第 5 篇

统计机器学习方法是如何实现分类与聚类的

艾博士导读

统计机器学习方法在人工智能发展历史上曾经起到过重要作用，当 20 世纪 90 年代初期人工智能陷入低谷时，也正是统计机器学习的发展才使得人工智能走出低谷，逐渐得到广泛的应用。当前的人工智能发展高潮应该与统计机器学习方法的发展紧密相关，即便在今天，统计机器学习方法也有着广泛的应用。

统计机器学习的最大特点是具有良好的理论基础，也可能正因为如此，本篇内容具有较多的公式，也打破了我在写作本书时定下的尽可能少用公式的禁忌，不过大家在学习本篇内容时，不要被大量出现的公式所吓倒，我会尽可能说明每个公式的含义及来龙去脉，即便你没有弄清楚具体的推导过程，也可以了解其中的思想。

本篇主要介绍统计机器学习方法中几个典型的监督与非监督学习方法，按照难易程度可以划分为 3 个不同的等级，读者可以根据自身需要有选择性地阅读其中几节或者全部内容。

第一级：5.1 节，介绍统计机器学习的基本概念。5.3 节到 5.3.1 小节之前的部分，介绍决策树的基本概念。5.4 节，介绍 k 近邻分类方法。5.5 节中的 5.5.1 小节，介绍支持向量机的基本概念。5.6 节开始部分，介绍聚类问题的基本概念。5.9 节，介绍统计机器学习中的验证与测试方法，以及分类问题的评价指标。

第二级：5.2 节，介绍朴素贝叶斯分类方法。5.3 节开始部分以及其中的 5.3.1 小节，介绍构建决策树的 ID3 方法。5.5 节中的 5.5.2 小节，介绍线性可分支持向量机。5.6 节介绍 k 均值聚类方法。5.7 节介绍层次聚类方法。

第三级：5.3 节中的 5.3.2～5.3.4 小节，介绍构建决策树的 C4.5 方法；介绍建造决策树过程中可能出现的过拟合问题，以及解决过拟合问题的剪枝方法；介绍什么是随机森林以及构建方法。5.5 节中的 5.5.3～5.5.6 小节，介绍线性支持向量机以及非线性支持向量机；介绍求解非线性支持向量机的核方法；介绍如何用二分类支持向量机求解多分类问题。5.8 节介绍基于密度的聚类方法 DBSCAN 算法。5.10 节结合文本分类问题和脱机手写汉字识别问题，分别介绍两种抽取特征的方法。

这天正在学习人工智能的小明，看到了这样一个题目。如表 5.1 所示，给出的是一个男女性别样本数据表。该表共有称作样本的 15 组数据，每组数据对应一个样本，每个样本有“年龄”“发长”“鞋跟”“服装”4 种特征，最后一列给出了是“男性”或者“女性”的性别分类。

其中年龄划分为老年、中年和青年，发长划分为短发、中发和长发，鞋跟划分为平底和高跟，服装划分为深色、浅色和花色，这些称为特征的取值。比如第一组数据，表示的是“一位老年人，留有短发，穿着平底鞋，身穿深色服装，其性别为男性”。而第三组数据，表示的是“一位老年人，头发中长，穿着平底鞋，身穿花色服装，其性别为女性”。题目要求根据这些样本数据建立一个人工智能系统，当任意输入一个人的“年龄”“发长”“鞋跟”“服装”这 4 个特征的取值后，即便是表 5.1 中不存在的样本，系统也可以判断出该人的类别，即是男性还是女性。

表 5.1　男女性别样本数据表

ID	年龄	发长	鞋跟	服装	性别
1	老年	短发	平底	深色	男性
2	老年	短发	平底	浅色	男性
3	老年	中发	平底	花色	女性
4	老年	长发	高跟	浅色	女性
5	老年	短发	平底	深色	男性
6	中年	短发	平底	浅色	男性
7	中年	短发	平底	浅色	男性
8	中年	长发	高跟	花色	女性
9	中年	中发	高跟	深色	女性
10	中年	中发	平底	深色	男性
11	青年	长发	高跟	浅色	女性
12	青年	短发	平底	浅色	女性
13	青年	长发	平底	深色	男性
14	青年	短发	平底	花色	男性
15	青年	中发	高跟	深色	女性

