

# 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



- ① 数学回顾
- ② 朴素贝叶斯算法原理
- ③ 朴素贝叶斯算法 Iris 数据集实践
- ④ 参考文献

## ① 数学回顾

条件概率

独立性

全概率定理

## ② 朴素贝叶斯算法原理

## ③ 朴素贝叶斯算法 Iris 数据集实践

## ④ 参考文献

## ① 数学回顾

条件概率

独立性

全概率定理

## ② 朴素贝叶斯算法原理

## ③ 朴素贝叶斯算法 Iris 数据集实践

## ④ 参考文献

# 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  $P(A|B)$ 。

# 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：

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

## ① 数学回顾

条件概率

独立性

全概率定理

## ② 朴素贝叶斯算法原理

## ③ 朴素贝叶斯算法 Iris 数据集实践

## ④ 参考文献



# 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

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

$P(A|B) = P(A)$  或  $P(A, B) = P(A)P(B)$  是 A, B 两个事件独立的充分必要条件。

# 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  $A = \{\text{第一次抛掷后得到 1}\}$ ,  $B = \{\text{两次抛掷的总和为 5}\}$
- ②  $A = \{\text{第一次抛掷后得到 1}\}$ ,  $B = \{\text{两次抛掷的总和为 3}\}$
- ③  $A = \{\text{两次抛掷的最大数为 2}\}$ ,  $B = \{\text{两次抛掷的最小数为 2}\}$

## 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  $P(A|B, C) = P(A|C)$ ：

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

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  $A_1, A_2, \dots, A_n$  为  $n$  个事件, 若它们满足

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

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

## ① 数学回顾

条件概率

独立性

全概率定理

## ② 朴素贝叶斯算法原理

## ③ 朴素贝叶斯算法 Iris 数据集实践

## ④ 参考文献

# 全概率定理

设  $A_1, A_2, \dots, A_n$  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

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

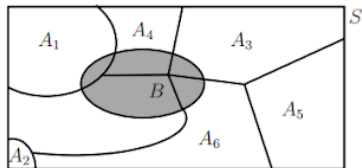


图 1: 全概率定理原理图

## 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

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

# 课后作业

- 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 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？
- ③ 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？



## 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 Iris 数据集实践

## 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 Iris 数据集实践

## 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

$$P(\text{Boy}) = 60\% \quad P(\text{Girl}) = 40\%,$$

$$P(\text{Pants}|\text{Boy}) = 1, \quad P(\text{Pants}|\text{Girl}) = \frac{1}{2}$$

、

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

## 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 Iris 数据集实践

## ④ 参考文献

# 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称为后验概率。 $P(y_i)$  随机选取样本为类别  $i$  的概率称为先验概率。 $P(x|y_i)$  已知样本类别下样本属性  $x$  的概率称为似然函数 (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

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

因为对于每个类别  $i$  来说，都存在分母  $P(x)$ ，所以化简得：

$$i_{MAP} = \operatorname{argmax}_{i \in I} P(y_i|x) \propto \operatorname{argmax}_{i \in I} P(y_i)P(x|y_i)$$

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

## 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 Iris 数据集实践

## ④ 参考文献



# 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,  $X = x_1, \dots, x_n$ , 那么

$$\begin{aligned} i_{MAP} &= \operatorname{argmax}_{i \in I} P(y_i | X) \propto \operatorname{argmax}_{i \in I} P(y_i) P(X | y_i) \\ &= \operatorname{argmax}_{i \in I} P(y_i) P(\langle x_1, \dots, x_n \rangle | y_i) \end{aligned}$$

当特征向量维度过高时, 可用数据会变得稀疏。

## 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 Iris 数据集实践

## ④ 参考文献

## 条件独立性假设

当样本的特征向量较多时,  $P(< x_1, \dots, x_n > | y_i)$  不易被计算 (当测试样本不在训练样本中时), 所以我们加入条件独立性假设, 所以:

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

到此就是完整的朴素贝叶斯分类器的核心原理。

## 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 Iris 数据集实践

## ④ 参考文献

# 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

$$\begin{aligned}
 P(\text{不购买} | <\text{青}, \text{中}, \text{是}, \text{中}>) &\propto P(X | \text{不购买})P(\text{不购买}) \\
 &= P(\text{青} | \text{不购买})P(\text{中} | \text{不购买})P(\text{是} | \text{不购买}) \\
 &\quad P(\text{中} | \text{不购买})P(\text{不购买}) \\
 &= 0.6 * 0.4 * 0.2 * 0.4 * 0.357 = 0.007
 \end{aligned}$$

