第 5 章

分　　类

分类知识反映同类事物的共同特征和不同事物的差异特征。分类(classification)是人类认识世界的一种重要方法,对于事物的认识大多是通过分门别类进行的。在数据挖掘领域,分类是从一个已知类别的数据集到一组预先定义的、非交叠的类别的映射过程。其中映射关系的生成以及映射关系的应用是数据挖掘分类算法主要的研究内容。映射关系就是常说的分类器(也称分类模型或分类函数),映射关系的应用就是使用分类器将测试数据集中的数据划分到给定类别中的某个类别的过程。

分类从历史的特征数据中构造出特定对象的分类模型(分类器),用来对未来数据进行预测分析,属于数据挖掘中的预测任务挖掘。分类技术使用的历史训练样本数据有确定的类标号,所以分类属于机器学习中的有监督学习。分类技术具有非常广泛的应用领域,如医疗诊断、信用卡系统的信用分级、图像模式识别、网络数据分类等。机器学习、专家系统、统计学和神经网络等领域的研究人员已经提出了许多具体的分类方法。目前比较常用的分类方法有 K 近邻(KNN)分类、贝叶斯分类、决策树和神经网络等。本章介绍分类的基本概念和基本的分类算法,第 6 章将介绍一些高级的分类算法。

5.1　基 本 概 念

分类的目的是得到一个分类器(也称作分类模型),通过得到的分类器能把测试集中的测试数据映射到给定类别中的某个类别,实现对该数据的预测性描述。分类可用于提取描述重要数据类的模型或预测未来的数据趋势。本节首先介绍分类的概念,然后介绍分类的过程、分类器常见的构造方法以及分类器的评价标准。

5.1.1　什么是分类

对于餐饮企业而言,数据分析极为重要,如搞清楚不同时节菜品的历史销售情况,分析不同因素影响下顾客的增加、流失情况等。数据分析的一项任务就是分类。分类方法用于预测数据对象的离散类别。如顾客对菜品种类 A、B、C 的喜好,贷款申请的“安全”或“危险”。这些类别可以用离散值表示,这些值之间没有次序。例如,可以使用值 1、2、3 表示菜品种类 A、B、C,这 3 个菜品种类之间并不存在次序。下面介绍与分类相关的几个重要概念。

1. 数据对象和属性

数据集由数据对象组成。数据集中的数据对象代表一个实体。

属性也称字段，表示数据对象的一个特征。每个数据对象都由若干个属性组成。

2. 分类器

分类的关键是找出一个合适的分类器，也就是分类函数或分类模型。分类的过程是依据已知的样本数据构造一个分类函数或者分类模型。该分类函数或分类模型能够把数据库中的数据对象映射到某个给定的类别中，从而确定数据对象的类别。

3. 训练集

分类的样本数据集合称为训练集，是构造分类器的基础。训练集由数据对象组成，每个数据对象的所属类别已知。在构造分类器时，需要输入包含一定样本数据的训练集。选取的训练集是否合适，直接影响到分类器性能的好坏。

4. 测试集

与训练集一样，测试集也是由类别属性已知的数据对象组成的。测试集用来测试基于训练集构造的分类器的性能。在分类器产生后，由分类器判定测试集对象的所属类别，再与测试集中已知的所属类别进行比较，得出分类器的正确率等一系列评价性能。

下面给出分类问题的形式化定义。给定一个数据集 $D=\{t_1,t_2,\cdots,t_n\}$ 和一组类 $C=\{C_1,C_2,\cdots,C_m\}$，分类问题是确定一个映射 f：$D\rightarrow C$，使得每个数据对象 t_i 被分配到某个类中。一个类 C_j 包含映射到该类中的所有数据对象，C 构成数据集 D 的一个划分，即 $C_j=\{t_i \mid f(t_i)=C_j, 1\leqslant i\leqslant n, t_i\in D\}$，并且 $C_j\cap C_i=\varnothing$。

