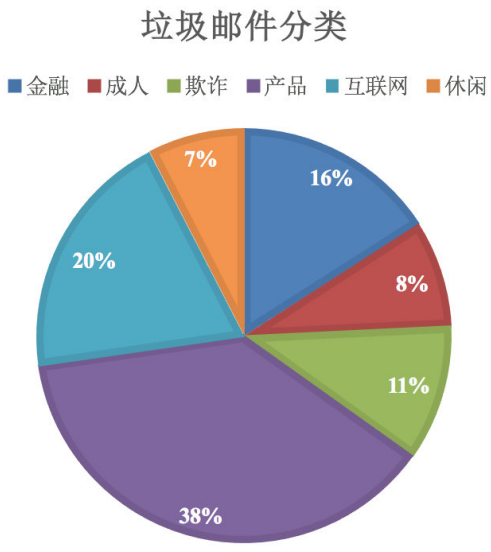
朴素贝叶斯

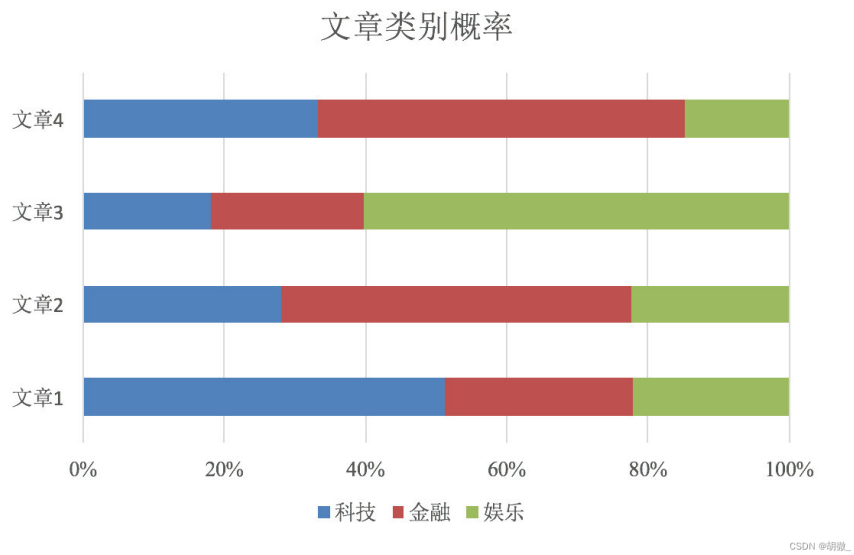
朴素贝叶斯算法是一类基于贝叶斯定理的分类算法。它**基于特征之间的条件独立性假设**，这是一个“朴素”的假设，因此得名为朴素贝叶斯。尽管这个独立性的假设在实际问题中并不总是成立，但朴素贝叶斯在实际应用中表现出令人满意的性能。

朴素贝叶斯算法的基本思想可以通过以下步骤概括：

1. **贝叶斯定理：** 朴素贝叶斯算法基于贝叶斯定理，该定理描述了在给定先验信息的情况下，如何更新对未知事件的概率估计。
2. **条件独立性假设：** 朴素贝叶斯算法假设特征之间是条件独立的，即给定类别的情况下，特征之间不存在相互影响。
3. **计算概率：** 对于给定的类别，计算每个特征在该类别下的条件概率。这涉及到计算类别的先验概率和每个特征在各个类别下的条件概率。
4. **分类：** 对于新的输入样本，基于贝叶斯定理计算其属于每个类别的后验概率，然后选择具有最高概率的类别作为样本的预测类别。

朴素贝叶斯是一种分类算法，经常被用于文本分类，它的输出结果是某个样本属于某个类别的概率。





贝叶斯公式

概率基础复习

- 联合概率：包含多个条件，且所有条件同时成立的概率
  - 记作：`P(A,B)`
- 条件概率：就是事件A在另外一个事件B已经发生条件下的发生概率
  - 记作：`P(A|B)`
- 相互独立：如果`P(A, B) = P(A)P(B)`，则称事件A与事件B相互独立

