第 5 章 统计机器学习方法

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，正是统计机器学习的发展才使得人工智能走出了低谷，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，虽然热潮来自深度学习。即便在今天，统计机器学习方法也有广泛的应用。

5.1 什么是统计机器学习方法

人之所以能做很多事情，重要的是具有学习能力。我们从小到大一直在学习，通过学习提高做事情的能力。计算机也一样，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："如果一个系统能通过执行某个过程改进它的性能，这就是学习。"

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。一个过程是学习，又称为训练，是从数据抽象模型的过程。另一个过程是使用，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，又称为训练集，由训练样本组成的集合，是学习、训练的依据。

表 5.1 给出了一个数据集，数据集中的每个样本由若干特征和类别标签组成，其中的"年龄""发长""鞋跟"和"服装"就是特征，而性别是类别标签。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。这就是统计学习方法所要解决的问题。

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

续表

ID	年龄	发长	鞋跟	服装	性别
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

当然,这里只是给出一个例子,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。

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

有监督学习又称为监督学习、有教师学习,也就是说给定数据集中的样本具有类别标签,如图5.1所示。这就好比小孩认识动物一样,看到了一只猫,妈妈告诉小孩这是一只猫,看到了一只狗,妈妈又告诉小孩这是一只狗,慢慢地小孩就认识了猫和狗。

监督指的就是类别标签信息。这类任务的目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,一般称为分类。

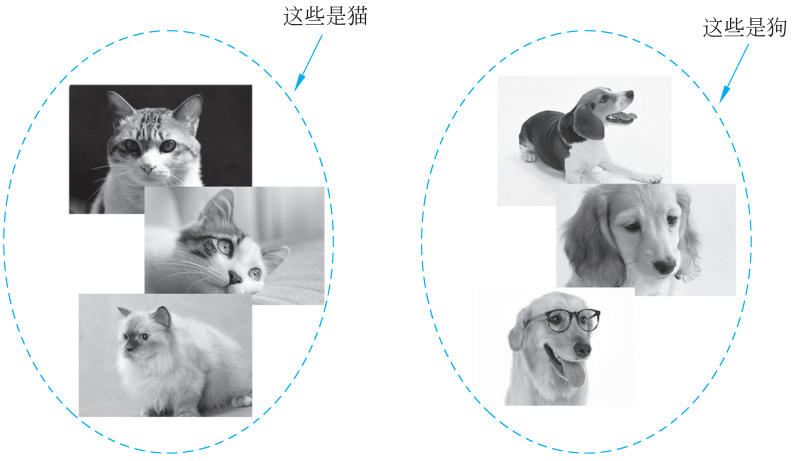

图5.1 有监督学习示意图

2. 无监督学习

无监督学习又称为无教师学习,与有监督学习刚好相反,给定的数据集中的样本只有特

征没有类别标签，如图 5.2 所示。例如，假设一个人从没看到过狗和猫，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，虽然他并不知道每类是什么动物。

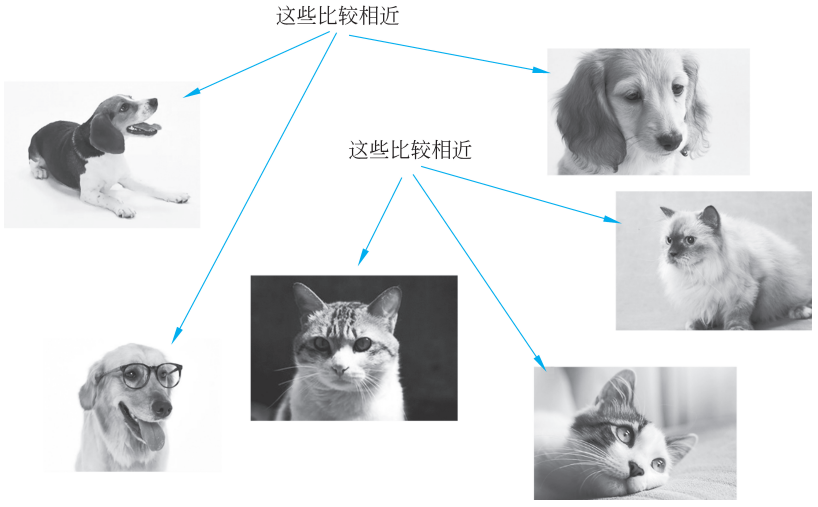

图 5.2　无监督学习示意图

由于没有标签信息，这类任务就是将特征比较接近的东西聚集为一类，一般称为聚类。

3. 半监督学习

顾名思义，半监督学习就是数据集中有部分样本有标签信息，部分样本没有标签信息，如图 5.3 所示。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。例如，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。