分类问题的样本数据(训练集)是由一个个数据对象组成的。每一个数据对象包含若干个属性，组成一个特征向量。训练集中的每一个数据对象有一个特定的类别属性(类标号)。该类标号是分类系统的输入，通常是历史经验数据。分析样本数据，通过训练集中的样本数据表现出的特性，为每个类找到一种准确的描述或者模型。基于此模型对未来的测试数据进行分类，就是类别预测。

另外，如果上面介绍的过程所构造的模型预测的是连续值或有序值，而不是离散的类标号，这种模型通常叫预测器(predictor)。在通常情况下，将离散的类标号预测叫分类，将连续的数值预测叫回归分析(regression analysis)。分类和回归分析是预测问题的两种主要类型。例如，销售经理希望预测一位顾客在一次购物过程中将花多少钱，该数据分析任务就是数值预测，也叫回归分析。

5.1.2 分类的过程

从例子中学习(learning from examples)是机器学习中最常用的方法。对于分类来说，就是根据带有类标号的样本例子建立分类模型，应用此分类模型对测试样本进行类标号预测。分类过程主要包含两个步骤：建立模型和应用模型进行分类。

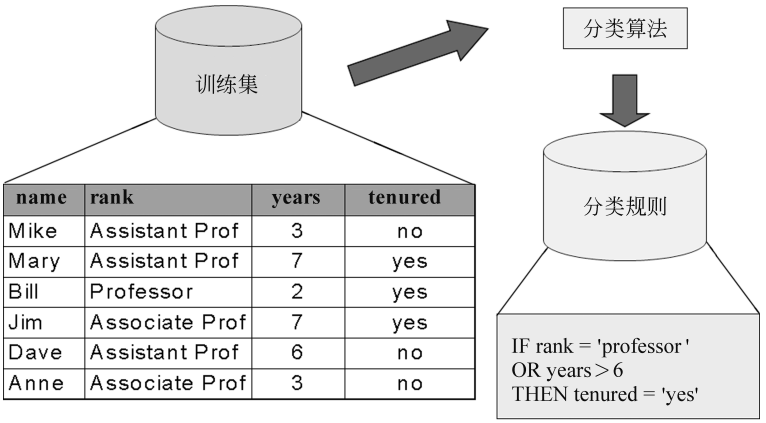
第一步：建立模型。

建立模型就是通过分析由属性描述的数据集来构造分类器模型。每个样本数据都属于一个预定义的类，由一个称作类标号的属性确定。例如，对于样本数据 X，x 是该数据的常规属性，y 是该数据的类标签属性，X 就可以简单地表示为二维关系 $X(x,y)$。其中，x 往往包含多个特征值，是多维向量。由于训练集提供了每个训练样本的类标号，所以建模过程是有监督学习，即模型的学习过程是在被告知每个训练样本属于哪个类的监

督下进行的。

分类模型的表示形式有分类规则、决策树以及等式、不等式、规则式等。分类模型对历史数据的分布进行了归纳,可以用于对测试数据样本进行分类,也有助于更好地理解数据集的内容或含义。

图 5.1 给出了利用某校教师情况数据库建立分类模型的过程。训练数据的常规属性有 name(名字)、rank(职称)和 years(工龄),类标号属性是 tenured(是否获得终身职位)。分类模型以分类规则的形式提供。

