验概率）
- $P(E)$：证据 E 的边际概率
- $P(H)$：假设 H 的先验概率
- $P(E|H)$：在假设 H 成立时证据 E 出现的可能（似然）
- $P(H|E)$：在证据 E 出现后假设 H 的后验概率

贝叶斯法则在

- 直接从训练数据估计 $P(H|E)$ 较难
- 估计 $P(E|H)$、$P(H)$ 和 $P(E)$ 较易

时非常有用，因为它提供了一种从 $P(E|H)$、$P(H)$ 和 $P(E)$ 中计算 $P(H|E)$ 的方法。通过贝叶斯法则，我们可以根据新证据更新对某一假设的信念

推导：

$$P(H|E) = \frac{P(H,E)}{P(E)} \quad \text{and} \quad P(E|H) = \frac{P(H,E)}{P(H)}$$

其中，$P(H,E)$ 表示 E 和 H 同时发生时的联合概率，下面的步骤就不写了

通过一个例子来说：
假设我们有一个疾病诊断的问题：

- H 是某人有某种疾病的事件
- E 是某人测试结果为阳性的事件

显然，给定测试结果为阳性时直接计算某人有这种疾病的概率 $P(H|E)$ 非常困难；
但是在我们知道

- $P(H)$ 是某人有这种疾病的先验概率
- $P(E)$ 是测试结果为阳性的边际概率
- $P(E|H)$ 某人有这种疾病时测试结果为阳性的概率

时（这三个数据非常容易获得），通过贝叶斯法则计算 $P(H|E)$ 就非常简单了

例子1 单一特征

问题描述：

- 脑膜炎（Meningitis）有50%的概率导致脖子僵硬（stiff neck）
- 脑膜炎的发生概率是 $\frac{1}{50000}$，脖子僵硬的发生概率是 $\frac{1}{20}$
- 计算在患者脖子僵硬的情况下，患者患有脑膜炎的概率

定义事件：

- H：脑膜炎
- E：脖子僵硬

给定概率：

- $P(H) = \frac{1}{50000}$

- $P(E) = \frac{1}{20}$
- $P(E|H) = 0.5$

计算：

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{0.5 \times \frac{1}{50000}}{\frac{1}{20}} = 0.0002$$

在高维空间中的问题：
如果我们有一维空间（只有一个特征 $X = (a)$），那么我们可以直接从训练数据集中估计 $P(H|X)$

当我们有更高维度的数据时，例如 $X = (a_1, \ldots, a_d)$，$d > 1$，情况就更复杂了。这是因为在高维空间中直接估计 $P(H|X)$ 变得更加困难，需要更多的数据和更复杂的模型

Naive Bayes Approximation 朴素贝叶斯近似

C 是一个类随机变量，用来表示一个未见过的 d 维度的测试对象 $X = (x_1, \ldots, x_d)$。给定一个具体的测试对象 $(a_1, \ldots, a_d)$，目标是估计

$$P(C = c \mid X = (a_1, \ldots, a_d)) = P(C = c \mid x_1 = a_1, \ldots, x_d = a_d)$$

根据贝叶斯法则，我们有

$$P(C = c \mid x_1 = a_1, \ldots, x_d = a_d) = \frac{P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)}{P(x_1 = a_1, \ldots, x_d = a_d)}$$

其中，分母 $P(x_1 = a_1, \ldots, x_d = a_d)$ 不依赖于类变量 C

最大后验概率:
具有最大分子 $P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)$ 的类 c 具有最大的后验概率：

$$P(C = c \mid x_1 = a_1, \ldots, x_d = a_d)$$

朴素贝叶斯假设：

- $P(C = c)$：可以估计为属于类 c 的训练数据对象的比例，是其先验概率
- $\prod_{i=1}^{d} P(x_i = a_i \mid C = c)$ 是在类别 c 下各个特征值 a_i 出现的联合概率
- 估计 $P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)$ 可以使用朴素假设：不同特征 $x_1, \ldots, x_d$ 的值在给定类的条件下相互独立

$$P(x_1 = a_1, \ldots, x_d = a_d \mid C = c) = \prod_{i=1}^{d} P(x_i = a_i \mid C = c)$$

估计 $P(x_i = a_i \mid C = c)$
可以通过计算训练数据中属于类 c 的对象中具有特征 $x_i = a_i$ 的比例来估计，这种方法比直接从训练数据中估计 $P(x_1 = a_1, \ldots, x_d = a_d)$ 要容易的多

根据以上思路，在独立性假设下，$P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)$ 的估计变为：

$$P(C = c) \prod_{i=1}^{d} P(x_i = a_i \mid C = c)$$

例子2 两个特征

定义事件：

- H = 引擎无法启动
- 证据 A = 电池电量不足；证据 B = 没有油

直接估计：
为了直接估计 $P(H|A,B)$，我们需要只考虑数据集中那些电池电量不足且没有油的汽车。在这些汽车中，我们需要统计引擎无法启动的汽车数量。这种情况在数据集中可能很少见，导致 $P(H|A,B)$ 的估计不可靠或为零（在最坏情况下）

$$P(H|A,B) = \frac{\text{电池电量不足 且 没有油 且 引擎无法启动 的 汽车数量}}{\text{电池电量不足 且 没有油 的 汽车数量}}$$

