贝叶斯分类器习题

浙江大学

赵洲

题目

- 1. 以下关于朴素贝叶斯分类器的描述，哪一项是正确的?

A. 朴素贝叶斯假设特征之间完全相关

B. 朴素贝叶斯只能用于二分类问题

C. 朴素贝叶斯假设特征之间条件独立

D. 朴素贝叶斯不适用于文本分类

解答

- A. 朴素贝叶斯假设特征之间完全相关

- **选项A错误**
- 朴素贝叶斯（Naive Bayes）分类器的核心假设是特征之间条件独立，而不是特征完全相关。具体来说，朴素贝叶斯模型假设，在给定类别标签的条件下，特征之间是独立的。这一假设被称为“朴素性假设”。
- 即假设在给定类别的情况下，特征之间不再有任何依赖关系。这是其名称中“朴素”（naive）的来源。

解答

B. 朴素贝叶斯只能用于二分类问题

- **选项B错误：** 朴素贝叶斯可以用于多分类问题，不仅限于二分类。
- 二分类问题：这是指分类任务只有两个类别的情况，例如垃圾邮件分类（垃圾邮件与非垃圾邮件）。
- 多分类问题：这是指分类任务有三个或更多类别的情况，例如新闻文章分类（政治、体育、娱乐等多个类别）。
- 多类朴素贝叶斯（Multinomial Naive Bayes）
- 高斯朴素贝叶斯（Gaussian Naive Bayes）

解答

- C. 朴素贝叶斯假设特征之间条件独立
- **选项C正确：** 这正是朴素贝叶斯分类器的核心假设。

- 朴素贝叶斯分类器的核心假设是 **条件独立性假设**，即在给定类别的情况下，特征之间是相互独立的。这个假设简化了概率的计算，使得计算联合概率变得可行。
- 虽然现实中很多特征并不是完全独立的，但在许多应用中，朴素贝叶斯算法仍然表现得很好，特别是当特征之间的依赖关系较弱时。

解答

- D. 朴素贝叶斯不适用于文本分类

- **选项D错误**
- 朴素贝叶斯在文本分类中广泛应用，尤其是在垃圾邮件检测、情感分析等任务中。它基于特征条件独立的假设，在文本分类中通常假设每个单词（特征）在给定类别的条件下是独立的，能够有效处理这种高维稀疏数据。。

题目

- **2. 在使用朴素贝叶斯分类器时，如果遇到某一特征值在训练数据中从未出现过，导致条件概率为零，可以采用以下哪种方法来解决?**

A. 增加训练数据量。
B. 使用拉普拉斯平滑（Laplace Smoothing）。
C. 删除该特征。
D. 将条件概率设为极大值。

题目

- A. 增加训练数据量。
- 错误

- **解释**：增加训练数据量可能会让更多的特征值在数据中出现，从而减少未观察到的特征值问题。
- 但是，增加数据量并不能保证所有可能的特征值都会出现在训练集中，特别是对于高维稀疏数据（如文本分类）中的罕见特征。
- **结论**：此方法是有益的，但并不是直接解决问题的可靠手段，无法保证避免条件概率为零的情况。
- **不够直接和有效，不能作为首选方法。**

题目

- B. 使用拉普拉斯平滑（Laplace Smoothing）。
- 正确

- **解释**：拉普拉斯平滑是解决此类问题的标准方法。它通过给所有可能的特征值增加一个固定的平滑值（通常为 1），从而避免条件概率为零。
- 通过拉普拉斯平滑，即使某一特征值在训练集中从未出现过，其条件概率也会被赋予一个很小的值。
- **结论**：**最常用、最直接有效的解决方案。**

题目

- C. 删除该特征。
- 错误

- **解释**：如果某个特征在训练数据中从未出现，删除该特征似乎是一种可行的方法。
- 然而，删除特征可能会丢失有潜在价值的信息，尤其是在高维数据中。
- 此外，删除特征不能从根本上解决概率为零的问题，因为其他未出现的特征依然可能导致同样的问题。
- **结论**：**不推荐作为主要解决方案。**

题目

- D. 将条件概率设为极大值。
- 错误

- **解释**：将某一特征值的条件概率直接设为极大值是错误的，因为这会对分类结果产生误导性影响，完全违背概率规则。
- 如果一个特征从未在训练数据中出现，就将其概率设为极大值，不仅不合理，还可能严重偏离实际数据分布。
- **结论**：**完全错误的方法，不应采用。**