小明第一次遇到这样的题目，思考了一会儿后也没有想到什么好的求解方法，又来请教艾博士。

5.1　统计学习方法

了解了小明的来意之后，艾博士讲解起来：这个问题属于统计学习方法研究的范畴，我们先简单介绍一下什么是统计学习方法。

统计学习方法又称作统计机器学习方法，属于机器学习的一种。

小明：什么是机器学习呢？

艾博士：我们人之所以能做很多事情，重要的就是具有学习能力。我们从小到大一直在学习，通过学习提高我们做事情的能力。计算机也是一样，我们也希望计算机能像人一样，拥有学习能力，一旦拥有了这种学习能力，计算机就可以帮助人类做更多的事情。这也是人工智能所追求的目标。

著名学者司马贺(赫伯特・西蒙)教授曾经对机器学习给出过一个定义:“如果一个系统能够通过执行某个过程改进它的性能,这就是学习”。

小明:这和我们人类的学习也差不多,我们学习不就是提升自我做事的能力吗?

艾博士:统计机器学习就是计算机系统通过运用数据及统计方法提高系统性能的机器学习,其特点是运用统计方法,从数据出发提取数据的特征,抽象出问题的模型,发现数据中所隐含的知识,最终用得到的模型对新的数据进行分析和预测。

统计机器学习一般具有两个过程:一个是学习,又称作训练,是从数据抽象模型的过程;另一个是使用,用学习到的模型对数据进行分析和预测。为了实现第一个过程,一般需要一个供学习用的数据集,又称作训练集,由训练样本组成的集合,这是学习、训练的依据。

像小明刚刚提到的这个题目,表 5.1 给出的就是数据集,数据集中的每一个样本由若干特征和类别标签组成,其中的“年龄”“发长”“鞋跟”“服装”就是特征,而性别就是类别标签。依据这个数据集采用某个统计学习方法建立一个男女性别分类模型,当任意给定一个人的“年龄”“发长”“鞋跟”和“服装”特征时,模型输出该人的性别。

当然这里只是给出了一个例子,对于实际问题来说,这个数据集太小了,需要更多的数据,特征数目也不够多,取值也需要再细化。

小明:统计机器学习都有哪些方法呢?

艾博士:统计机器学习具有很多种方法,从是否有类别标签的角度,可以划分为以下几种。

1. 有监督学习

艾博士:有监督学习(见图 5.1)又称作监督学习、有教师学习,也就是说给定数据集中的样本具有类别标签。这就好比是小孩认识动物一样,看到了一只猫,妈妈告诉小孩这是一只猫;看到了一只狗,妈妈又告诉说这是一只狗。慢慢地小孩就学会了认识猫和狗。

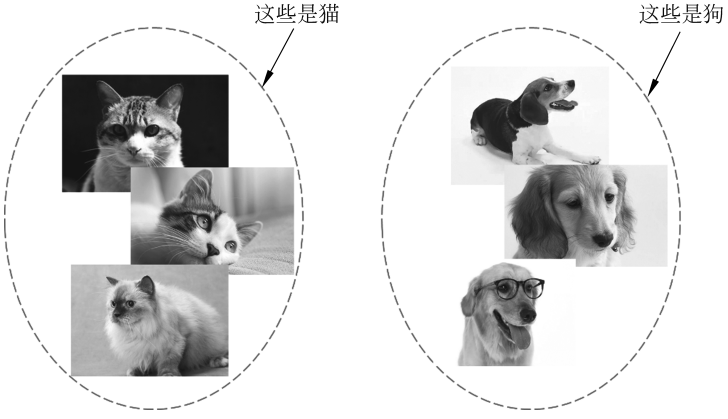

图 5.1　有监督学习示意图

小明:这里的“监督”指的就是类别标签吗?