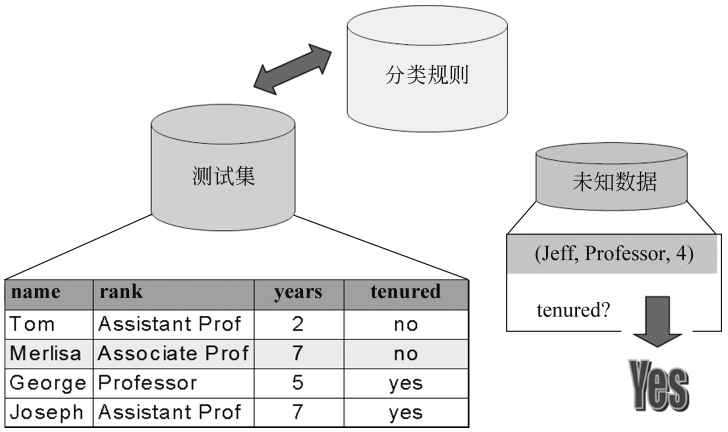


name	rank	years	tenured
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

图 5.1　建立分类模型示例

第二步:应用分类模型进行分类。

首先根据特定领域对分类模型的性能要求,对第一步建立的分类模型的性能进行科学评估,具体评估方法在 5.5 节中详细描述。如果该分类模型满足研究领域的性能要求,就可以用该分类模型对类标号未知的数据或对象进行分类。例如,在图 5.2 中,通过分析现有教师数据得到的分类规则可以用来预测新入职的或未来招聘的教师是否能够获得终身职位。



name	rank	years	tenured
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

图 5.2　用分类模型进行分类

简单地说，模型的建立就是使用训练数据进行学习的过程，模型的应用就是对类标号未知的数据进行分类的过程。

5.1.3 分类器常见构造方法

从构造分类器依据的理论来源看，分类器常见的构造方法有数理统计方法、机器学习方法和神经网络方法等。

(1) 数理统计方法包括贝叶斯方法和非参数方法。常见的 KNN 算法就属于非参数方法。

(2) 机器学习方法包括决策树法和规则归纳法。

(3) 神经网络方法主要是 BP 算法。

(4) 其他方法包括粗糙集方法、遗传算法等。

从构造分类器使用的技术来分，可以将分类器构造方法分为 4 种类型：

(1) 基于距离的分类方法，主要是 KNN 算法。

(2) 决策树分类方法，主要有 ID3、C4.5 等。

(3) 贝叶斯方法，主要包括朴素贝叶斯方法、最大期望算法。

(4) 规则归纳方法，主要包括 AQ 算法、CN2 算法和 FOIL 算法。

5.2 KNN 分 类

古语“物以类聚，人以群分”和“近朱者赤，近墨者黑”都解释了周围环境对人的巨大影响，对于一个人，可以从其朋友和亲属的品性作出大概的判断。同理，分类时也可以通过与测试数据最接近的训练样本的类别来进行判断。Cover 和 Hart 于 1968 年提出了 KNN 算法。该算法通过计算每个训练数据到待分类数据的距离，选择与待分类数据距离最近的 k 个训练数据，k 个训练数据中哪个类别的训练数据占多数，则待分类数据就属于哪个类别。所谓 k 最近邻，就是 k 个最近的邻居的意思。KNN 算法的基本思想是每个样本都可以用与它最接近的 k 个邻居来代表。其中的 k 表示最接近待分类样本的 k 个训练数据，一般 k 取奇数，这是为了保证在投票的时候不会出现票数相同的情况。KNN 分类算法是一种有监督的学习方法，是根据不同特征值之间的距离进行分类的一种简单的机器学习方法，是数据挖掘技术中比较常用的分类算法，其思路简单、直观。由于其实现的简单性及较高的分类准确性，因此该算法在很多领域得到了广泛应用。

如果一个样本在特征空间中的 k 个最相似(即在特征空间中最邻近)的样本中的大多数属于某个类别，则该样本也属于这个类别。在 KNN 算法中，待分类样本选择的邻居都是已经正确分类的对象。该方法在分类决策上只依据最邻近的 k 个样本的类别来决定待分样本所属的类别。由于 KNN 算法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样类本集来说，KNN 算法较其他方法更为适合。