题目

- **3. 下面哪种情况下，朴素贝叶斯分类器的性能可能会受到严重影响?**

A. 特征之间高度相关。
B. 训练数据量充足。
C. 类别数量较少。
D. 特征是离散的。

解答

- A. 特征之间高度相关。**选项A正确**

- **解释**：朴素贝叶斯分类器假设特征是条件独立的，这意味着它假定各个特征之间没有依赖关系。然而，在现实世界的许多任务中，特征之间可能存在很强的相关性。像文本分类，某些单词可能总是同时出现，或者某些特征在类别间的分布高度重合。
- 当特征之间高度相关时，朴素贝叶斯模型的假设就不再成立，模型的预测性能可能会显著下降。因为在高度相关的特征中，某些特征的信息会被重复计算，导致不准确的概率估计。
- **结论：这是朴素贝叶斯分类器性能最容易受到严重影响的情况。**

解答

- B. 训练数据量充足。**选项B错误**

- **解释**：充足的训练数据量通常能够帮助模型更好地拟合数据，减少过拟合的风险，并提高模型的泛化能力。朴素贝叶斯算法的计算效率较高，数据量增加时，它依然能够高效地进行训练和分类。
- 数据量充足不会直接影响朴素贝叶斯的表现，反而会有助于模型更好地估计类别条件概率。
- **结论**：数据量充足通常是有利的，不会对模型性能造成严重影响。

解答

- C. 类别数量较少。**选项C错误**

- **解释**:

  - 类别数量较少一般不会对朴素贝叶斯的性能造成严重影响。即使类别较少，朴素贝叶斯依然能够有效地分类，因为其核心思想是基于每个类别的条件概率进行分类。类别数量的多少主要影响模型的计算复杂度，但不会直接导致性能下降。
  - 在类别数量少的情况下，朴素贝叶斯依然能够快速并且有效地进行训练和预测。

- **结论**:

  - **类别数量较少不会对模型性能产生严重影响**

解答

- C. 类别数量较少。**选项C错误**

- **解释**：类别数量较少一般不会对朴素贝叶斯的性能造成严重影响。即使类别较少，朴素贝叶斯依然能够有效地分类，因为其核心思想是基于每个类别的条件概率进行分类。类别数量的多少主要影响模型的计算复杂度，但不会直接导致性能下降。
- 在类别数量少的情况下，朴素贝叶斯依然能够快速并且有效地进行训练和预测。
- **结论**：**类别数量较少不会对模型性能产生严重影响**

解答

- D. 特征是离散的。

- **解释**：朴素贝叶斯特别适用于离散特征，尤其在文本分类任务中，单词通常被视为离散特征。离散特征使得条件概率估计更加直接和有效。
- 对于离散特征，朴素贝叶斯能够直接计算每个特征值的条件概率，而不需要进行复杂的分布估计（如连续特征的概率密度函数）。因此，离散特征实际上非常适合朴素贝叶斯。

- **结论**：离散特征不会对朴素贝叶斯的性能造成严重影响，反而是它的优势之一。

题目

- 判断1.
- 贝叶斯分类器属于生成式模型，而逻辑回归属于判别式模型。

解答

- **正确**

贝叶斯分类器通过学习数据的联合概率分布 P(X,Y)，然后利用贝叶斯定理**计算后验概率** P(Y｜X)，因此属于 **生成式模型**。

而决策树、SVM等直接学习条件概率分布 P(Y｜X)P(Y|X)P(Y｜X)，属于 **判别式模型**。

题目

- 判断2.
- 朴素贝叶斯分类器不适用于高维数据集，因为计算复杂度太高。

解答

- 错误

- 实际上，朴素贝叶斯分类器在高维数据集上表现良好。由于条件独立性假设，计算联合概率时不需要考虑特征之间的组合，计算复杂度较低，适合处理高维数据。

题目

- 判断3.
- 在朴素贝叶斯分类器中，连续型特征必须离散化才能处理。

解答

- 错误

- 朴素贝叶斯分类器可以通过假设连续型特征服从某种概率分布（如高斯分布）来处理连续型特征，而不一定需要将其离散化。

题目

- 1. 已知在一个邮件分类问题中，有垃圾邮件（Spam）和正常邮件（Ham）两类。在训练集中，垃圾邮件占40%，正常邮件占60%。现在有一封新邮件，包含关键词“优惠”（Offer）。已知在垃圾邮件中，20%的邮件包含“优惠”，在正常邮件中，只有5%的邮件包含“优惠”。问这封邮件是垃圾邮件的概率是多少?