艾博士肯定地说:是的,监督指的就是类别标签信息。这类任务为的是让人工智能系统学会认识某个事物属于哪个类别,也就是根据特征划分到指定类别,一般称作为分类。

2. 无监督学习

艾博士：无监督学习(见图 5.2)又称作无教师学习，与监督学习刚好相反，给定的数据集中的样本只有特征没有类别标签。比如假设一个人从没有看到过狗和猫，给他一些猫和狗的照片，他虽然不认识哪个是猫哪个是狗，但是该人观看了一会儿照片后，根据两种动物的特点，他可以区分出这是两种不同的动物，进而可以将这些照片划分为两类：一类是狗，另一类是猫，虽然他并不知道每一类是什么动物。

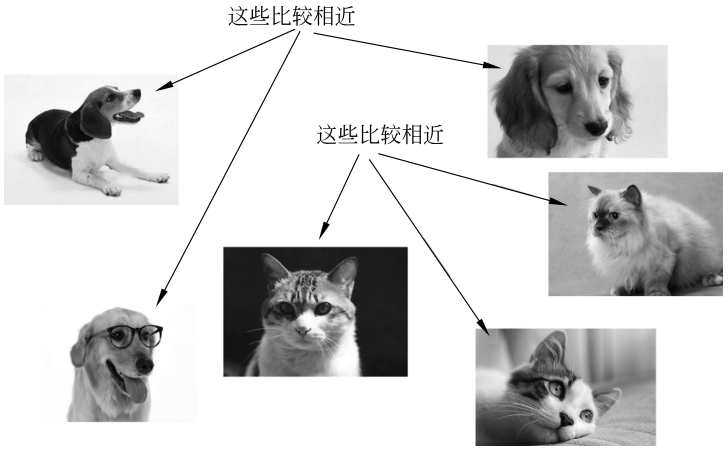

图 5.2　无监督学习

小明：由于没有标签信息，这类任务只能做到把类似的东西归纳为一个类别吧？

艾博士：确实如此，这类任务就是将特征比较接近的东西聚集为一类，一般称作聚类。

3. 半监督学习

艾博士：顾名思义，半监督学习(见图 5.3)就是数据集中部分样本有标签信息，部分样本没有标签信息。半监督学习就是如何利用这些无标签数据，提高学习系统的性能。比如在一些猫和狗的照片中，一部分照片标注是猫或者是狗，但是也有一部分照片没有任何类别标注。

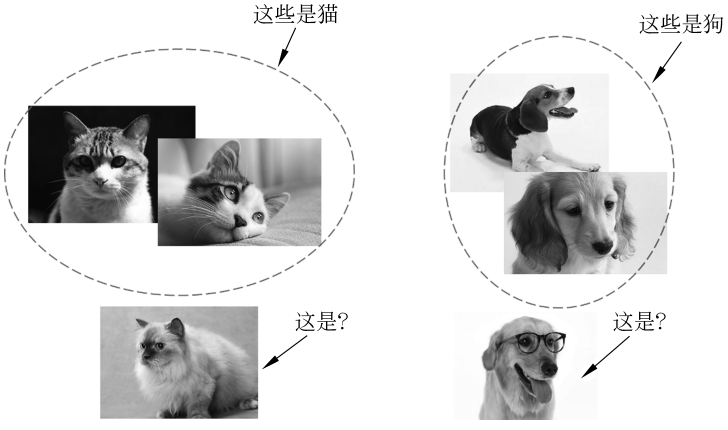

图 5.3　半监督学习示意图

小明：如何利用无标签样本呢？

艾博士：一般来说，半监督学习中大部分样本还是有标签的，利用有标签样本可以大概预测出那些无标签样本的类别，利用预测结果可以进一步优化系统的分类性能。当然预测结果会存在一定的错误，这是半监督学习要解决的问题。

4. 弱监督学习

艾博士：弱监督学习指的是提供的学习样本中标签信息比较弱，这又可以分为几种情况。第一种是不完全监督学习（见图 5.4），其特点是标签信息不充分，只有少量样本具有类别标签，而大部分样本没有标签信息。