KNN 算法中的基本要素有距离计算和 k 值确定。两个点的距离越近，意味着这两个点属于一个分类的可能性越大。距离计算方法包括欧几里得距离、曼哈顿距离、余弦距

离等,表 5.1 给出了这 3 种距离计算方法的基本情况。当然,要根据具体的分类对象选择合适的距离度量方法。采用不同的距离度量方法,对最终的结果影响很大。在利用 KNN 算法判断类别时 k 的取值很重要。可以在测试数据集测试完毕后计算分类误差率,然后设定不同的 k 值重新进行训练,最后取分类误差率最小的 k 值作为以后分类使用的 k 值。

表 5.1 3 种常用距离计算方法的基本情况

距离计算方法	说 明	公 式
欧几里得距离	两点间的直线距离	$D(a,b)=\sqrt{\sum_i (a_i-b_i)^2}$
曼哈顿距离	两点的横向距离加上纵向距离	$d(x,y)=\sum_{i=1}^{n} \lvert x_i-y_i \rvert$
余弦距离	特征向量夹角的余弦值,更适合解决异常值和数据稀疏问题	$\cos<a,b>=\frac{a\cdot b}{\lvert a \rvert \lvert b \rvert}$

KNN 算法的步骤如下:

(1) 计算距离。给定测试对象,计算它与训练集中的每个训练样本的距离。

(2) 寻找邻居。圈定距离最近的 k 个训练样本,作为测试对象的近邻。

(3) 进行分类。根据这 k 个近邻归属的主要类别,对测试对象进行分类。

在实际的算法实现中,可以维护一个大小为 k 的按距离由大到小排列的优先级队列,用于存储最近邻训练样本。随机从训练集中选取 k 个样本作为初始的最近邻样本,分别计算测试数据到这 k 个训练样本的距离,将训练样本标号和距离存入优先级队列;遍历训练集,计算当前训练样本与测试数据的距离,将所得距离 L 与优先级队列中的最大距离 L_{max} 进行比较。若 $L\geqslant L_{max}$,则舍弃该训练样本,遍历下一个训练样本;若 $L<L_{max}$,则删除优先级队列中距离最大的训练样本,将当前训练样本存入优先级队列。遍历完毕,计算优先级队列中 k 个样本的多数类,并将其作为测试数据的类别。

【例 5.1】 分析表 5.2 中学生的高度,"身高"用于计算距离,$k=5$,对<Pat,女,1.6>分类。

表 5.2 学生身高信息表

序 号	姓 名	性 别	身高/m	高 度
1	Kristina	女	1.6	矮
2	Jim	男	2	高
3	Maggie	女	1.9	中等
4	Martha	女	1.88	中等
5	Stephanie	女	1.7	矮
6	Bob	男	1.85	中等
7	Kathy	女	1.6	矮
8	Dave	男	1.7	矮

续表

序　号	姓　名	性　别	身高/m	高　度
9	Worth	男	2.2	高
10	Steven	男	2.1	高
11	Debbie	女	1.8	中等
12	Todd	男	1.95	中等
13	Kim	女	1.9	中等
14	Amy	女	1.8	中等
15	Wynette	女	1.75	中等

对前 $k=5$ 个记录，$N=\{<$Kristina，女，1.6$>,<$ Jim，男，2$>,<$ Maggie，女，1.9$>,<$ Martha，女，1.88$>,<$ Stephanie，女，1.7$>\}$。

对第 6 个记录$<$ Bob，男，1.85$>$计算距离，得到 $N=\{<$Kristina，女，1.6$>,<$ Bob，男，1.85$>,<$ Maggie，女，1.9$>,<$ Martha，女，1.88$>,<$ Stephanie，女，1.7$>\}$。

对第 7 个记录$<$ Kathy，女，1.6$>$计算距离，得到 $N=\{<$Kristina，女，1.6$>,<$ Bob，男，1.85$>,<$ Kathy，女，1.6$>,<$ Martha，女，1.88$>,<$ Stephanie，女，1.7$>\}$。