解答

- 这个问题可以通过**贝叶斯定理**来解决。贝叶斯定理给出了如何根据已知条件计算某一事件发生的概率。
- 在这个问题中，我们需要计算邮件是垃圾邮件（Spam）给定关键词“优惠”（Offer）出现的条件概率，即 $P(Spam \mid Offer)$。

解答

贝叶斯定理：

$$P(\text{Spam}|\text{Offer}) = \frac{P(\text{Offer}|\text{Spam}) \cdot P(\text{Spam})}{P(\text{Offer})}$$

其中：

- $P(\text{Spam}|\text{Offer})$ 是在邮件包含关键词"优惠"时，邮件为垃圾邮件的概率（这是我们要求的结果）。
- $P(\text{Offer}|\text{Spam})$ 是垃圾邮件中包含"优惠"关键词的概率（给定垃圾邮件，包含"优惠"的概率）。
- $P(\text{Spam})$ 是邮件是垃圾邮件的先验概率。
- $P(\text{Offer})$ 是邮件包含"优惠"关键词的总体概率。

解答

步骤：

1. **已知条件**：
   - $P(\text{Spam}) = 0.40$（垃圾邮件的先验概率，40%）
   - $P(\text{Ham}) = 0.60$（正常邮件的先验概率，60%）
   - $P(\text{Offer}|\text{Spam}) = 0.20$（垃圾邮件中包含“优惠”关键词的条件概率，20%）
   - $P(\text{Offer}|\text{Ham}) = 0.05$（正常邮件中包含“优惠”关键词的条件概率，5%）

解答

2. **计算** $P(\text{Offer})$： $P(\text{Offer})$ 是邮件中包含"优惠"关键词的总体概率，可以通过全概率公式计算：

$$P(\text{Offer}) = P(\text{Offer}|\text{Spam}) \cdot P(\text{Spam}) + P(\text{Offer}|\text{Ham}) \cdot P(\text{Ham})$$

代入已知的值：

$$P(\text{Offer}) = (0.20 \cdot 0.40) + (0.05 \cdot 0.60)$$

$$P(\text{Offer}) = 0.08 + 0.03 = 0.11$$

解答

3. **计算** $P(\text{Spam}|\text{Offer})$： 现在，我们可以使用贝叶斯定理计算在邮件包含“优惠”时，它是垃圾邮件的概率：

$$P(\text{Spam}|\text{Offer}) = \frac{P(\text{Offer}|\text{Spam}) \cdot P(\text{Spam})}{P(\text{Offer})}$$

代入已知的值：

$$P(\text{Spam}|\text{Offer}) = \frac{0.20 \cdot 0.40}{0.11}$$

$$P(\text{Spam}|\text{Offer}) = \frac{0.08}{0.11} \approx 0.727$$

结果：

这封邮件是垃圾邮件的概率 $P(\text{Spam}|\text{Offer})$ 大约为 **0.727**，即 **72.7%**。

题目

- 2. 在某个分类问题中，有三个类别 C1,C2,C3，其先验概率分别为 P(C1)=0.3,P(C2)=0.4,P(C3)=0.3。给定一个特征 XXX，其在各类别下的条件概率为：

$P(X \mid C1)=0.5$

$P(X \mid C2)=0.2$

$P(X \mid C3)=0.4$

根据贝叶斯分类器，判断该样本最可能属于哪个类别。

解答

- 这个问题要求我们使用**贝叶斯定理**来判断给定特征 XXX 最可能属于哪个类别。我们可以通过计算每个类别的后验概率，选择具有最高后验概率的类别作为预测结果。

解答

贝叶斯定理

贝叶斯定理给出的公式是：

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

其中：

- $P(C_i|X)$ 是在给定特征 X 后，样本属于类别 C_i 的后验概率。
- $P(X|C_i)$ 是在类别 C_i 下，特征 X 的条件概率。
- $P(C_i)$ 是类别 C_i 的先验概率。
- $P(X)$ 是特征 X 的总概率（对于所有类别的条件概率的加权和）。

由于我们关心的是类别 C_1、C_2、和 C_3 中哪个类别的后验概率最大，且 $P(X)$ 对所有类别来说是相同的，我们可以忽略 $P(X)$ 进行比较，只需要计算：