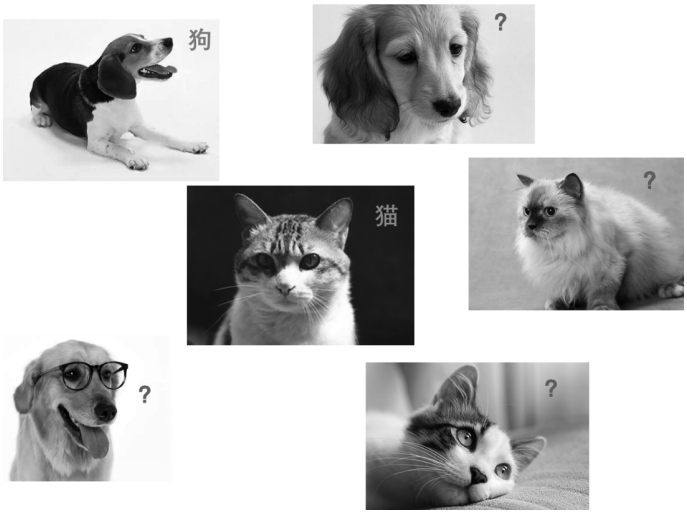

图 5.4　弱监督学习——不完全监督学习

小明：这与半监督学习有什么区别呢？

艾博士：严格来说半监督学习可以归类到这类弱监督学习中，都属于不完全监督学习。但是一般情况下，半监督学习带标签样本会更多一些，而弱监督学习中的带标签样本会更少。

艾博士：第二种弱监督学习是不确切监督学习（见图 5.5），其特点是具有类别标签信息，但是标注对象不明确，只给了一个粗粒度的标注。比如一张遛狗的照片，照片中有狗，也有人，也有其他的东西，标签只说明照片中有狗，但是没有明确地指明具体哪个是狗。

小明：感觉这类学习难度就更大了，因为虽然具有标签信息，但是属于粗粒度的标注，学习过程中需要明确具体的标注对象。

艾博士：是的，增加了不少学习难度。这类学习可以把样本想象成一个包，标签信息只说明了包内有什么，而没有说明包内的具体所指。

艾博士：还有一类弱监督学习就是强化学习（见图 5.6）。在强化学习中没有明确的数据告诉计算机学习什么，但是可以设置奖惩函数，当结果正确时获得奖励，而结果错误时遭受惩罚，通过不断试错的方法获得数据，从而进行学习。

小明：在第 2 篇中讲过的下围棋的 AlphaGo 就用到了强化学习吧？

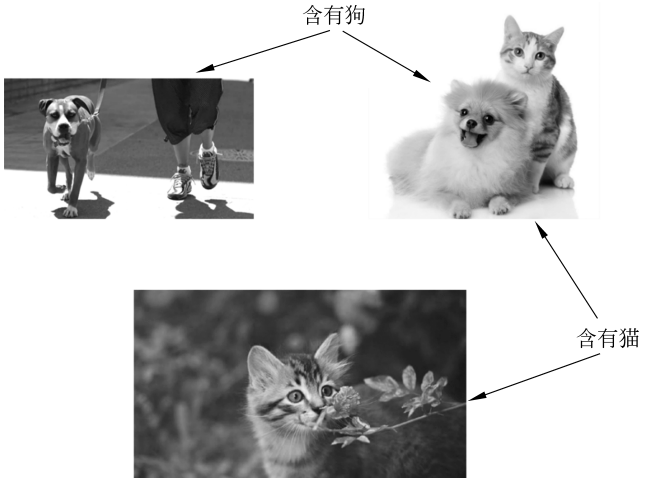

图 5.5　弱监督学习——不确切监督学习

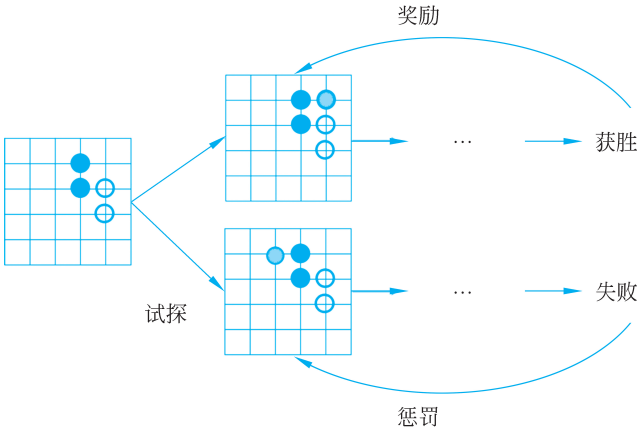

图 5.6　弱监督学习——强化学习

艾博士： AlphaGo 中用到了强化学习，而 AlphaGo Zero 则摆脱了人类数据，完全依靠强化学习达到了人类棋手所不能达到的下棋水平。

除此之外，还有不精确监督学习也属于弱监督学习，其特点是标签信息存在错误标注，比如将个别狗的照片标记成了猫，或者将个别猫的照片标记成了狗。一般来说当数据集大了以后都不可避免地会存在一些标注错误，有些机器学习方法对少量标注错误并不敏感，有些方法可能比较敏感，即便存在少量错误标注的样本，也可能会带来比较大的问题，这就涉及如何剔除这些错误标注样本的问题。