对第 8 个记录$<$ Dave，男，1.7$>$计算距离，得到 $N=\{<$Kristina，女，1.6$>,<$ Dave，男，1.7$>,<$ Kathy，女，1.6$>,<$ Martha，女，1.88$>,<$ Stephanie，女，1.7$>\}$。

对第 9 个和第 10 个记录计算距离，优先级队列没有变化。

对第 11 个记录$<$ Debbie，女，1.8$>$计算距离，得到 $N=\{<$Kristina，女，1.6$>,<$ Dave，男，1.7$>,<$ Kathy，女，1.6$>,<$ Debbie，女，1.8$>,<$ Stephanie，女，1.7$>\}$。

对第 12～14 个记录计算距离，优先级队列没有变化。

对第 15 个记录$<$ Wynette，女，1.75$>$计算距离，得到 $N=\{<$Kristina，女，1.6$>,<$ Dave，男，1.7$>,<$ Kathy，女，1.6$>,<$ Wynette，女，1.75$>,<$ Stephanie，女，1.7$>\}$。

最后的优先级队列是$<$Kristina，女，1.6$>,<$Kathy，女，1.6$>,<$Stephanie，女，1.7$>,<$Dave，男，1.7$>,<$Wynette，女，1.75$>$。在这 5 个训练样本中，4 个属于“矮”，1 个属于“中等”。最终利用 KNN 算法认为 Pat 是“矮”类别。

KNN 算法主要依据邻近的 k 个样本进行类别的判断。然后依据 k 个样本中出现次数最多的类别作为待分类样本的类别。

KNN 算法的优点如下：

(1) 简单，易于理解，易于实现，无须估计参数，无须训练。

(2) 适合对稀有事件进行分类。

(3) 特别适用于多分类问题。

KNN 算法有如下缺点：

(1) 对每一个待分类样本都要计算它到全体训练样本的距离，才能求得它的 k 个最近邻。对测试样本进行分类时计算量大，内存开销大。

(2) 可解释性较差，无法给出像决策树那样的规则。

(3) 当样本不平衡(例如,一个类的样本容量很大,而其他类的样本容量很小)时,有可能导致以下情况:当输入一个新样本时,该样本的 k 个近邻中大容量类的样本占多数,影响分类结果。

5.3 贝叶斯分类

贝叶斯分类是一种基于统计的分类方法,它利用概率统计知识进行分类。贝叶斯分类使用概率表示各种形式的不确定性。在先验概率与类条件概率已知的情况下,计算给定样本属于特定类的概率,并选定其中概率最大的一个类别作为该样本的最终类别。朴素贝叶斯分类算法逻辑简单,并在大型数据库中表现出高准确率和高计算速度等特点。贝叶斯定理是朴素贝叶斯分类算法的理论依据。

5.3.1 贝叶斯定理

贝叶斯定理的基础是概率论中的乘法公式:

$$P(AB)=P(A)P(B\mid A)=P(B)P(A\mid B)$$

可以变形为

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

贝叶斯定理由英国数学家贝叶斯(Thomas Bayes,1702—1761)提出的,用来描述两个条件概率之间的确定关系。贝叶斯定理是乘法公式的变形。

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

A 和 B 是独立的事件。$P(B|A)$称为后验概率或在条件 A 下 B 的后验概率,$P(B)$称为先验概率或 B 的先验概率。贝叶斯定理提供了一种由 $P(B)$、$P(A)$和 $P(A|B)$计算后验概率的方法。

例如,假设数据样本是由属性 age 和 income 描述的顾客,X 是一位 25 岁、收入为 5000 美元的顾客,即 X 的属性值为:age=25,income=$5000。对应的假设 H 是顾客 X 将购买计算机。

- $P(H|X)$表示在已知某顾客信息 age=25、income=$5000 的条件下,该顾客会买计算机的概率。
- $P(H)$表示任意顾客购买计算机的概率。
- $P(X|H)$表示已知顾客会购买计算机,那么该顾客属性为 age=25、income=$5000 的概率。
- $P(X)$表示在所有顾客信息的集合中,顾客属性为 age=25,income=$5000 的概率。

【例 5.2】 现分别有 A、B 两个容器,在容器 A 里有 6 个红球和 4 个黑球,在容器 B 里有 2 个红球和 8 个黑球。现从这两个容器之一里任意取了一个球,且是红球,这个红球来自容器 A 的概率是多少?