$$P(C_i|X) \propto P(X|C_i)P(C_i)$$

即对于每个类别 C_i，计算 $P(X|C_i) \cdot P(C_i)$，然后比较结果。

解答

已知条件：

- $P(C_1) = 0.3$
- $P(C_2) = 0.4$
- $P(C_3) = 0.3$
- $P(X|C_1) = 0.5$
- $P(X|C_2) = 0.2$
- $P(X|C_3) = 0.4$

解答

计算后验概率：

1. **类别 C_1 的后验概率：**

$$P(C_1|X) \propto P(X|C_1)P(C_1) = 0.5 \times 0.3 = 0.15$$

2. **类别 C_2 的后验概率：**

$$P(C_2|X) \propto P(X|C_2)P(C_2) = 0.2 \times 0.4 = 0.08$$

3. **类别 C_3 的后验概率：**

$$P(C_3|X) \propto P(X|C_3)P(C_3) = 0.4 \times 0.3 = 0.12$$

解答

比较后验概率：

- $P(C_1|X) = 0.15$
- $P(C_2|X) = 0.08$
- $P(C_3|X) = 0.12$

结论：

从计算结果来看，$P(C_1|X)$ 的值最大，因此根据贝叶斯分类器，该样本最可能属于类别 C_1。

题目

- **3. 在一个朴素贝叶斯分类器中，有两个连续特征 X1 和 X2，以及两个类别 Y={0,1} 。已知：**
- P(Y=1)=0.5，P(Y=0)=0.5
- 当 Y=1时，X1服从均值为5，方差为1的高斯分布；X2 服从均值为5，方差为1的高斯分布。
- 当 Y=0时，X1服从均值为0，方差为1的高斯分布；X2 服从均值为0，方差为1的高斯分布。
- **给定一个样本 (X1=4,X2=4)，计算其属于类别 Y=1的后验概率。**

解答

要计算给定样本 $(X_1 = 4, X_2 = 4)$ 属于类别 $Y = 1$ 的后验概率，我们可以使用**贝叶斯定理**，并根据特征是连续的高斯分布来计算每个类别的似然概率。

贝叶斯定理：

贝叶斯定理给出后验概率的计算公式：

$$P(Y = 1|X_1 = 4, X_2 = 4) = \frac{P(X_1 = 4, X_2 = 4|Y = 1)P(Y = 1)}{P(X_1 = 4, X_2 = 4)}$$

我们需要计算分子 $P(X_1 = 4, X_2 = 4|Y = 1)P(Y = 1)$，并且可以忽略分母部分 $P(X_1 = 4, X_2 = 4)$ 因为它对所有类别都是相同的。最终，我们通过比较不同类别的后验概率来决定样本属于哪个类别。

解答

1. 计算 $P(X_1 = 4, X_2 = 4|Y = 1)$

在 $Y = 1$ 时，特征 X_1 和 X_2 都服从均值为 5，方差为 1 的高斯分布。因此，给定 $Y = 1$，X_1 和 X_2 的条件概率可以通过高斯分布的概率密度函数 (PDF) 计算。

高斯分布的概率密度函数为：

$$P(X|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

对于 X_1 和 X_2 在 $Y = 1$ 的情况下，均值 $\mu = 5$，方差 $\sigma^2 = 1$，所以：

$$P(X_1 = 4|Y = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-5)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

$$P(X_2 = 4|Y = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4-5)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

解答

因为 X_1 和 X_2 是独立的，我们可以将它们的条件概率相乘：

$$P(X_1 = 4, X_2 = 4|Y = 1) = P(X_1 = 4|Y = 1) \cdot P(X_2 = 4|Y = 1)$$

代入上述公式：

$$P(X_1 = 4, X_2 = 4|Y = 1) = \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)\right) \cdot \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)\right)$$

$$P(X_1 = 4, X_2 = 4|Y = 1) = \frac{1}{2\pi} \exp(-1)$$

解答

2. 计算 $P(Y = 1)$

已知 $P(Y = 1) = 0.5$，这是类别 $Y = 1$ 的先验概率。

解答

3. 计算 $P(Y=0|X_1=4, X_2=4)$

类似地，我们也可以计算类别 $Y=0$ 时的似然概率 $P(X_1=4, X_2=4|Y=0)$。在 $Y=0$ 时，X_1 和 X_2 都服从均值为 0，方差为 1 的高斯分布。

$$P(X_1=4|Y=0)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(4-0)^2}{2}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{16}{2}\right)=\frac{1}{\sqrt{2\pi}}\exp(-8)$$

$$P(X_2=4|Y=0)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(4-0)^2}{2}\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{16}{2}\right)=\frac{1}{\sqrt{2\pi}}\exp(-8)$$