以上从样本标签的角度对机器学习方法做了分类，每类还有不同的机器学习方法。下面几节中，我们将介绍其中几个典型的监督和非监督统计机器学习方法。

小明读书笔记

统计机器学习属于机器学习的一种，其特点是运用数学统计方法，抽象出问题的模型，发现数据中蕴含的内在规律，用得到的模型实现对新数据的分析和预测。

统计机器学习分为两个过程：一个是训练过程，从数据抽象模型的过程；另一个是使用过程，用学习到的模型对数据进行分析和预测。为了实现训练过程，需要一个数据集，称作训练集，它是由训练用样本组成的集合，这是训练的依据。

按照是否有标注信息以及标注信息的多少，统计机器学习可以划分为有监督学习、无监督学习、半监督学习和弱监督学习等。

5.2　朴素贝叶斯方法

艾博士：朴素贝叶斯方法是一种基于概率的分类方法，其基本思想是：对于一个以若干特征表示的待分类样本，依次计算样本属于每个类别的概率，其中所属概率最大的类别作为分类结果输出。

为了叙述方便，我们给出如下符号表示：设共有 K 个类别，分别用 $y_1,y_2,\cdots,y_K$ 表示。每个样本具有 N 个特征，分别为 $A_1,A_2,\cdots,A_N$，每个特征 A_i 又有 S_i 个可能的取值，分别为 $a_{i1},a_{i2},\cdots,a_{iS_i}$。

小明：这些看起来有些抽象，您结合例子具体说明一下吧。

艾博士：好的，我们还是以前面说过的男女性别分类的例子加以说明。在该例子中，共有男性和女性两个类别，所以类别数 K 为 2，我们可以用 y_1 表示男性，用 y_2 表示女性。每个样本有“年龄”“发长”“鞋跟”和“服装”共 4 种特征，可以用 A_1 表示“年龄”，A_2 表示“发长”，A_3 表示“鞋跟”，A_4 表示“服装”。年龄特征 A_1 可以有“老年”“中年”“青年”3 种取值，则特征 A_1 的取值个数 S_1 为 3，分别可以用 a_{11} 表示老年，用 a_{12} 表示中年，用 a_{13} 表示青年。同样地，发长特征 A_2 可以有“长发”“中发”“短发”3 种取值，则特征 A_2 的取值个数 S_2 为 3，分别可以用 a_{21} 表示长发，用 a_{22} 表示中发，用 a_{23} 表示短发。以此类推，对于特征“鞋跟”和“服装”也可以用类似的表示方法进行表示，我们就不一一说明了。

小明：有了这几个例子就清楚各个符号的具体含义了。

艾博士：对于待分类样本我们用 x 表示：

$$x=(x_1,x_2,\cdots,x_N)$$

其中，x_i 为待分类样本的第 i 个特征 A_i 的取值。

比如：$x=$(青年，中发，平底，花色)，表示的是一个年龄特征为青年，发长特征为中发，鞋跟特征为平底，服装特征为花色的样本。

我们的目的就是计算给定的待分类样本 x 属于某个类别 y_i 的概率 $P(y_i|x)$，然后将 x 分类到概率值最大的类别中。

小明：这个概率怎么计算呢？

艾博士：一般来说这个概率并不是太好计算，需要转换一下。小明你还记得我们在第 3 篇求解拼音输入法问题时提到过的贝叶斯公式吗？

小明回想了一会儿回答说：我印象中贝叶斯公式是这样的：

$$P(B\mid A)=\frac{P(A\mid B)P(B)}{P(A)} \tag{5.1}$$

艾博士：对，就是这个贝叶斯公式。我们看看这个贝叶斯公式是否可以帮助到我们。

假设待分类样本的出现表示事件 A，而属于类别 y_i 表示事件 B，则根据贝叶斯公式我们有：

$$P(y_i \mid x)=\frac{P(x \mid y_i)P(y_i)}{P(x)} \tag{5.2}$$

式(5.2)中，$P(y_i)$表示类别 y_i 出现的概率，$P(x)$表示 x 出现的概率，$P(x \mid y_i)$表示在类别 y_i 中出现特征取值为 $x=(x_1,x_2,\cdots,x_N)$的概率。