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)}$$

注：w为给定文档的特征值(频数统计,预测文档提供)，c为文档类别

下面通过一个案例“ 判断女神对你的喜欢情况 ”来理解这个公式

样本数	职业	体型	女神是否喜欢
1	程序员	超重	不喜欢
2	产品	匀称	喜欢
3	程序员	匀称	喜欢
4	程序员	超重	喜欢
5	美工	匀称	不喜欢
6	美工	超重	不喜欢
7	产品	匀称	喜欢

问题如下：

女神喜欢的概率？
职业是程序员并且体型匀称的概率？
在女神喜欢的条件下，职业是程序员的概率？
在女神喜欢的条件下，职业是程序员、体重超重的概率？

计算结果为：

P(喜欢) = 4/7
P(程序员, 匀称) = 1/7(联合概率)
P(程序员|喜欢) = 2/4 = 1/2(条件概率)
P(程序员, 超重|喜欢) = 1/4

思考题：

- 在小明是产品经理并且体重超重的情况下，如何计算小明被女神喜欢的概率？即`P(喜欢|产品, 超重) = ？`

P(喜欢|产品, 超重) = P(产品, 超重|喜欢)P(喜欢)/P(产品, 超重)

计算上式可以发现：

P(产品, 超重|喜欢) 和 P(产品, 超重) 的结果均为0，导致无法计算结果。这是因为我们的样本量太少了，不具有代表性。
本来现实生活中，肯定是存在职业是产品经理并且体重超重的人的，P(产品, 超重)不可能为0；而且事件 职业是产品经理 和事件 体重超重 通常被认为是相互独立的事件，但是，根据我们有限的7个样本计算 P(产品, 超重) = P(产品)P(超重) 不成立。
而朴素贝叶斯可以帮助我们解决这个问题：

朴素贝叶斯，简单理解，就是假定了特征与特征之间相互独立的贝叶斯公式。
也就是说，朴素贝叶斯，之所以朴素，就在于假定了特征与特征相互独立。

所以，思考题如果按照朴素贝叶斯的思路来解决，就可以是

P(产品, 超重) = P(产品) * P(超重) = 2/7 * 3/7 = 6/49
p(产品, 超重|喜欢) = P(产品|喜欢) * P(超重|喜欢) = 1/2 * 1/4 = 1/8
P(喜欢|产品, 超重) = P(产品, 超重|喜欢)P(喜欢)/P(产品, 超重) = 1/8 * 4/7 / 6/49 = 7/12

拉普拉斯平滑系数

贝叶斯公式如果应用在**文章分类**的场景当中，我们可以这样看：

公式分为三个部分：

- P(C)：每个文档类别的概率(某文档类别数 / 总文档数量)
- P(W | C)：给定类别下特征（被预测文档中出现的词）的概率
  - 计算方法：P(F1 | C)=Ni/N （训练文档中去计算）
    - Ni为该F1词在C类别所有文档中出现的次数
    - N为所属类别C下的文档所有词出现的次数和
- P(F1,F2,…) 预测文档中每个词的概率

如果计算两个类别概率比较：

所以我们只要比较前面的大小就可以，得出谁的概率大

下面通过一个案例进行理解

需求：通过前四个训练样本（文章），判断第五篇文章，是否属于China类

	文档ID	文档中的词	属于c=China类
训练集	1	Chinese Beijing Chinese	Yes
	2	Chinese Chinese Shanghai	Yes
	3	Chinese Macao	Yes
	4	Tokyo Japan Chinese	No
测试集	5	Chinese Chinese Chinese Tokyo Japan	?


```
P(C|Chinese, Chinese, Chinese, Tokyo, Japan)
= P(Chinese, Chinese, Chinese, Tokyo, Japan|C) * P(C) / P(Chinese, Chinese, Chinese, Tokyo, Japan)
= P(Chinese|C)^3 * P(Tokyo|C) * P(Japan|C) * P(C) / [P(Chinese)^3 * P(Tokyo) * P(Japan)]

# 这个文章是需要计算是不是China类，是或者不是最后的分母值都相同:

# 首先计算是China类的概率:
P(Chinese|C) = 5/8
P(Tokyo|C) = 0/8
P(Japan|C) = 0/8

# 接着计算不是China类的概率:
P(Chinese|C) = 1/3
P(Tokyo|C) = 1/3
P(Japan|C) = 1/3
```