解：假设取出红球为事件 B，从容器 A 里取出一个球为事件 A，则有

$$P(B)=\frac{8}{20}=\frac{2}{5},\quad P(A)=\frac{1}{2},\quad P(B\mid A)=\frac{6}{10}=\frac{3}{5}$$

根据贝叶斯定理，有

$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)}=\frac{\frac{3}{5}\times\frac{1}{2}}{\frac{2}{5}}=\frac{3}{4}$$

【例 5.3】 表 5.3 是某医院近期门诊病人情况。一个头疼的学生被诊断为感冒的概率有多大？

表 5.3 某医院近期门诊病人情况

症　状	职　业	诊断结果
头疼	学生	感冒
头疼	农民	过敏
打喷嚏	工人	脑震荡
打喷嚏	工人	感冒
头疼	护士	感冒
打喷嚏	学生	脑震荡

令 $B=\{$既头疼又是学生$\}$，$B_1=\{$头疼$\}$，$B_2=\{$学生$\}$，$A=\{$感冒$\}$，则问题转化为求 $P(A|B)$。

由于“头疼”和“是学生”是两个独立的对象，贝叶斯定理就转化为

$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)}=\frac{P(B_1\mid A)P(B_2\mid A)P(A)}{P(B_1)P(B_2)}$$

$$=\frac{\frac{2}{3}\times\frac{1}{3}\times\frac{1}{2}}{\frac{1}{2}\times\frac{2}{6}}=\frac{2}{3}\approx 66.67\%$$

因此，这个头疼的学生有大约 66.67%的概率被诊断为感冒。

请自己计算一下这个头疼的学生被诊断为脑震荡的概率是多少。

5.3.2 朴素贝叶斯分类算法

朴素贝叶斯分类算法是一种十分简单的分类算法，它以对象各特征属性相互独立为假设，以贝叶斯定理为基础，对于给出的待分类对象，求解在此对象出现的条件下各个类别出现的概率，哪个概率最大，就认为此待分类对象属于哪个类别。例如，在街上看到一个黑人，猜他是从哪里来的，我们十有八九猜他来自非洲。这是因为黑人来自非洲的概率最高。当然他也可能是美洲人、欧洲人或亚洲人等，但在没有其他可用信息时，人们会选择条件概率最大的类别。

朴素贝叶斯分类算法如下：

(1) 设 $x=\{a_1,a_2,\cdots,a_m\}$为一个待分类项，而每个 a 为 x 的一个特征属性。

(2) 有类别集合 $C=\{y_1,y_2,\cdots,y_n\}$。

(3) 计算 $P(y_1|x),P(y_2|x),\cdots,P(y_n|x)$。

(4) 如果 $P(y_k|x)=\max\{P(y_1|x),P(y_2|x),\cdots,P(y_n|x)\}$，则 $x\in y_k$。

第(3)步是朴素贝叶斯分类算法的关键。首先根据训练集，统计得到在各类别下各个特征属性的条件概率，如下所示：

$P(a_1|y_1),P(a_2|y_1),\cdots,P(a_m|y_1)$

$P(a_1|y_2),P(a_2|y_2),\cdots,P(a_m|y_2)$

$\vdots$

$P(a_1|y_n),P(a_2|y_n),\cdots,P(a_m|y_n)$

因为各个特征属性是条件独立的，则根据贝叶斯定理有

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

因为分母对于所有类别均为常数，所以只要将分子最大化即可。又因为各特征属性是条件独立的，所以有

$$P(x\mid y_i)P(y_i)=P(a_1\mid y_i)P(a_2\mid y_i)\cdots P(a_m\mid y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j\mid y_i)$$

