数据挖掘原理与应用

第四章-朴素贝叶斯分类器

叶志鹏

南京理工大学泰州科技学院

1st Jan, 2023

思考以下几个问题？

1. 已知明天下雨的可能性是 $\frac{1}{10}$，刮风天下雨的可能性也是 $\frac{1}{10}$ 吗？
2. 四个选项的选择题蒙对的概率是 $\frac{1}{4}$，对于不同学生来说做对选择题的概率也是 $\frac{1}{4}$ 吗？
3. 掷均匀的硬币，掷一次硬币正面朝上的可能性为 $\frac{1}{2}$，掷 10 次正面朝上后，下一次还是正面朝上的可能性还是 $\frac{1}{2}$ 吗？
4. 在连续两次掷骰子的实验中，已知两次抛掷的点数和为 9，第一次抛掷的点数为 6 的可能性有多大？

条件概率公式

- 条件概率 [BT02] 是在给定部分信息的基础上对实验结果的一种推断。
- 假设事件 B 已经发生，求 B 事件发生条件下，A 事件发生的可能性，记为 $P(A|B)$。公式如下：

$$P(A|B)=\frac{\text{事件}A\cap B\text{的实验结果数}}{\text{事件}B\text{的试验结果数}}=\frac{P(A,B)}{P(B)}$$

条件概率的例子

连续三次抛掷一个两面均匀的硬币的实验中，A 事件为正面出现的次数多于反面出现的次数，B 事件为第一次抛掷得到正面，求 P(A|B)。

样本空间由下列 8 个实验结果组成：

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

因此 $P(B) = \frac{4}{8}$, 而事件 $A \cap B$ 的结果由 HHH,HHT,HTH 组成，其概率为：

$$P(A, B) = \frac{3}{8}$$

最后可以得到：

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{3/8}{4/8} = \frac{3}{4}$$

独立性的基本概念

条件概率 $P(A|B)$ 刻画了事件 B 的发生给事件 A 带来了信息。如果 B 事件的发生不给 A 事件带来信息，没有改变事件 A 发生的概率。则称 A,B 两个事件独立，即

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

因为 $P(A|B) = \frac{P(A,B)}{P(B)}$，所以

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

独立性的示例

考虑连续两次抛掷一个具有 4 个面对称的骰子，求下列 A，B 事件是否独立?

1. A={第一次抛掷后得到 i}, B={第二次抛掷后得到 j}
2. A={第一次抛掷后得到 1}，B={两次抛掷的总和为 5}
3. A={两次抛掷的最大数为 2}，B={两次抛掷的最小数为 2}

条件独立

在给定 C 事件下，若事件 A 和事件 B 满足

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

则称 A 和 B 在给定 C 之下条件独立。利用条件概率的定义和乘法规则，可以推导条件独立的另一个公式 $P(A|B, C) = P(A|C)$：

$$\begin{aligned} P(A, B|C) &= \frac{P(A, B, C)}{P(C)} \\ &= \frac{P(C)P(B|C)P(A|B, C)}{P(C)} \\ &= P(B|C)P(A|B, C) \end{aligned}$$

只要 $P(B|C) \neq 0$，那么 $P(A|B, C) = P(A|C)$。

一组事件的独立性

设 $A_1, A_2, ..., A_n$ 为 n 个事件，若它们满足

$$P(A_1, A_2, ..., A_n) = P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

对 {1,2,...,n } 的任意子集 S 成立，则称 $A_1, ..., A_n$ 为相互独立的事件。

全概率定理

设 $A_1, A_2, ..., A_n$ 是一组互不相容的事件，形成样本空间的一个分割（每一个试验结果必定使得其中一个事件发生），又假定对每一个 $i, P(A_i) > 0$。则对于任何事件 B，下列公式成立

$$
\begin{aligned}
P(B) &= P(A_1, B) + ... + P(A_n, B) \\
&= P(A_1)P(B|A_1) + ... + P(A_n)P(B|A_n)
\end{aligned}
$$

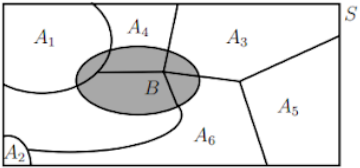

图 1: 全概率定理原理图

全概率定理例题

你参加一个棋类比赛，其中 50% 是一类棋手，赢他们的概率为 0.3，25% 是二类棋手，赢他们的概率为 0.4，剩下三类棋手，赢他们的概率为 0.5，求你胜算的概率？