问题：从上面的例子我们可以得到 `P(Tokyo|C)` 和 `P(Japan|C)` 都为0，这是不合理的，如果词频列表里面有很多次数都为0，很可能计算结果都为0。

解决办法：拉普拉斯平滑系数

$$P(F1|C) = \frac{Ni+\alpha}{N+\alpha m}$$

α为指定的系数一般为1，m为训练文档中统计出的特征词个数

```
# 这个文章是需要计算是不是China类:
# 该例中，m=6（训练集中特征词的个数，重复不计）

首先计算是China类的概率:
    P(Chinese|C) = 5/8 --> 6/14
    P(Tokyo|C) = 0/8 --> 1/14
    P(Japan|C) = 0/8 --> 1/14

接着计算不是China类的概率:
    P(Chinese|C) = 1/3 --> 2/9
    P(Tokyo|C) = 1/3 --> 2/9
    P(Japan|C) = 1/3 --> 2/9
```

用sklearn使用朴素贝叶斯

```
sklearn.naive_bayes.MultinomialNB(alpha = 1.0)
朴素贝叶斯分类
alpha：拉普拉斯平滑系数
```

朴素贝叶斯优缺点

（1）优点

朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率
对缺失数据不太敏感，算法也比较简单，常用于文本分类
分类准确度高，速度快

（2）缺点

由于使用了样本属性独立性的假设，所以如果特征属性有关联时其效果不好
需要计算先验概率，而先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳；

什么是条件概率，边缘概率和联合概率

1. **条件概率（Conditional Probability）：** 条件概率表示在某个事件已经发生的条件下，另一个事件发生的概率。给定事件 B，事件 A 的条件概率表示为 P(A|B)，读作"在 B 发生的情况下 A 发生的概率"。
   具体计算方式为：
   $P(A|B) = \frac{P(A \cap B)}{P(B)}$
   其中，$P(A \cap B)$ 是 A 和 B 同时发生的概率，而 $P(B)$ 是事件 B 发生的概率。
2. **边缘概率（Marginal Probability）：** 边缘概率是指在某个事件的条件下，另一个事件的概率。通常，它是通过对联合概率进行边缘化（Marginalization）得到的。对于事件 A 和 B，边缘概率可以表示为：
   $P(A) = \sum_i P(A \cap B_i)$
   或者
   $P(B) = \sum_i P(A_i \cap B)$
   其中，B_i 是事件 B 的一个可能的取值，A_i 是事件 A 的一个可能的取值。
3. **联合概率（Joint Probability）：** 联合概率表示多个事件同时发生的概率。对于事件 A 和 B，联合概率表示为 $P(A \cap B)$。
   联合概率和条件概率之间的关系可以通过以下等式表示：
   $P(A \cap B) = P(A|B) \cdot P(B)$

这些概率概念在概率统计、机器学习等领域广泛应用，特别是在贝叶斯统计和概率图模型中。

朴素贝叶斯模型如何学习的，训练过程是怎么样的

朴素贝叶斯模型的学习过程主要包括训练和预测两个阶段。在训练阶段，模型学习先验概率和条件概率；在预测阶段，模型基于学到的参数进行分类。

以下是朴素贝叶斯模型的训练过程：

1. **准备数据：** 收集并准备带有标签的训练数据。每个样本都有一组特征和一个标签，特征表示样本的属性，标签表示样本所属的类别。
2. **计算类别的先验概率：** 对于每个类别，计算训练集中该类别的样本数量占总样本数量的比例。这个比例就是类别的先验概率。
   $P(C_i) = \frac{\text{类别}\,i\,\text{的样本数量}}{\text{总样本数量}}$
3. **计算特征在各类别下的条件概率：** 对于每个特征，计算在每个类别下该特征的条件概率。这涉及到计算每个类别中具有该特征的样本数量占该类别总样本数量的比例。
   $P(F_j|C_i) = \frac{\text{类别}\,i\,\text{中具有特征}\,j\,\text{的样本数量}}{\text{类别}\,i\,\text{的样本总数量}}$
4. **存储参数：** 将计算得到的先验概率和条件概率存储起来，作为模型的参数。