同样由于 X_1 和 X_2 独立：

$$P(X_1=4, X_2=4|Y=0)=\left(\frac{1}{\sqrt{2\pi}}\exp(-8)\right)\cdot\left(\frac{1}{\sqrt{2\pi}}\exp(-8)\right)$$

$$P(X_1=4, X_2=4|Y=0)=\frac{1}{2\pi}\exp(-16)$$

解答

4. 计算后验概率

我们现在可以使用贝叶斯定理计算后验概率：

$$P(Y=1|X_1=4,X_2=4) \propto P(X_1=4,X_2=4|Y=1)P(Y=1)$$

$$P(Y=1|X_1=4,X_2=4) \propto \frac{1}{2\pi}\exp(-1)\cdot 0.5 = \frac{1}{2\pi}\exp(-1)\cdot 0.5$$

$$P(Y=0|X_1=4,X_2=4) \propto P(X_1=4,X_2=4|Y=0)P(Y=0)$$

$$P(Y=0|X_1=4,X_2=4) \propto \frac{1}{2\pi}\exp(-16)\cdot 0.5 = \frac{1}{2\pi}\exp(-16)\cdot 0.5$$

解答

比较后验概率

现在，我们可以比较 $P(Y=1|X_1=4,X_2=4)$ 和 $P(Y=0|X_1=4,X_2=4)$ 的大小。我们不需要计算 $1/2\pi$ 因为它对两个后验概率是相同的。比较两者的主要因素是指数项：

- $\exp(-1)$ 和 $\exp(-16)$ 显然，$\exp(-1)>\exp(-16)$，因此：

$$P(Y=1|X_1=4,X_2=4)>P(Y=0|X_1=4,X_2=4)$$

结果：

因此，该样本更可能属于类别 $Y=1$。

题目

- 4.现有一位候选人，其特征为“本科”、“有编程经验”、“项目数量多”，使用朴素贝叶斯分类器计算该候选人通过面试的概率。

学历	编程经验	项目数量	是否通过面试
硕士	有	多	是
硕士	有	少	是
本科	有	多	是
本科	无	少	否
本科	无	少	否
大专	无	少	否

解答

1. 计算先验概率 $P(是)$ 和 $P(否)$

从数据表中：

- $P(是) = \frac{通过面试样本数}{总样本数} = \frac{3}{6} = 0.5$
- $P(否) = \frac{未通过面试样本数}{总样本数} = \frac{3}{6} = 0.5$

解答

2. 条件概率计算

特征 1：学历

- $P(\text{学历}=\text{本科}|\text{是}) = \frac{\text{通过面试中学历为本科的样本数}}{\text{通过面试样本数}} = \frac{1}{3} = 0.333$
- $P(\text{学历}=\text{本科}|\text{否}) = \frac{\text{未通过面试中学历为本科的样本数}}{\text{未通过面试样本数}} = \frac{2}{3} = 0.667$

特征 2：编程经验

- $P(\text{编程经验}=\text{有}|\text{是}) = \frac{\text{通过面试中编程经验为有的样本数}}{\text{通过面试样本数}} = \frac{3}{3} = 1.0$
- $P(\text{编程经验}=\text{有}|\text{否}) = \frac{\text{未通过面试中编程经验为有的样本数}}{\text{未通过面试样本数}} = \frac{0}{3} = 0.0$

特征 3：项目数量

- $P(\text{项目数量}=\text{多}|\text{是}) = \frac{\text{通过面试中项目数量为多的样本数}}{\text{通过面试样本数}} = \frac{2}{3} = 0.667$
- $P(\text{项目数量}=\text{多}|\text{否}) = \frac{\text{未通过面试中项目数量为多的样本数}}{\text{未通过面试样本数}} = \frac{0}{3} = 0.0$

解答

3. 使用贝叶斯公式计算后验概率

贝叶斯公式：

$$P(是|数据) = \frac{P(是) \cdot P(学历{=}本科|是) \cdot P(编程经验{=}有|是) \cdot P(项目数量{=}多|是)}{P(数据)}$$

其中：

$$P(数据) = P(是|数据) + P(否|数据)$$

计算 $P(是|数据)$

$$P(是|数据) \propto P(是) \cdot P(学历{=}本科|是) \cdot P(编程经验{=}有|是) \cdot P(项目数量{=}多|是)$$

$$P(是|数据) \propto 0.5 \cdot 0.333 \cdot 1.0 \cdot 0.667 = 0.111$$

计算 $P(否|数据)$

$$P(否|数据) \propto P(否) \cdot P(学历{=}本科|否) \cdot P(编程经验{=}有|否) \cdot P(项目数量{=}多|否)$$

$$P(否|数据) \propto 0.5 \cdot 0.667 \cdot 0.0 \cdot 0.0 = 0$$

归一化后计算最终概率

因为 $P(否 \mid 数据) = 0$，所以：

$$P(数据) = P(是|数据) + P(否|数据) = 0.111 + 0 = 0.111$$

$$P(是|数据) = \frac{0.111}{0.111} = 1$$

解答

4. 结论

候选人通过面试的概率为 1，未通过的概率为 0，因此判断该候选人**肯定会通过面试**。

End