贝叶斯估计：

$$P(H|A,B) = \frac{P(A,B|H)P(H)}{P(A,B)}$$

通常 $P(A,B|H)$ 并不好计算，因此我们可以使用朴素贝叶斯近似来估计上式中的后验概率（这种方法更容易实现）

朴素贝叶斯估计：

$$P(H|A,B) = \frac{P(A,B|H)P(H)}{P(A,B)} = \frac{P(A|H)P(B|H)P(H)}{P(A,B)}$$

其中

- $P(A|H)$ 可以估计为引擎无法启动的汽车中电池电量不足的汽车比例
- $P(B|H)$ 可以估计为引擎无法启动的汽车中没有油的汽车比例

比例形式

假设 $X = (a_1, \ldots, a_k)$ 是输入的测试对象，我们需要从集合 $\{c_1, \ldots, c_k\}$ 中选择或预测 X 的类

计算后验概率：

$$P(C = c \mid x_1 = a_1, \ldots, x_d = a_d) = \frac{P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)}{P(x_1 = a_1, \ldots, x_d = a_d)}$$

比例形式：
由于分母 $P(x_1 = a_1, \ldots, x_d = a_d)$ 对所有类 C 是相同的，因此后验概率

$$P(C = c \mid x_1 = a_1, \ldots, x_d = a_d)$$

与分子

$$P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)$$

成正比。因此，

$$P(C = c \mid x_1 = a_1, \ldots, x_d = a_d) \propto P(C = c)P(x_1 = a_1, \ldots, x_d = a_d \mid C = c)$$

上式可以被简化为

$$\begin{array}{ccccc} P(C|X) & \propto & P(C) & \times & P(X|C) \\ \text{posterior} & \propto & \text{prior} & \times & \text{likelihood} \end{array}$$

例子

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

实例：$X = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$\begin{aligned} P(\text{Play} = \text{yes} \mid X) &\propto P(X \mid \text{Play} = \text{yes}) P(\text{Play} = \text{yes}) \\ &= P(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{yes}) \times P(\text{Temp} = \text{cool} \mid \text{Play} = \text{yes}) \\ &\quad \times P(\text{Humidity} = \text{high} \mid \text{Play} = \text{yes}) \times P(\text{Windy} = \text{true} \mid \text{Play} = \text{yes}) \\ &\quad \times P(\text{Play} = \text{yes}) \\ &= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} \\ &= 0.00529 \end{aligned}$$

$$\begin{aligned} P(\text{Play} = \text{no} \mid X) &\propto P(X \mid \text{Play} = \text{no}) P(\text{Play} = \text{no}) \\ &= P(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{no}) \times P(\text{Temp} = \text{cool} \mid \text{Play} = \text{no}) \\ &\quad \times P(\text{Humidity} = \text{high} \mid \text{Play} = \text{no}) \times P(\text{Windy} = \text{true} \mid \text{Play} = \text{no}) \\ &\quad \times P(\text{Play} = \text{no}) \\ &= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{5}{14} \times \frac{9}{14} \\ &= 0.020 \end{aligned}$$

因此，$\text{Play} = \text{no}$

分类器任务

给定一个测试对象 $x = (a_1, \ldots, a_k)$，我们想要预测它的类别 C

$$P(C = c \mid X = x) \propto P(C = c) \prod_{i=1}^{d} P(x_i = a_i \mid C = c)$$

这之后，选择使得上述概率最大化的类别 $c \in \{c_1, \ldots, c_k\}$

对象排序

给定一组类别 $\{c_1, \ldots, c_k\}$ 和一组测试对象 $x_1, \ldots, x_t$，目的是根据这些对象属于某个特定类别 c 的概率来对它们进行排名

计算每个测试对象属于类别 c 的实际概率：

$$P(C = c \mid X = x_1), \ldots, P(C = c \mid X = x_t)$$

对于一个固定的对象 $x = (a_1, \ldots, a_d)$，我们有

$$\sum_{i=1}^{k} P(C = c_i \mid X = x) = 1$$

x 属于类别 c 的概率为

$$P(C=c \mid X=x)=\frac{P(C=c \mid X=x)}{\sum_{i=1}^{k} P(C=c_i \mid X=x)}=\frac{P(C=c)P(X=x \mid C=c)}{\sum_{i=1}^{k} P(C=c_i)P(X=x \mid C=c_i)}$$

例子

Example: predicting whether to play or not

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Computing probabilities $\overline{X}$ = (Outlook = sunny, Temp = cool, Humidity = high, Windy = true)

$$P(\text{Play = yes} \mid \overline{X})=\frac{0.00529}{0.00529+0.02}=0.209 \qquad P(\text{Play = no} \mid \overline{X})=\frac{0.02}{0.00529+0.02}=0.791$$

朴素假设的注解

- 朴素贝叶斯分类器假设给定目标变量后，所有特征都是相互独立的。在实际应用中，这一假设通常是不成立的，因为真实世界的数据特征之间往往存在某种程度的相关性
- 尽管基于一个简化的假设，朴素贝叶斯分类器在某些领域（如文本分类和垃圾邮件检测）中表现出了出奇的效果。这表明即便假设不完全符合实际，该模型依然能够提供有用的结果
- 朴素贝叶斯提供了一种估计一组随机变量的联合分布的方法，而不必担心数据稀疏性问题。这主要得益于其将多维特征空间的复杂依赖关系简化为一维的独立关系，从而降低了模型从数据中学习的复杂性和计算需求
- 一些线性分类器（如感知机）也采用了关于特征独立性的假设。在这些模型中，虽然没有明确声明特征独立，但模型设计（如线性权重和）隐含了特征之间相互独立的影响，这些假设帮助简化了问题，使得模型易于理解和实现