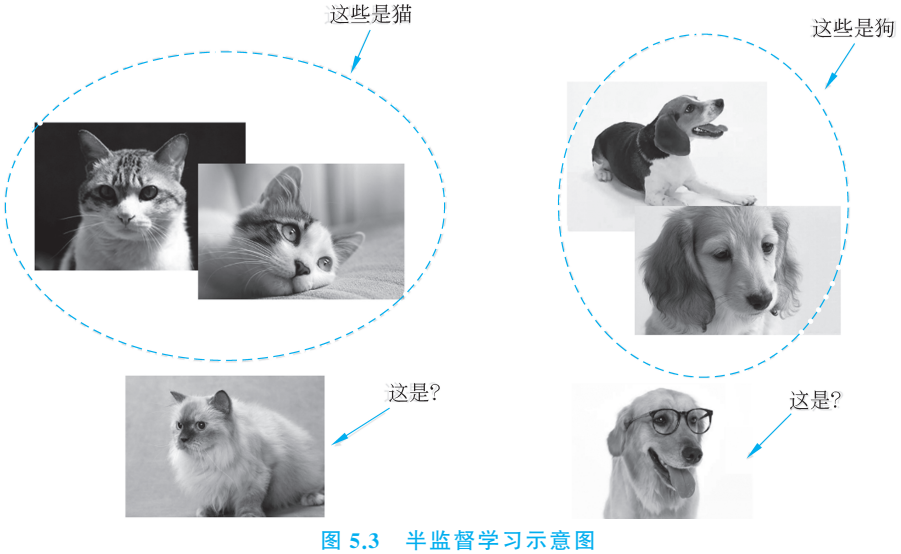

图 5.3　半监督学习示意图

一般来说，半监督学习中大部分样本是有标签的，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

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。

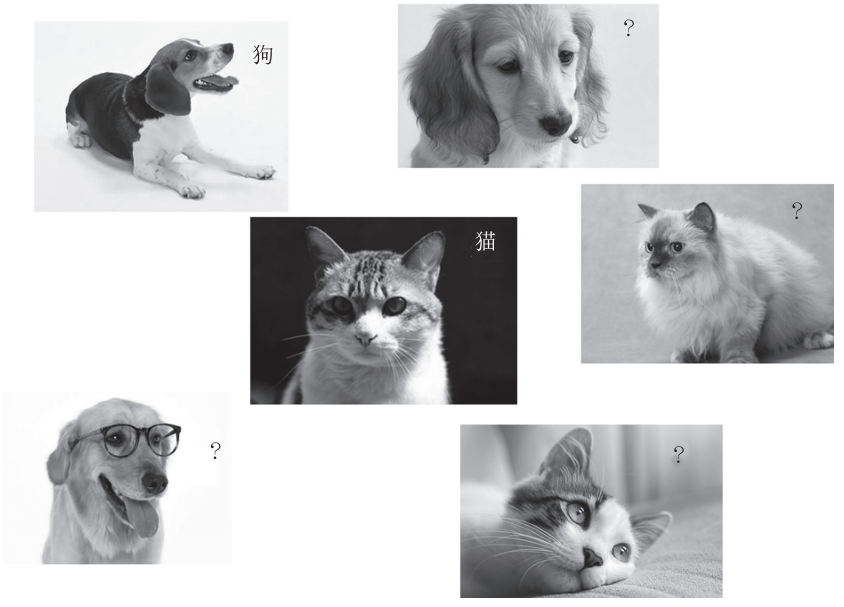

图5.4 弱监督学习——不完全监督学习

严格来说,半监督学习也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。

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,还有其他的东西,标签只说明照片中有狗,但是没有明确指明具体哪个是狗。