训练完成后，模型就可以用于预测新样本的类别。预测过程如下：

1. **计算后验概率：** 对于给定的新样本，根据贝叶斯定理计算其属于每个类别的后验概率。
   $P(C_i|\text{新样本}) = P(C_i) \cdot \prod_{j=1}^{n} P(F_j|C_i)$
   其中，n 是特征的数量，F_j 表示第 j 个特征。
2. **选择最可能的类别：** 根据后验概率选择具有最高概率的类别作为新样本的预测类别。

朴素贝叶斯模型的简单性和高效性使得它在文本分类、垃圾邮件过滤等应用中广泛使用。

朴素贝叶斯法将实例分到后验概率最大的类中。后验概率最大化这等价于期望风险最小化。

如何理解生成模型和判别模型

生成模型（Generative Model）和判别模型（Discriminative Model）是机器学习中两种不同类型的模型，它们的主要区别在于模型对数据的处理方式和学习目标。

1. **生成模型（Generative Model）：**
   - **定义：** **生成模型试图学习数据的生成过程，即模型尝试建模观察数据与潜在变量之间的联合分布。**
   - **目标：** 生成模型的目标是学习数据的整体分布，以便能够生成与训练数据类似的新样本。

- **应用：** **生成模型常用于生成新的样本**，如生成图像、文本、音频等。它们也可以用于密度估计、缺失数据填充等任务。

例子： 高斯混合模型（Gaussian Mixture Model，GMM）、变分自编码器（Variational Autoencoder，VAE）。

2. **判别模型（Discriminative Model）：**
   - **定义：** 判别模型关注的是对不同类别之间的决策边界的建模，即模型试图学习输入与标签之间的条件分布。
   - **目标：** 判别模型的目标是**学习类别之间的决策边界，以便能够对新输入进行分类**。
   - **应用：** 判别模型广泛用于分类、回归等任务，其中主要关注的是对输入数据进行标签预测。

   例子： 逻辑回归、支持向量机（Support Vector Machine，SVM）、深度学习中的大多数神经网络。

对比和理解：

- **生成模型关注整体分布：** 生成模型对整个数据分布进行建模，从而能够生成新的数据样本。它们通常对数据的隐含结构有更全面的认识。
- **判别模型关注决策边界：** 判别模型关注对不同类别之间的决策边界进行建模，以便进行分类。它们更直接地关注输入与输出之间的映射。

朴素贝叶斯模型朴素在哪里？存在什么问题？有哪些优化方向？

"朴素"在朴素贝叶斯模型中指的是对特征之间的条件独立性的假设。具体来说，朴素贝叶斯**假设给定类别的情况下，各个特征之间是相互独立的**。这是一个朴素的假设，因为在实际问题中，很多情况下特征之间并不是完全独立的。尽管这个独立性的假设在实际问题中不总是成立，朴素贝叶斯在实际应用中表现出令人满意的性能，尤其在文本分类等领域。

存在的问题和优化方向如下：

存在的问题：

独立性假设问题： 朴素贝叶斯模型假设特征之间是条件独立的，这在某些实际问题中可能不成立，导致模型精度下降。

对缺失数据敏感： 朴素贝叶斯对缺失数据敏感，如果某个特征在训练数据中未出现过，会导致概率估计为零，影响模型性能。

优化方向：

处理独立性假设问题： 有一些改进的朴素贝叶斯模型尝试放宽对特征独立性的假设，如半朴素贝叶斯模型。这些模型尝试通过考虑一定程度上的特征相关性来改善模型的性能。

处理缺失数据： 可以采用各种方法来处理缺失数据，如使用插值方法进行填充，或者使用更复杂的模型来更好地处理缺失数据。

使用平滑技术： 为了避免在估计概率时出现零概率的问题，可以使用平滑技术，如拉普拉斯平滑（Laplace smoothing）或Lidstone平滑，以确保所有特征值都有非零的概率。

结合其他模型： 可以考虑结合朴素贝叶斯模型与其他模型，形成混合模型，以平衡不同模型的优缺点，提高整体性能。

尽管朴素贝叶斯模型有其局限性，但在许多实际问题中，它仍然是一个简单而有效的分类算法，尤其在处理文本分类等应用中常取得不错的结果。