【例 5.4】 表 5.4 给出了一个顾客数据库标记类的训练集 D。使用朴素贝叶斯分类算法预测待分类数据的类标号。数据元组用属性 age、income、student、credit_rating 和 buys_comp 来描述。类标号属性 buys_comp 具有两个不同的值 no 和 yes。设 C_1 对应于类 buys_comp=yes，C_2 对应于类 buys_comp=no。待分类数据为

$$X=(\text{age}\leqslant 25,\text{income}=\text{medium},\text{student}=\text{yes},\text{credit\_rating}=\text{fair})$$

表 5.4 顾客数据库标记类的训练集

ID	age	income	student	credit_rating	buys_comp
1	≤25	high	no	fair	no
2	≤25	high	no	excellent	no
3	26～35	high	no	fair	yes
4	>35	medium	no	fair	yes
5	>35	low	yes	fair	yes
6	>35	low	yes	excellent	no
7	26～35	low	yes	excellent	yes
8	≤25	medium	no	fair	no
9	≤25	low	yes	fair	yes
10	>35	medium	yes	fair	yes

续表

ID	age	income	student	credit_rating	buys_comp
11	≤25	medium	yes	excellent	yes
12	26～35	medium	no	excellent	yes
13	26～35	high	yes	fair	yes
14	>35	medium	no	excellent	no
15	<25	medium	yes	fair	yes

解：(1) 计算每个类的先验概率 $P(C_i)$。

$$P(C_1)=\frac{10}{15}=0.667$$

$$P(C_2)=\frac{5}{15}=0.333$$

(2) 计算每个特征属性对于每个类别的条件概率。

$$P(\text{age}\leqslant 25 \mid \text{buys\_comp}=\text{yes})=\frac{3}{10}=0.3$$

$$P(\text{income}\leqslant \text{medium} \mid \text{buys\_comp}=\text{yes})=\frac{5}{10}=0.5$$

$$P(\text{student}\leqslant \text{yes} \mid \text{buys\_comp}=\text{yes})=\frac{7}{10}=0.7$$

$$P(\text{credit\_rating}\leqslant \text{fair} \mid \text{buys\_comp}=\text{yes})=\frac{7}{10}=0.7$$

$$P(\text{age}\leqslant 25 \mid \text{buys\_comp}=\text{no})=\frac{3}{5}=0.6$$

$$P(\text{income}\leqslant \text{medium} \mid \text{buys\_comp}=\text{no})=\frac{2}{5}=0.4$$

$$P(\text{student}\leqslant \text{yes} \mid \text{buys\_comp}=\text{no})=\frac{1}{5}=0.2$$

$$P(\text{credit\_rating}\leqslant \text{fair} \mid \text{buys\_comp}=\text{no})=\frac{2}{5}=0.4$$

(3) 计算条件概率 $P(X|C_i)$。

$$P(X \mid \text{buys\_comp}=\text{yes})=0.3\times 0.5\times 0.7\times 0.7=0.0735$$

$$P(X \mid \text{buys\_comp}=\text{no})=0.6\times 0.4\times 0.2\times 0.4=0.0192$$

(4) 计算对于每个类 C_i 的 $P(X|y_i)P(y_i)$。

$$P(X \mid \text{buys\_comp}=\text{yes})P(\text{buys\_comp}=\text{yes})=0.0735\times 0.667=0.049$$

$$P(X \mid \text{buys\_comp}=\text{no})P(\text{buys\_comp}=\text{no})=0.0192\times 0.333=0.00639$$

因此，对于样本 X，朴素贝叶斯分类算法的预测为

$$\text{buys\_comp}=\text{yes}$$

朴素贝叶斯分类算法有以下优点：