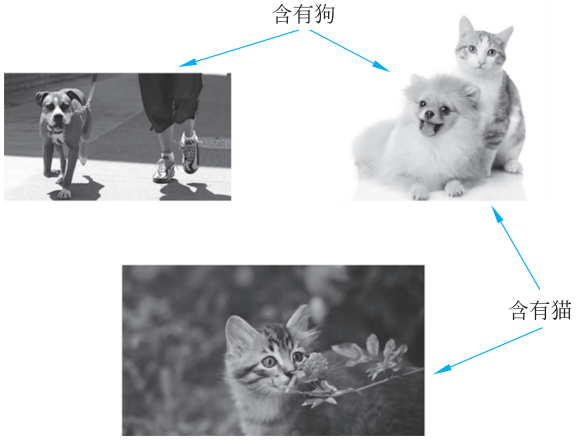

图5.5 弱监督学习——不确切监督学习

这类学习任务难度更大，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，增加了学习难度。这类学习可以把样本想成一个包，标签信息只说明了包内有什么，而没有说明包内具体所指。

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。

下围棋的 AlphaGo 就用到强化学习，而 AlphaGo Zero 则摆脱了人类数据，完全依靠强化学习达到人类棋手所不能达到的下棋水平。

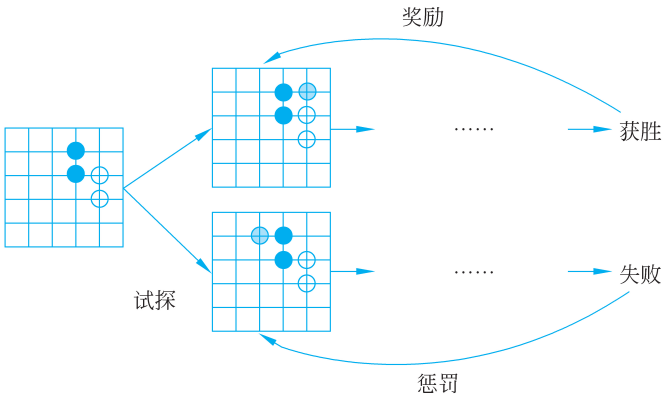

图 5.6　弱监督学习——强化学习

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，数据集大了以后不可避免地存在一些标注错误，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。下面几节介绍其中几个典型的监督和非监督统计机器学习方法。

5.2　朴素贝叶斯方法

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，我们给出如下的符号表示：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 $a_{i1},a_{i2},\cdots,a_{iS_i}$。

下面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，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”4 种特征，可以用 A_1 表示“年龄”，用 A_2 表示“发长”，用 A_3 表示“鞋跟”，用 A_4 表示“服装”。年龄特征 A_1 可以有“老年”“中年”和“青年”3 种取值，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,发长特征 A_2 可以有“长发”“中发”和“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表示,这里就不一一说明了。

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。

例如,$x=$(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。

一般来说,这个概率并不是太容易计算,需要转换一下。根据贝叶斯公式:

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

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,则根据贝叶斯公式有:

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

其中,$P(y_i)$表示类别 y_i 出现的概率,$P(x)$表示 x 出现的概率,$P(x|y_i)$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 $P(y_i|x)$ 最大与求 $P(x|y_i)P(y_i)$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式(5.3)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 $P(x|y_i)$、$P(y_i)$,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。

其中的概率一般都根据数据统计计算。如果有了训练集,通过训练集就可以计算出这两个概率。

对于男、女性别分类的例子,假设有如表5.1所示的数据集。

依据数据集,用属于类别 y_i 的样本数除以总样本数计算出类别概率 $P(y_i)$:

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

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),由于数据集中没有这样的样本,所以按照该数据集计算,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。

对于这个例题来说,数据集中的样本太少,出现 0 概率在所难免,但是本质上并不是数据集大小的问题,而是组合爆炸问题。

为什么会有组合爆炸问题呢?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,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,类别数、特征数和特征的取值数比较多时,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。