我们的目的就是通过贝叶斯公式，计算待分类样本 x 在每个类别中的概率，然后以取得最大概率的类别作为分类结果。

由于待分类样本是给定的，所以对于这个问题来说，$P(x)$是固定的，所以求概率 $P(y_i \mid x)$最大与求 $P(x \mid y_i)P(y_i)$最大是等价的。因为我们并不关心属于哪个类别的概率具体是多少，而只关心属于哪个类别的概率最大。

因此，问题转换为如何计算下式在哪个类别 y_i 下最大问题：

$$P(x \mid y_i)P(y_i) \tag{5.3}$$

这是两个概率的乘积，如果分别可以计算出两个概率值 $P(x \mid y_i)$、$P(y_i)$，那么这个问题也就解决了。这样的分类方法称为贝叶斯方法。

讲到这里艾博士询问小明：你觉得这两个概率值怎么计算呢？

小明： 概率一般都是根据数据统计计算。如果有了训练集，通过训练集是否就可以计算出这两个概率？

我计算一下试试看。比如还是前面的男女性别分类的例子，为了计算方便，我们将表 5.1 复制过来，如表 5.2 所示(方便读者阅读)。

表 5.2 男女性别样本数据表

ID	年龄	发长	鞋跟	服装	性别
1	老年	短发	平底	深色	男性
2	老年	短发	平底	浅色	男性
3	老年	中发	平底	花色	女性
4	老年	长发	高跟	浅色	女性
5	老年	短发	平底	深色	男性
6	中年	短发	平底	浅色	男性
7	中年	短发	平底	浅色	男性
8	中年	长发	高跟	花色	女性
9	中年	中发	高跟	深色	女性
10	中年	中发	平底	深色	男性
11	青年	长发	高跟	浅色	女性
12	青年	短发	平底	浅色	女性
13	青年	长发	平底	深色	男性
14	青年	短发	平底	花色	男性
15	青年	中发	高跟	深色	女性

我觉得类别概率 $P(y_i)$ 比较容易计算，属于类别 y_i 的样本数除以总样本数就可以了，即

$$P(y_i)=\frac{\text{属于类别 } y_i \text{ 的样本数}}{\text{总样本数}} \tag{5.4}$$

表 5.2 中共 15 个样本，其中 8 个类别为男性，7 个类别为女性。所以有：

$$P(y_1)=P(\text{男性})=\frac{8}{15}=0.5333$$

$$P(y_2)=P(\text{女性})=\frac{7}{15}=0.4667$$

概率 $P(x|y_i)$ 体现的是类别 y_i 中具有 x 特征的概率，与具体的待分类样本有关，前面给的待分类样本的例子 $x=$（青年，中发，平底，花色）。我发现数据集中没有一个这样的样本，所以按照该数据集计算的话，得到的概率为 0。这就出现问题了，因为无论是男性类别还是女性类别，式(5.3)的计算结果都为 0，无法判断属于哪个类别的概率更大。是不是训练集数据量太少了？

艾博士：对于我们这个例题来说，数据集确实有点小，但是小明你提到的问题本质上并不是数据集大小的问题，而是组合爆炸问题。

小明：为什么会有组合爆炸问题呢？

艾博士：小明你看，一个样本由多个特征组成，而每个特征又有多个取值，这样每个特征的每一个可能取值都会组成一个样本，再考虑不同的类别，都需要计算其概率值，其总数是每个特征取值数的乘积再乘以类别数，当类别数、特征数和特征的取值数比较多时，就出现了组合爆炸问题。以这个例题为例，特征包含了年龄、发长、鞋跟和服装 4 种特征，而年龄、发长和服装 3 个特征均有 3 个取值，鞋跟特征有两个取值，类别分为男性和女性两个类别，这样可能的组合数就是 $3\times3\times3\times2\times2=108$ 种。由于这个例题中特征数、特征的取值数和类别数都比较小，组合爆炸问题还不太明显，如果类别数、特征数和特征可能的取值数比较多时，需要估计的概率值将会呈指数增加，从而造成组合爆炸。这样，需要非常多的样本才有可能比较准确地估计每种情况下的概率值，而对于实际问题来说，很难做到如此全面地采集数据。