设 A_i 表示你于 i 类棋手相遇的事件：

$$P(A_1) = 0.5, P(A_2) = 0.25, P(A_3) = 0.25$$

B 为你赢得比赛的事件，则

$$P(B|A_1) = 0.3, P(B|A_2) = 0.4, P(B|A_3) = 0.5$$

所以

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \\ &= 0.5 * 0.3 + 0.25 * 0.4 + 0.25 * 0.5 \\ &= 0.375 \end{aligned}$$

1 数学回顾

2 朴素贝叶斯算法原理
- 贝叶斯示例
- 贝叶斯定理
- 多维属性的联合概率
- 条件独立性假设
- 贝叶斯分类案例
- 连续型朴素贝叶斯算法
- 朴素贝叶斯算法的总结

3 朴素贝叶斯算法文本分类实践

4 参考文献

1 数学回顾

2 朴素贝叶斯算法原理
- 贝叶斯示例
- 贝叶斯定理
- 多维属性的联合概率
- 条件独立性假设
- 贝叶斯分类案例
- 连续型朴素贝叶斯算法
- 朴素贝叶斯算法的总结

3 朴素贝叶斯算法文本分类实践

4 参考文献

贝叶斯示例

一所学校的学生中，60% 为男生，40% 为女生。男生总是爱穿长裤，女生则一半长裤，一半裙子。随机选取一个穿长裤的学生，他（她）是女生的概率为多大？

贝叶斯示例

$$P(Boy) = 60\%\ P(Girl) = 40\%,$$
$$P(Pants|Boy) = 1, P(Pants|Girl) = \frac{1}{2}$$

、

$$\begin{aligned}
P(Girl|Pants) &= \frac{P(Girl, Pants)}{P(Pants)} = \frac{P(Girl)P(Pants|Girl)}{P(Pants)} \\
&= \frac{P(Girl)P(Pants|Girl)}{P(Girl)P(Pants|Girl) + P(Boy)P(Pants|Boy)} \\
&= \frac{0.4 * 0.5}{0.4 * 0.5 + 0.6 * 1} \\
&= 0.25
\end{aligned}$$

贝叶斯定理

假设 x 为我们的样本属性，y_i 为样本的标签。根据贝叶斯公式:

$$P(y_i|x) = \frac{P(y_i)P(x|y_i)}{P(x)}$$

我们将 $P(y_i|x)$，已知样本属性信息求样本对用类别的概率记为后验概率。$P(y_i)$ 随机选取样本为类别 i 的概率为先验概率。
$P(x|y_i)$ 已知样本类别下样本属性 x 的概率为似然函数 (Likelihood)

极大后验假设

贝叶斯分类器的方法就是通过比较 $P(y_i|x)$ 间的概率大小，继而选择概率最大的对应的类别。记为

$$i_{MAP} = argmax_{i \in I} P(y_i|x) = argmax_{i \in I} \frac{P(y_i)P(x|y_i)}{P(x)}$$

因为对于每个类别 i 来说，都存在分母 P(x)，所以化简得：

$$i_{MAP} = argmax_{i \in I} P(y_i|x) = argmax_{i \in I} P(y_i)P(x|y_i)$$

我们将 i_{MAP} 称为极大后验假设（Maximum A Posteriori, MAP）

多维属性的联合概率

已知样本对象是由多个属性组成的特征向量，$X = x_1, ..., x_n$，那么

$$
\begin{aligned}
i_{MAP} = & argmax_{i \in I} P(y_i|X) = argmax_{i \in I} P(y_i) P(X|y_i) \\
= & argmax_{i \in I} P(y_i) P(< x_1, ..., x_n > |y_i)
\end{aligned}
$$

当特征向量维度过高时，可用数据会变得稀疏。

条件独立性假设

当样本的特征向量较多时，$P(< x_1, ..., x_n > |y_i)$ 不易被计算，所以我们加入条件独立性假设，所以：

$$\begin{aligned} i_{MAP} &= argmax_{i \in I} P(y_i) P(< x_1, ..., x_n > |y_i) \\ &= argmax_{i \in I} P(y_i) \prod_j P(x_j|y_i) \end{aligned}$$