为解决这个问题,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

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

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

引入独立性假设后,式(5.3)可以写为

$$\begin{aligned}P(x\mid y_i)P(y_i)&=\prod_{j=1}^{N}P(x_j\mid y_i)\cdot P(y_i)\\&=P(y_i)\prod_{j=1}^{N}P(x_j\mid y_i)\end{aligned}\tag{5.5}$$

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。这种引入独立性假设后的贝叶斯分类方法称作朴素贝叶斯方法。

由于引入了独立性假设,对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,计算起来也更加简单。下面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:

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

这样,即便是在表5.2这样的小数据集的情况下,也可以求出这几个概率值,而不会出现概率为0的情况。这就是引入独立性假设带来的好处。

下面依据表5.2的数据计算一下这几个概率。

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:

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

将以上结果代入式(5.3)中,有

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

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

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

其中3个样本发长为中发,所以有

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

其中2个样本鞋跟为平底,所以有

$$P(\text{平底} \mid \text{女性}) = \frac{2}{7} = 0.286$$

其中2个样本服装为花色,所以有

$$P(\text{花色} \mid \text{女性}) = \frac{2}{7} = 0.286$$

再加上我们前面已经计算过的:

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

将以上结果代入式(5.3)中,有

$$\begin{aligned} P(x \mid \text{女性})P(\text{女性}) = {} & P(\text{青年} \mid \text{女性})P(\text{中发} \mid \text{女性})P(\text{平底} \mid \text{女性})P(\text{花色} \mid \text{女性})P(\text{女性}) \\ = {} & 0.429 \times 0.429 \times 0.286 \times 0.286 \times 0.4667 \\ = {} & 0.007030 \end{aligned} \tag{5.8}$$

式(5.8)的计算结果大于式(5.7)的计算结果，说明待分类样本 $x=$(青年，中发，平底，花色)属于女性的概率大于属于男性的概率，所以应该被分类为女性。

引入独立性假设后问题确实简单了不少，但是实际问题中特征之间一般具有一定的相关性，并不完全满足独立性假设。比如年龄特征和鞋跟特征，对于老年人来说，由于行走不方便，自然穿高跟鞋的就少，二者是有一定相关性的。但是如果不引入独立性假设，参数量也就是需要估计的概率值太多，很难有足够的数据集支持这些参数的估计。所以，引入独立性假设也是不得已采用的一种简化手段，以便于真正将这种方法用于解决实际问题，而且朴素贝叶斯方法也确实在解决实际问题中取得了很好的效果。

实际使用朴素贝叶斯方法进行分类时，一般会根据训练数据集事先计算好所有的概率值，存储起来，这个过程属于训练过程。在具体分类时直接调用所需要的概率值就可以了，这个过程属于分类过程。

另外，由于概率值一般都比较小，式(5.5)是多个概率值的连乘运算，当特征比较多时，连乘运算的结果会变得越乘越小，可能出现计算结果下溢的情况，即当运算结果小于计算机所能表示的最小值之后，就被当作 0 处理了。为此，一般通过取对数的方式将连乘运算转换为累加运算，即用式(5.9)代替式(5.5)，二者取得最大值的类别 y_i 是一样的，不影响分类结果。

$$\begin{aligned}\log(P(x\mid y_i)P(y_i)) &= \log\Big(P(y_i)\prod_{j=1}^{N}P(x_j\mid y_i)\Big)\\ &= \log(P(y_i))+\sum_{j=1}^{N}\log(P(x_j\mid y_i))\end{aligned} \tag{5.9}$$

即便引入特征的独立性假设后，当用式(5.6)计算概率值时，也不能完全排除概率为 0 的情况出现。比如对于表 5.2 所示的数据集，当类别为男性时鞋跟特征取值为高跟的数据一个也没有，这样就会导致概率 P(高跟|男性)为 0 的情况出现。为此，可以采用拉普拉斯平滑方法避免概率为 0 的情况发生。