小明：那么如何解决这个问题呢？

艾博士：为解决这个问题，我们可以假设各特征间是独立的。在独立性假设下，特征每个取值的概率可以单独估计，不存在组合问题，也就消除了组合爆炸问题。在这样的假设下，给定类别 y_i 时某个特征组合的联合概率等于该类别下各个特征单独取值概率的乘积，即

$$P(x\mid y_i)=P((x_1,x_2,\cdots,x_N)\mid y_i)=\prod_{j=1}^{N}P(x_j\mid y_i)$$

其中，$P(x_j|y_i)$ 是类别为 y_i 时，第 j 个特征 A_j 取值为 x_j 的概率；N 为特征个数。

在引入独立性假设后，式(5.3)可以写为：

$$P(x\mid y_i)P(y_i)=\prod_{j=1}^{N}P(x_j\mid y_i)\cdot P(y_i)$$

$$= P(y_i)\prod_{j=1}^{N} P(x_j \mid y_i) \tag{5.5}$$

这样分类问题就变成了求式(5.5)最大时所对应的类别问题。这种引入了独立性假设后的贝叶斯分类方法称作朴素贝叶斯方法。

小明：这样处理会带来哪些好处呢？

艾博士：这样对于特征每个取值的概率就可以单独计算了，不需要考虑与其他特征的组合情况，减少了对训练集数据量的需求，计算起来更加简单。下面我们给出具体的计算方法。

在给定类别 y_i 的情况下，特征 A_k 取值为 a_{kj} 的概率 $P(a_{kj} \mid y_i)$ 可以通过训练集计算得到：

$$P(a_{kj} \mid y_i) = \frac{\text{类别 } y_i \text{ 的样本中特征 } A_k \text{ 值为 } a_{kj} \text{ 的样本数}}{\text{标记为类别 } y_i \text{ 的样本数}} \tag{5.6}$$

回到我们的例题，由于 $x=$(青年，中发，平底，花色)，就是要分别计算以下几个概率的乘积：

$$P(\text{青年} \mid y_i)P(\text{中发} \mid y_i)P(\text{平底} \mid y_i)P(\text{花色} \mid y_i)$$

小明你再试试，看是否可以根据数据集计算出这几个概率来？

小明对照表 5.2 所示的数据认真计算起来。

当类别 y_i 为男性时共有 8 个样本，其中 2 个样本年龄为青年，所以有：

$$P(\text{青年} \mid \text{男性}) = \frac{2}{8} = 0.25$$

其中 1 个样本发长为中发，所以有：

$$P(\text{中发} \mid \text{男性}) = \frac{1}{8} = 0.125$$

其中 8 个样本鞋跟全部为平底，所以有：

$$P(\text{平底} \mid \text{男性}) = \frac{8}{8} = 1$$

其中 1 个样本服装为花色，所以有：

$$P(\text{花色} \mid \text{男性}) = \frac{1}{8} = 0.125$$

再加上我们前面已经计算过的：

$$P(\text{男性}) = \frac{8}{15} = 0.5333$$

以上结果代入式(5.3)中，有：

$$\begin{aligned} P(x \mid \text{男性})P(\text{男性}) &= P(\text{青年} \mid \text{男性})P(\text{中发} \mid \text{男性})P(\text{平底} \mid \text{男性}) \\ &\quad P(\text{花色} \mid \text{男性})P(\text{男性}) \\ &= 0.25 \times 0.125 \times 1 \times 0.125 \times 0.5333 \\ &= 0.002083 \end{aligned} \tag{5.7}$$

当类别 y_i 为女性时共有 7 个样本，其中 3 个样本年龄为青年，所以有：

$$P(\text{青年} \mid \text{女性}) = \frac{3}{7} = 0.429$$