到此就是完整的朴素贝叶斯分类器的核心原理。

贝叶斯分类案例

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

通过朴素贝叶斯分类器，判断一个收入中等，信用中的青年爱好游戏顾客，是否会购买电脑？训练集如左表所示。

朴素贝叶斯分类案例

$$P(\text{不购买} \mid <\text{青，中，是，中}>) \Longrightarrow P(X| \text{不购买})P(\text{不购买})$$
$$= P(\text{青} \mid \text{不购买})P(\text{中} \mid \text{不购买})P(\text{是} \mid \text{不购买})$$
$$P(\text{中} \mid \text{不购买})P(\text{不购买})$$
$$= 0.6 * 0.4 * 0.2 * 0.4 * 0.357 = 0.007$$

同理，$P(\text{购买} \mid <\text{青，中，是，中}>) \Longrightarrow 0.028$。
$0.028 > 0.007$
所以通过朴素贝叶斯分类器可得出一个收入中等，信用中的青年爱好游戏顾客会购买电脑。

1 数学回顾

2 朴素贝叶斯算法原理
- 贝叶斯示例
- 贝叶斯定理
- 多维属性的联合概率
- 条件独立性假设
- 贝叶斯分类案例
- 连续型朴素贝叶斯算法
- 朴素贝叶斯算法的总结

3 朴素贝叶斯算法文本分类实践

4 参考文献

连续型朴素贝叶斯算法

id	收入	购买
1	125	否
2	100	否
3	70	否
4	120	否
5	95	是
6	60	否
7	220	否
8	85	是
9	75	否
10	90	是

对于连续性朴素贝叶斯算法有两种方法处理：

1. 将连续数据离散化
2. 假设数据服从 $N(\mu,\sigma^2)$ 的正态分布，通过训练数据估计出 μ,σ，再用求得的分布，算出对于的条件概率。

$$P\left(X_i \mid c_j\right)=\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}e^{-\frac{\left(x_i-\mu_{ij}\right)^2}{2\sigma_{ij}^2}}$$

$$P\left(X_i \mid c_j\right)=\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}e^{-\frac{\left(x_i-\mu_{ij}\right)^2}{2\sigma_{ij}^2}}$$

朴素贝叶斯算法的总结

- 本质上是同时考虑了先验概率和似然概率的重要性
- 属性可以离散、也可以连续
- 数学基础坚实、分类效率稳定
- 对缺失和噪声数据不太敏感
- 属性如果不相关，分类效果很好

1 数学回顾

2 朴素贝叶斯算法原理

3 朴素贝叶斯算法文本分类实践
- 数据集介绍
- 数据预处理
- 数据变换/特征工程
- 模型训练
- 模型评估

4 参考文献

数据集介绍

- Iris [Fis88] 鸢尾花数据集内包含 3 种类别，分别为山鸢尾（Iris-setosa）、变色鸢尾（Iris-versicolor）和维吉尼亚鸢尾（Iris-virginica）。数据集共 150 条记录，每类各 50 个数据，每条记录有花萼长度、花萼宽度、花瓣长度、花瓣宽度 4 项特征 (cm)。

图 2: Setosa

图 3: Versicolour

图 4: Virginica

1. 数学回顾
2. 朴素贝叶斯算法原理
3. 朴素贝叶斯算法文本分类实践
   - 数据集介绍
   - 数据预处理
   - 数据变换/特征工程
   - 模型训练
   - 模型评估
4. 参考文献

1. 数学回顾
2. 朴素贝叶斯算法原理
3. 朴素贝叶斯算法文本分类实践
   - 数据集介绍
   - 数据预处理
   - 数据变换/特征工程
   - 模型训练
   - 模型评估
4. 参考文献

1. 数学回顾
2. 朴素贝叶斯算法原理
3. 朴素贝叶斯算法文本分类实践
4. 参考文献

[BT02] Dimitri P. Bertsekas and John N. Tsitsiklis.
Introduction to probability.
2002.

[Fis88] R.A. Fisher.
Iris.
UCI Machine Learning Repository, 1988.

Thanks!