同理,  $P(\text{购买} | <\text{青}, \text{中}, \text{是}, \text{中}>) \implies 0.028$ 。

$0.028 > 0.007$

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 Iris 数据集实践

## ④ 参考文献

连续型朴素贝叶斯算法

对于连续性朴素贝叶斯算法有两种方法处理：

id	收入	购茨
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

- 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  $\mu, \sigma$ ，再用求得的分布，算出对于的条件概率。

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

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



## 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 Iris 数据集实践

## ④ 参考文献

# 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 Iris 数据集实践
  - 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 Iris 数据集实践
  - 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

- 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



 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 Iris 数据集实践
  - 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 Iris 数据集实践
  - 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 Iris 数据集实践
  - 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 Iris 数据集实践
  - 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

# 模型评估

- 模型训练好了之后还需要对模型进行评估好坏。通过模型评估进一步选择不同模型，策略和算法来达到数据挖掘的目的。
- 常见的分类算法的评估指标有混淆矩阵、准确率、精确率、召回率、ROC 曲线等。

# 混淆矩阵

		Actual class	
		positive class	negative class
Predicted class	positive class	True Positive(TP)	False Positive(FP)
	negative class	False Negative(FN)	True Negative(TN)

图 5: 混淆矩阵

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用 n 行 n 列的矩阵来表示，主要包括如下 4 个要素：

# 混淆矩阵

- 对于 2 分类来说，如 A, B 两类，选择 A 为正类的话。
- 若一个实例是正类，且被预测为正类，即为真正类 (True Positive TP)
  - 若一个实例是负类，且被预测为负类，即为真负类 (True Negative TN)
  - 若一个实例是正类，且被预测为负类，即为假负类 (False Negative FN)
  - 若一个实例是负类，且被预测为正类，即为假正类 (False Positive FP)

对于多分类来说，将一个类作为正类，其他类作为负类。

# 准确率

准确率是最常见的分类任务评价指标，即被准确分类的样本数占总样本数的比例，一般而言，准确率越高，分类器性能越好。

$$\text{准确率}(\text{accuracy}) = \frac{TP + TN}{TP + TN + FP + FN}$$

但当数据分布不均衡时，准确率指标并不能很好的评价模型的性能。比如预测地震数据集，可能 100 个样本，才有 1 次地震样本。如果分类错误，分类器的准确率也达到 99%。这显然不是我们想要的，还需参考其他指标。

## 精确率、召回率与 $F_1$

- 精确率 (precision) 是针对预测结果而言。表示的是预测为正的样本中有多少是真正的正样本。预测为正类有两种可能性，一种是把正类预测为正类，另一种是把负类也预测为正类。公式如下：

$$\text{精确率}(\text{precision}) = \frac{TP}{TP + FP}$$

- 召回率 (recall) 是针对测试样本而言，表示样本中正例有多少被预测正确了。这里面也有两种可能性，一种是把原来的正类预测成了正类，另一种是把原来的正类预测成了负类。公式如下：

$$\text{召回率}(\text{recall}) = \frac{TP}{TP + FN}$$

# 权衡精确率与召回率

- 对于地震场景，我们希望召回率（recall）非常高，比如我们情愿发出 1000 次警报，10 次地震预测正确了（recall 100%，precesion 1%），也不要发出 100 次警报，8 次地震预测正确，2 次漏报（recall 80%，precesion 8%）。
- 对于买瓜场景，不可能挑边所有瓜，所以我们希望 precision 要高，recall 低也是值得的。
- 另外一些场景需要权衡就有了  $F_1$  score，即精确率与召回率的调和均值。

$$F_1 = \frac{2rp}{r + p} = \frac{2 * TP}{2 * TP + FP + FN}$$

# 练习题

假设我们有 60 个样本，40 个负样本，要找出所有的正样本。系统查找出 50 个正样本，其中 40 个是真正的正样本，计算上述各指标。



## 练习题

假设我们有 60 个样本，40 个负样本，要找出所有的正样本。系统查找出 50 个正样本，其中 40 个是真正的正样本，计算上述各指标。

$$TP = 40 \quad FN = 20$$

$$FP = 10 \quad TN = 30$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.7$$

$$precision = \frac{TP}{TP + FP} = 0.8$$

$$recall = \frac{TP}{TP + FN} = \frac{1}{3}$$

$$F_1 = \frac{2rp}{r + p} = 0.47$$

- ① 数学回顾
- ② 朴素贝叶斯算法原理
- ③ 朴素贝叶斯算法 Iris 数据集实践
- 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!*