拉普拉斯平滑方法的基本思想是，假定每种情况至少出现一次，而无论数据集中是否出现过。也就是说，在用式(5.6)计算概率 $P(a_{kj}\mid y_i)$时，对分子中的“类别 y_i 的样本中特征 A_k 值为 a_{kj} 的样本数”进行计数时，采用在原有计数的基础上再加 1 的方法，防止出现 0 的情况。对于具有 S_k 个取值的特征 A_k 来说，在类别 y_i 下其所有取值的概率和应该为 1，即

$$\sum_{j=1}^{S_k}P(a_{kj}\mid y_i)=1 \tag{5.10}$$

为此，式(5.6)的分母应该相应地增加 S_k 以满足概率和为 1 这一条件。这样，采用拉普拉斯平滑后，式(5.6)就变成了式(5.11)：

$$P(a_{kj}\mid y_i)=\frac{\text{类别 } y_i \text{ 的样本中特征 } A_k \text{ 值为 } a_{kj} \text{ 的样本数}+1}{\text{标记为类别 } y_i \text{ 的样本数}+\text{特征 } A_k \text{ 可能的取值数 } S_k} \tag{5.11}$$

这样就避免了出现概率等于 0 的情况。

因为特征 A_k 具有 S_k 个取值，计数时每个取值的样本数都增加了一个，相当于多了 S_k 个样本，这样，式(5.11)的分母中就需要加上 S_k。这样处理后才能满足式(5.10)概率和为 1 的条件。

对于类别概率，也采用类似的办法，假定每个类别至少存在一个样本，这样类别概率计

算式(5.4)就变成了式(5.12):

$$P(y_i)=\frac{\text{属于类别 } y_i \text{ 的样本数}+1}{\text{总样本数}+\text{类别数 } K} \tag{5.12}$$

与式(5.11)的道理相同,由于每个类别增加了一个样本数,共有 K 个类别,相当于增加了 K 个样本,所以分母中要加上类别数 K,以便满足每个类别的概率累加和为1的条件。

采用拉普拉斯平滑方法后的概率,按照表5.2给出的数据集,我们计算一下两个类别概率和在不同类别下发长特征几个取值的概率,计算结果如下。

表5.2中共有15个样本,男性和女性两个类别,其中男性有8个样本,女性有7个样本。按照式(5.12)计算得到类别概率:

$$P(\text{男性})=\frac{8+1}{15+2}=0.5294$$

$$P(\text{女性})=\frac{7+1}{15+2}=0.4706$$

同样,对于发长特征共有短发、中发和长发3个取值,在8个男性类别样本中有6个短发样本、1个中发样本和1个长发样本。按照式(5.11)计算得到概率:

$$P(\text{短发}\mid\text{男性})=\frac{6+1}{8+3}=0.6364$$

$$P(\text{中发}\mid\text{男性})=\frac{1+1}{8+3}=0.1818$$

$$P(\text{长发}\mid\text{男性})=\frac{1+1}{8+3}=0.1818$$

同样,对于发长特征,在7个女性类别样本中有1个短发样本、3个中发样本和3个长发样本。按照式(5.11)计算得到概率:

$$P(\text{短发}\mid\text{女性})=\frac{1+1}{7+3}=0.2$$

$$P(\text{中发}\mid\text{女性})=\frac{3+1}{7+3}=0.4$$

$$P(\text{长发}\mid\text{女性})=\frac{3+1}{7+3}=0.4$$

拉普拉斯平滑方法通过在原有计数基础上加1的方法,解决了因数据不足造成的概率为0问题,看起来是一个小技巧,实际上是有理论依据的,具体就不介绍了。

最后再举一个采用朴素贝叶斯方法做文本分类任务的例子。

所谓文本分类任务,就是对于一个给定文本,按照其内容分配到相应的类别中。比如有4个新闻类别分别为体育、财经、政治和军事,新来了一份新闻稿件,应该属于哪个类别呢?这就是文本分类任务所要完成的任务。

为了完成这个任务,首先要收集包含这4方面内容的新闻稿件作为训练数据集,我们称之为语料库。语料库中每篇新闻稿件作为一个训练样本。收集到的每篇新闻稿件要标注好所属的文本类别,以便用于计算朴素贝叶斯分类方法中所用到的各种概率。为了防止出现概率等于0的情况,我们采用拉普拉斯平滑方法。

首先按照式(5.12)计算4个类别的类别概率,以新闻稿件为单位进行计算: