第一章 统计学习及监督学习概论

**1.统计学习：**

统计学习（statistical learning）是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。统计学习也称为统计机器学习（statistical machine learning）。

**2.统计学习的主要特点：**

统计学习的主要特点是：（1）统计学习以计算机及网络为平台，是建立在计算机及网络上的：（2）统计学习以数据为研究对象，是数据驱动的学科：（3）统计学习的目的是对数据进行预测与分析：（4）统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析：（5）统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独自的理论体系与方法论。

**3.统计学习研究的对象：**

统计学习研究的对象是数据。

**4.统计学习的目的：**

用于对数据的预测与分析。

**5.统计学习的方法：**

由监督学习、无监督学习和强化学习等组成。

**6.统计学习方法的三要素：**

模型、策略和算法。

**7.监督学习：**

从标注数据中学习预测模型的机器学习问题。

**8.监督学习的公式表示：**

监督学习中输入实例的特征向量记作

表示第个特征。注意与不同，本书通常用表示多个输入变量中的第个变量，即

监督学习从训练数据（training data）集合中学习模型，对测试数据（test data）进行预测。训练数据由输入（或特征向量）与输出对组成，训练集通常表示为

**9.监督学习的模型：**

监督学习的模型可以是概率模型或非概率模型，由条件概率分布或决策函数表示，随具体学习方法而定。

**10.无监督学习：**

无监督学习是指从无标注数据中学习预测模型的机器学习问题。

**11.概率模型与非概率模型：**

本书介绍的决策树、朴素贝叶斯、隐马尔可夫模型、条件随机场、概率潜在语义分析、潜在狄利克雷分配、高斯混合模型是概率模型。感知机、支持向量机、*k*近邻、AdaBoost、*k*均值、潜在语义分析，以及神经网络是非概率模型。逻辑斯谛回归既可看作是概率模型，又可看作是非概率模型。

**12.逻辑斯蒂回归既是概率模型又是非概率模型的原因：**

从概率的角度来看，逻辑斯蒂回归模型假设样本的标记（输出）服从伯努利分布。从非概率的角度来看，逻辑斯蒂回归模型可以被看作是一种线性分类器，它通过计算输入样本的特征向量与参数的乘积，将样本映射到一个实数，然后通过一个阈值将其二值化为0或1，从而完成分类。

**13.参数化模型与非参数化模型：**

本书介绍的感知机、朴素贝叶斯、逻辑斯谛回归、*k*均值、高斯混合模型是参数化模型。决策树、支持向量机、AdaBoost、*k*近邻、潜在语义分析、概率潜在语义分析潜在狄利克雷分配是非参数化模型。

**14.参数空间：**

参数向量取值于维欧式空间

**15.损失函数和风险函数：**

输出的预测值与真实值Y可能一致也可能不一致，用一个损失函数或代价函数来度量预测错误的程度。

**16.常用的损失函数：**

（1）0-1损失函数（0-1 loss function）

（2）平方损失函数（quadratic loss function）

非负性、可微

（3）绝对损失函数（absolute loss function）

欧式距离的绝对值，一范数

（4）对数损失函数（logarithmic loss function）或对数似然函数（log-likelihood loss function）

损失函数值越小，模型就越好。由于模型的输入、输出是随机变量，遵循联合分布，所以损失函数的期望是

这是理论上模型关于联合分布的平均意义下的损失，称为风险函数（risk function）或期望损失（expected loss）。

**17.经验风险：**

模型关于训练数据集的平均损失称为经验风险（empirical risk）或经验损失（empirical loss），记作：

**18.监督学习的基本策略：**

经验风险最小化和结构风险最小化。

**19.结构风险：**

结构风险最小化（structural risk minimization，SRM）是为了防止过拟合而提出来的策略。结构风险最小化等价于正则化（regularization）。结构风险在经验风险上加上表示模型复杂度的正则化项（regularizer）或罚项（penalty term）。在假设空间、损失函数以及训练数据集确定的情况下，结构风险的定义是：

**20.经验风险和结构风险最小化所转化为的最优化问题：**

经验风险最小化：

结构风险最小化：

**21.过拟合：**

以为追求提高对训练数据的预测能力，所选模型的复杂度则往往会比真模型更高。学习时选择的模型所包含的参数过多，以致出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象。

**22.防止过拟合的两种模型选择方法：**

正则化与交叉验证。

**23.正则化：**

模型选择的典型方法是正则化（regularization）。正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项（regularizer）或罚项（penalty term）。正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大。比如，正则化项可以是模型参数向量的范数。

正则化符合奥卡姆剃刀（Occam’s razor）原理。奥卡姆剃刀原理应用于模型选择时变为以下想法：在所有可能选择的模型中，能够很好地解释已知数据并且十分简单才是最好的模型，也就是应该选择的模型。从贝叶斯估计的角度来看，正则化项对应于模型的先验概率。可以假设复杂的模型有较小的先验概率，简单的模型有较大的先验概率。

**24.交叉验证的基本想法：**

重复得使用数据，把给定的数据进行切分，将切分的数据集组合为训练集与测试集，在此基础上反复地进行训练、测试以及模型选择。

**25.交叉验证的三种方法：**

1.简单交叉验证

简单交叉验证方法是:首先随机地将已给数据分为两部分，一部分作为训练集，另一部分作为测试集（例如，70%的数据为训练集，30%的数据为测试集）；然后用训练集在各种条件下（例如，不同的参数个数）训练模型，从而得到不同的模型；在测试集上评价各个模型的测试误差，选出测试误差最小的模型。

2.S折验证

应用最多的是折交叉验证（*S*-fold cross validation），方法如下：首先随机地将已给数据切分为个互不相交大小相同的子集；然后利用个子集的数据训练模型，利用余下的子集测试模型；将这一过程对可能的种选择重复进行；最后选出次评测中平均测试误差最小的模型。

3.留一交叉验证

折交叉验证的特殊情形是称为留一交叉验证（leave-one-out cross validation），往往在数据缺乏的情况下使用。这里，是给定数据集的容量。

**26.泛化误差：**

首先给出泛化误差的定义。如果学到的模型是，那么用这个模型对未知数据预测的误差即为泛化误差（generalization error）：

**27.泛化能力：**

由方法学习到的模型对未知数据的预测能力。

**28.监督学习方法分类：**

监督学习方法又可以分成生成方法和判别方法。

**29.生成方法：**

生成方法由数据学习联合概率分布，然后求出条件概率分布作为预测的模型，即生成模型：

这样的方法之所以称为生成方法，是因为模型表示了给定输入产生输出的生成关系。典型的生成模型有朴素贝叶斯法和隐马尔可夫模型，将在后面章节进行相关讲述。

**30.判别方法：**

判别方法由数据直接学习决策函数或者条件概率分布作为预测的模型，即判别模型。判别方法关心的是对给定的输入，应该预测什么样的输出。典型的判别模型包括：近邻法、感知机、决策树、逻辑斯谛回归模型、最大熵模型、支持向量机、提升方法和条件随机场等，将在后面章节讲述。

**31.生成方法和判别方法的特点：**

生成方法的特点：生成方法可以还原出联合概率分布，而判别方法则不能;生成方法的学习收敛速度更快，即当样本容量增加的时候，学到的模型可以更快地收敛于真实模型；当存在隐变量时，仍可以用生成方法学习，此时判别方法就不能用。

判别方法的特点：判别方法直接学习的是条件概率或决策函数，直接面对预测，往往学习的准确率更高；由于直接学习或，可以对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习问题。

**32.评级分类器性能的指标：**

**给出数据，计算混淆矩阵**

表格

描述已自动生成

TP——将正类预测为正类数；

FN——将正类预测为负类数；

FP——将负类预测为正类数；

TN——将负类预测为负类数。

精确率定义为

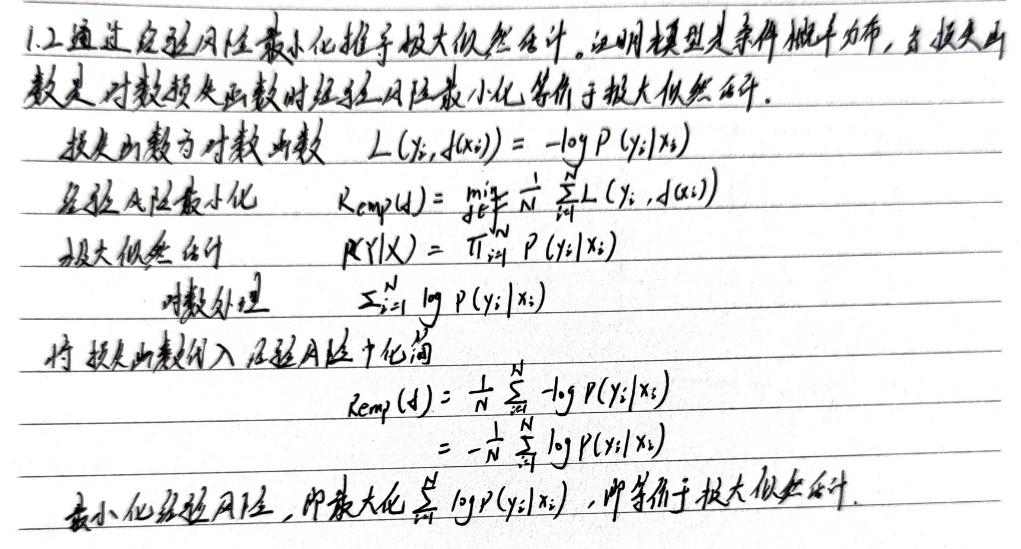
召回率定义为

此外，还有值，是精确率和召回率的调和均值，即

精确率和召回率都高时，值也会高。

**33.习题1.2通过经验风险最小化推导极大似然估计：**

证明模型是条件概率分布，当损失函数是对数损失时，经验风险最小化等价于极大似然估计。



答：

当模型是条件概率分布且损失函数是对数损失时，最小化结构风险是：

我们公式(1)里面的极小化转化为极大化，于是有

因为**以为底数的对数函数是增函数**，那么就相当于极大化下面的似然函数：

将似然函数取对数变为对数似然方程，再利用当底数时，对数函数是增函数，就可以完成由极大似然估计推导最小化经验风险。

第二章 感知机

**1.感知机：**

感知机（perceptron）是二类分类的线性分类模型，其输入为实例的特征向量，输出为实例的类别，取和二值。

**2.感知机属于判别模型：**

感知机对应于输入空间（特征空间）中将实例划分为正负两类的分离超平面，属于判别模型。

**3.感知机学习的目标：**

感知机学习旨在求出将训练数据进行线性划分的分离超平面，导入基于误分类的损失函数，利用梯度下降法对损失函数进行极小化，求得感知机模型。

**4.感知机的两种形式：**

感知机分为原始形式和对偶形式。

**5.感知机的定义：**

**定义2.1（感知机）** 假设输入空间（特征空间）是，输出空间是。输入表示实例的特征向量，对应于输入空间（特征空间）的点；输出表示实例的类别。由输入空间到输出空间的如下函数：

称为感知机。其中，和为感知机模型参数，叫作权值（weight）或权值向量（weight vector），叫作偏置（bias），表示和的内积。是符号函数，即

感知机是一种线性分类模型，属于判别模型。感知机模型的假设空间是定义在特征空间中的所有线性分类模型（linear classification model）或线性分类器（linear classifier），即函数集合。

感知机有如下几何解释：线性方程

对应于特征空间中的一个超平面，其中是超平面的法向量，是超平面的截距。这个超平面将特征空间划分为两个部分。位于两部分的点（特征向量）分别被分为正、负两类。因此，超平面称为分离超平面（separating hyperplane），如图2.1所示。

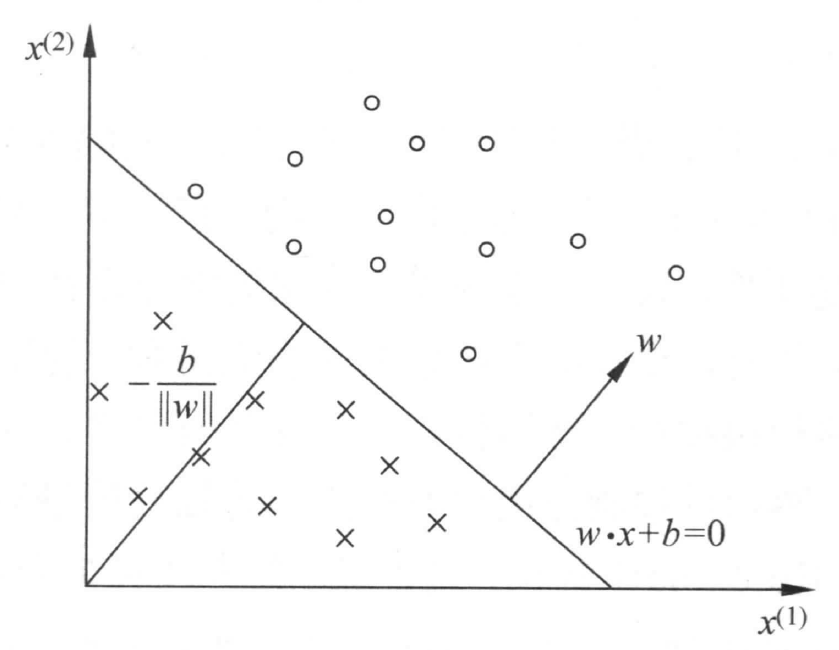


图2.1 感知机模型

**6.数据集的线性可分性：**

对所有的实例，有，对所有的实例，有，则称数据集为线性可分数据集（linearly separable data set）；否则，称数据集线性不可分。

**7.感知机的学习策略：**

假设训练数据集是线性可分的，感知机学习的目标是求得一个能够将训练集正实例点和负实例点完全正确分开的分离超平面。为了找出这样的超平面，即确定感知机模型参数，，需要确定一个学习策略，即定义（经验）损失函数并将损失函数极小化。

**8.感知机学习的损失函数：**

损失函数的一个自然选择是误分类点的总数。但是，这样的损失函数不是参数，的连续可导函数，不易优化。损失函数的另一个选择是误分类点到超平面的总距离，这是感知机所采用的。为此，首先写出输入空间中任一点到超平面的距离：

这里，是的范数。

其次，对于误分类的数据来说，

成立。因为当时，；而当时，。因此，误分类点到超平面的距离是

这样，假设超平面的误分类点集合为，那么所有误分类点到超平面的总距离为

不考虑，就得到感知机学习的损失函数。

**9.优化过程：**

给定训练数据集

其中，，。感知机学习的损失函数定义为

其中为误分类点的集合。这个损失函数就是感知机学习的经验风险函数。

显然，损失函数是非负的。如果没有误分类点，损失函数值是0。而且，误分类点越少，误分类点离超平面越近，损失函数值就越小。一个特定的样本点的损失函数：在误分类时是参数，的线性函数，在正确分类时是0。因此，给定训练数据集，损失函数是，的连续可导函数。

**10.感知机的学习算法：**

随机梯度下降法

感知机学习算法是对以下最优化问题的算法。给定一个训练数据集

其中，，，求参数，，使其为以下损失函数极小化问题的解

其中为误分类点的集合。

**算法2.1（感知机学习算法的原始形式）**

输入：训练数据集，其中，；学习率；

输出，；感知机模型。

（1）选取初值，；

（2）在训练集中选取数据；

（3）如果，

（4）转至（2），直至训练集中没有误分类点。

**11.理解掌握例2.1：**

**例2.1** 如图2.2所示的训练数据集，其正实例点是，，负实例点是，试用感知机学习算法的原始形式求感知机模型。这里，，。

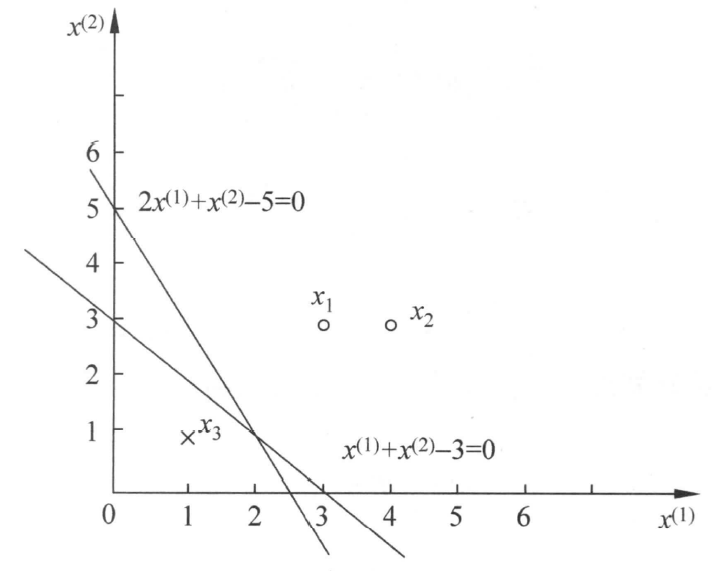


图2.2 感知机示例

**解** 构建最优化问题：

按照算法2.1求解，。。

（1）取初值，

（2）对，，未能被正确分类，更新，

得到线性模型

（3）对，，显然，，被正确分类，不修改，；

对，，被误分类，更新，。

得到线性模型

如此继续下去，直到

对所有数据点，没有误分类点，损失函数达到极小。

分离超平面为：

感知机模型为：

迭代过程见表2.1。

表2.1 例2.1求解的迭代过程

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 迭代次数 | 误分类点 |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

这是在计算中误分类点先后取得到的分离超平面和感知机模型。如果在计算中误分类点依次取，那么得到的分离超平面是。

可见，感知机学习算法由于采用不同的初值或选取不同的误分类点，解可以不同。

**12.理解算法的收敛性：**

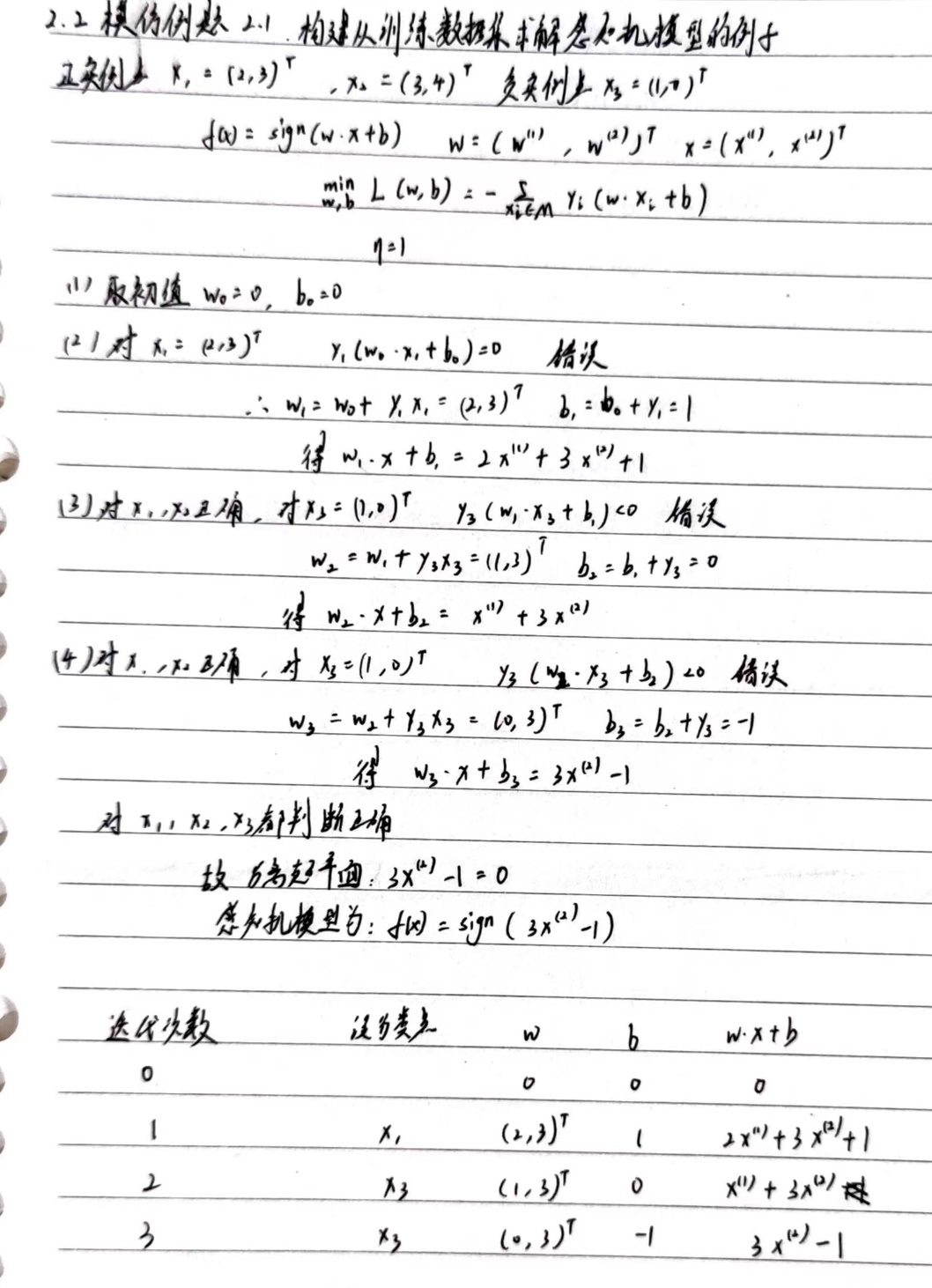
定理2.1（Novikoff） 设训练数据集是线性可分的，其中，，则

（1）存在满足条件的超平面将训练数据集完全正确分开；且存在，对所有

（2）令，则感知机算法2.1在训练数据集上的误分类次数满足不等式

**13.课后习题2.2：**

模仿例题2.1，构建从训练数据集求解感知机模型的例子。



第三章 *k*近邻法

**1.k近邻：**

*k*近邻法（*k*-nearest neighbor，*k*-NN）是一种基本分类与回归方法。本书只讨论分类问题中的*k*近邻法。*k*近邻法的输入为实例的特征向量，对应于特征空间的点；输出为实例的类别，可以取多类。*k*近邻法假设给定一个训练数据集，其中的实例类别已定。分类时，对新的实例，根据其*k*个最近邻的训练实例的类别，通过多数表决等方式进行预测。因此，*k*近邻法不具有显式的学习过程。*k*近邻法实际上利用训练数据集对特征向量空间进行划分，并作为其分类的“模型”。***k*值的选择**、**距离度量**及**分类决策规则**是*k*近邻法的三个基本要素。*k*近邻法1968年由Cover和Hart提出。

**2.k近邻算法：**

**算法3.1（*k*近邻法）**

输入：训练数据集

其中，为实例的特征向量，为实例的类别，；实例特征向量；

输出：实例所属的类。

（1）根据给定的距离度量，在训练集中找出与最近的*k*个点，涵盖这*k*个点的的邻域记作；

（2）在中根据分类决策规则（如多数表决）决定的类别：

式(3.1)中，为指示函数，即当时为1，否则为0。

*k*近邻法的特殊情况是的情形，称为最近邻算法。对于输入的实例点（特征向量），最近邻法将训练数据集中与最邻近的类作为的类。

**3.距离度量：**

设特征空间是维实数向量空间，，，，的距离定义为

这里。当时，称为欧氏距离（Euclidean distance），即

当时，称为曼哈顿距离（Manhattan distance），即

当，它是各个坐标距离的最大值，即

**4.掌握例3.1：**

**例3.1** 已知二维空间的3个点，，，试求在取不同值时，距离下的最近邻点。

**解** 因为和只有第一维的值不同，所以为任何值时，。而

于是得到：等于1或2时，是的最近邻点；大于等于3时，是的最近邻点。

**5.理解*k*值的选择：**

*k*值的选择会对*k*近邻法的结果产生重大影响。

如果选择较小的*k*值，就相当于用较小的邻域中的训练实例进行预测，“学习”的近似误差（approximation error）会减小，只有与输入实例较近的（相似的）训练实例才会对预测结果起作用。但缺点是“学习”的估计误差（estimation error）会增大，预测结果会对近邻的实例点非常敏感。如果邻近的实例点恰巧是噪声，预测就会出错。换句话说，*k*值的减小就意味着整体模型变得复杂，容易发生过拟合。

如果选择较大的*k*值，就相当于用较大邻域中的训练实例进行预测。其优点是可以减少学习的估计误差，但缺点是学习的近似误差会增大。这时与输入实例较远的（不相似的）训练实例也会对预测起作用，使预测发生错误。*k*值的增大就意味着整体的模型变得简单。

如果，那么无论输入实例是什么，都将简单地预测它属于在训练实例中最多的类。这时，模型过于简单，完全忽略训练实例中的大量有用信息，是不可取的。

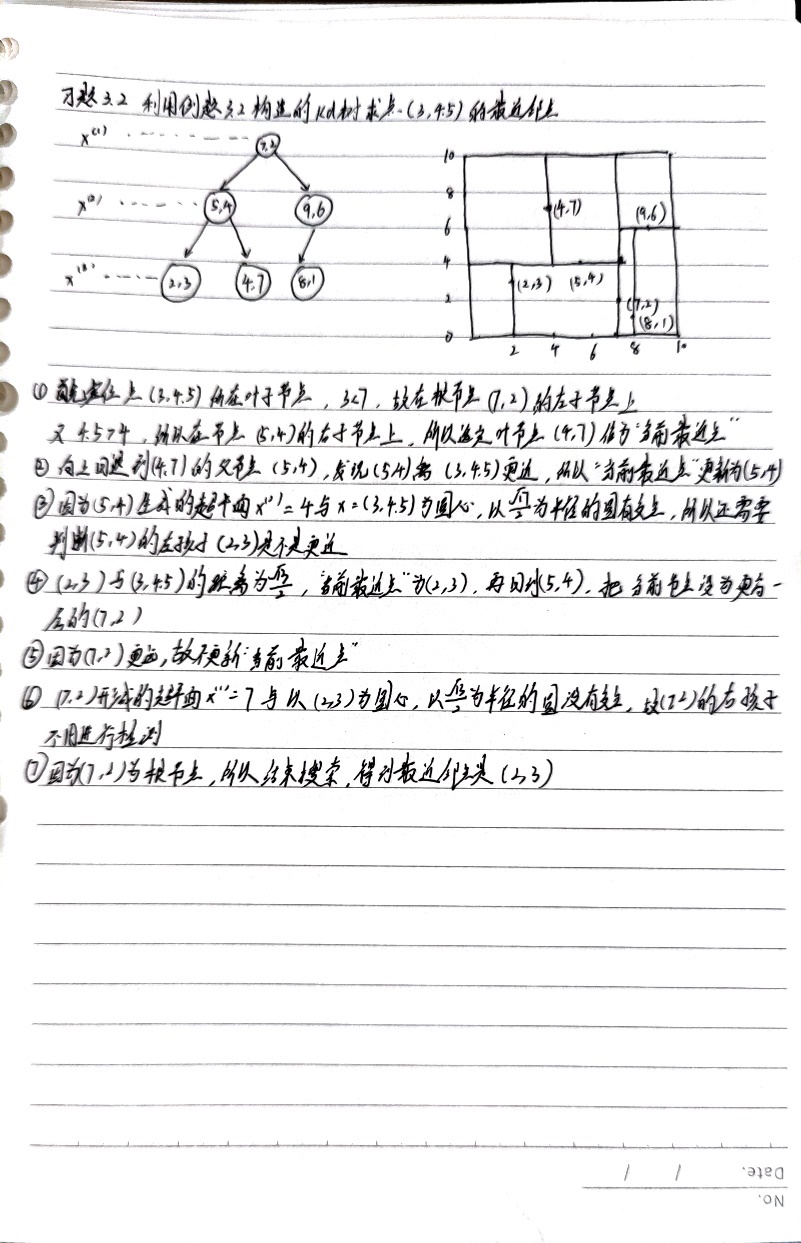
在应用中，*k*值一般取一个比较小的数值。通常采用交叉验证法来选取最优的*k*值。

**6.理解课后习题3.1：**

参照图3.1，在二维空间中给出实例点，画出*k*为1和2时的*k*近邻法构成的空间划分，并对其进行比较，体会*k*值的选择与模型复杂度及预测准确率的关系。

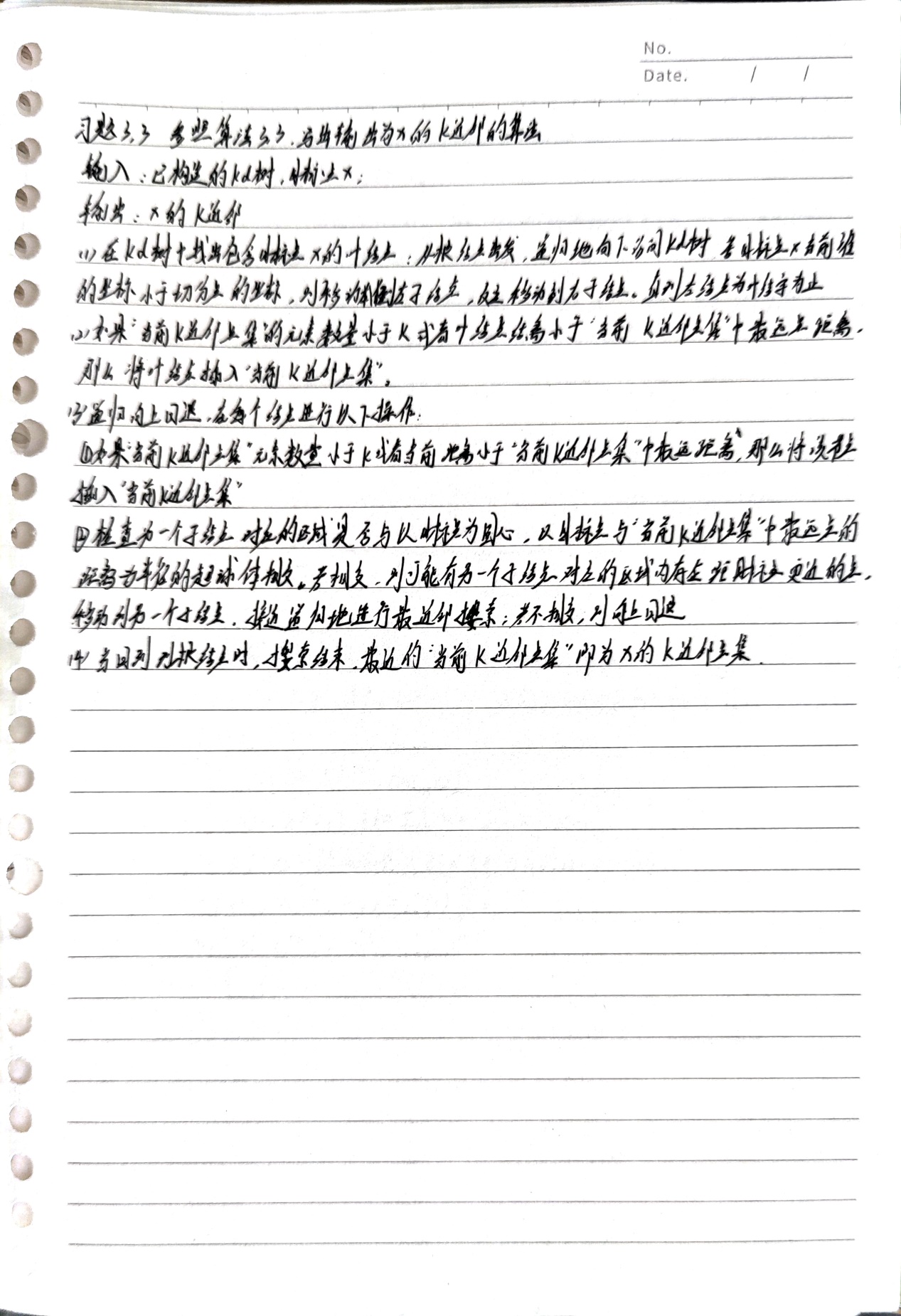
**7.掌握课后习题3.2：**

利用例题3.2构造的树求点的最近邻点。



**8.掌握课后习题3.3：**

参照算法3.3，写出输出为的*k*近邻算法。



第四章 朴素贝叶斯法

**1.朴素贝叶斯法：**

朴素贝叶斯法对条件概率分布作了条件独立性的假设。由于这是一个较强的假设，朴素贝叶斯法也由此得名。具体地，条件独立性假设是

朴素贝叶斯法实际上学习到生成数据的机制，所以属于生成模型。条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的。这一假设使朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率。

**2.后验概率计算根据贝叶斯定理：**

**3.朴素贝叶斯分类器：**

注意到，在式(4.6)中分母对所有都是相同的，所以，

**4.后验最大化：**

朴素贝叶斯将实例分到后验概率最大的类中。这等价于期望风险最小化。假设选择0-1损失函数：

式中是分类决策函数。这时，期望风险函数为

期望是对联合分布取的。由此取条件期望

为了使期望风险最小化，只需对逐个极小化，由此得到：

这样一来，根据期望风险最小化准则就得到了后验概率最大化准则：

**5.极大似然估计：**

在朴素贝叶斯法中，学习意味着估计和。可以应用极大似然估计法估计相应的概率。先验概率的极大似然估计是

设第个特征可能取值的集合为，条件概率的极大似然估计是

式中，是第个样本的第个特征；是第个特征可能取的第个值；为指示函数。

**6.掌握例4.1：**

**例4.1** 试由表4.1的训练数据学习一个朴素贝叶斯分类器并确定的类标记。表中，为特征，取值的集合分别为，，为类标记，。

表4.1 训练数据

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

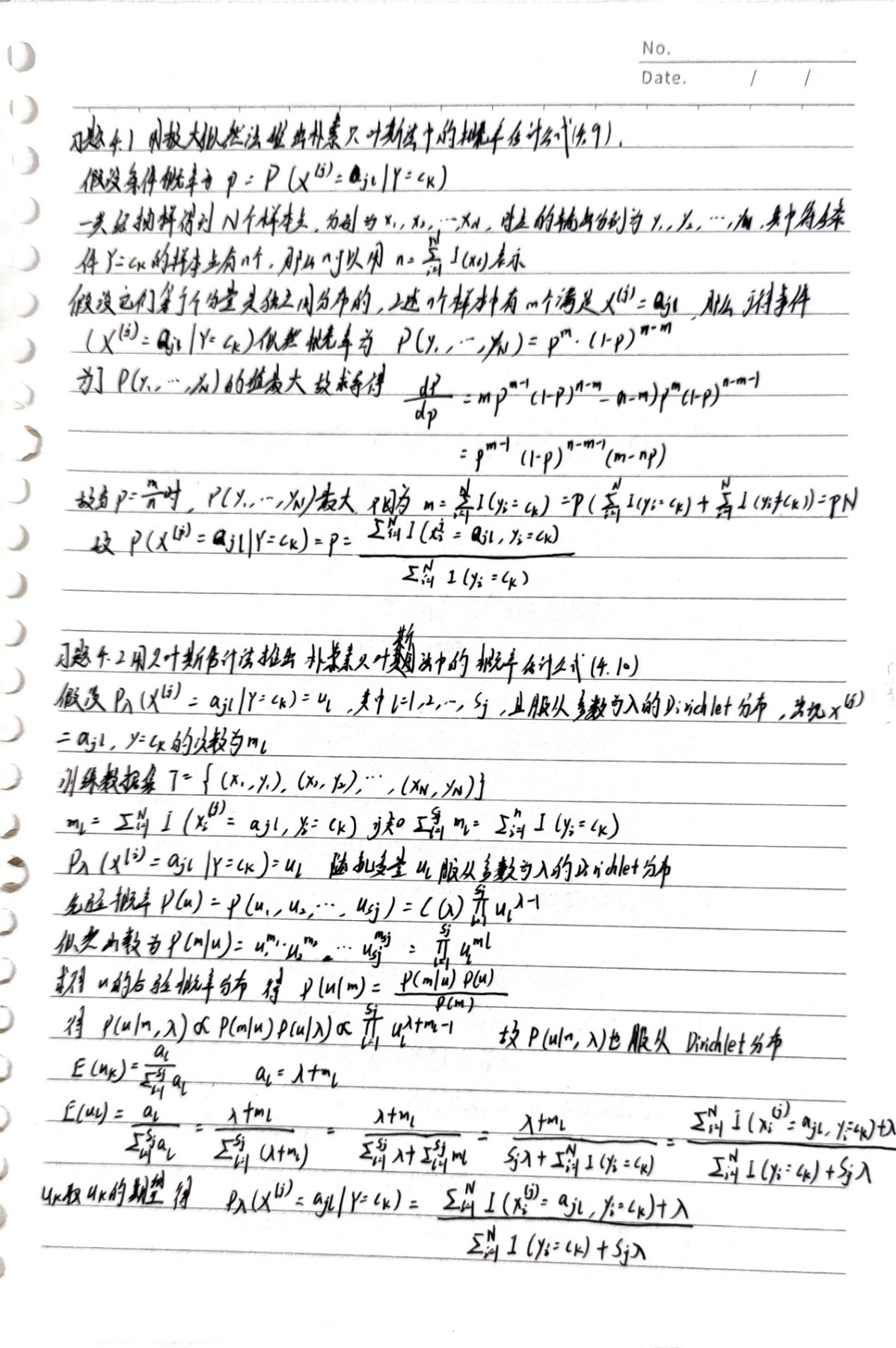
**解** 根据算法4.1，由表4.1，容易计算下列概率：

对于给定的计算：

因为最大，所以。

**7.习题4.1：**

使用极大似然估计法推出朴素贝叶斯法中的概率估计公式(4.9)



我们先将条件概率转化为联合概率

因为我们在1.1小节，已经推导出了，那么我们现在只需要对进行估计即可。同样的，我们假设

那么使得中任何一个等号不成立的概率就是。

所以，我们根据训练数据集可以得到似然函数是

其中，是使成立的数据的个数，也就是

同样的为了计算方便，我们对公式(8)取自然对数得到对数似然函数，然后再关于求导数得到驻点，验证下为唯一的极大值，就可以得到

然后联合公式(4.8)，就可以得到(4.9)。

第五章决策树

**1.决策树：**

**定义5.1（决策树）** 分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点（node）和有向边（directed edge）组成。结点有两种类型：内部结点（internal node）和叶结点（leaf node）。内部结点表示一个特征或属性，叶结点表示一个类。

用决策树分类，从根结点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子结点;这时，每一个子结点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直至达到叶结点。最后将实例分到叶结点的类中。

**2.熵：**

随机变量的熵定义为

**3.熵的性质：**

熵越大，随机变量的不确定性就越大。从定义可验证

**4.条件熵：**

条件熵表示在已知随机变量的条件下随机变量的不确定性。随机变量给定的条件下随机变量的条件（conditional entropy），定义为给定条件下的条件概率分布的熵对的数学期望

这里，。

**5.信息增益：**

信息增益（information gain）表示得知特征的信息而使得类的信息的不确定性减少的程度。

**定义5.2（信息增益）** 特征对训练数据集的信息增益，定义为集合的经验熵与特征给定条件的经验条件熵，即

一般地，熵与条件熵之差称为互信息（mutual information）。决策数学习中的信息增益等价于训练数据集中类与特征的互信息。

**6.信息增益的算法：**

输入：训练数据集和特征；

输出：特征对训练数据集的信息增益。

（1）计算数据集的经验熵

（2）计算特征对数据集的经验条件熵

（3）计算信息增益

**7.运用信息增益算法的例题5.2：**

**例5.2** 对表5.1所给的训练数据集，根据信息增益准则选择最优特征。

解 首先计算经验熵。

然后计算各特征对数据集的信息增益。分别以，，，表示年龄、有工作、有自己的房子和信贷情况4个特征，则

（1）

这里，，分别是中（年龄）取值为青年、中年和老年的样本子集。类似地，

（2）

（3）

（4）

最后，比较各特征的信息增益值。由于特征（有自己的房子）的信息增益值最大，所以选择特征作为最优特征。

**8.信息增益比：**

以信息增益作为划分训练数据集的特征，存在偏向于选择取值较多的特征的问

题。使用信息增益比（information gain ratio）可以对这一问题进行校正。这是特征选

择的另一准则。

**定义5.3（信息增益比）** 特征对训练数据集的信息增益比定义为其信息增益与训练数据集关于特征的值的熵之比，即

其中，，是特征取值的个数。

**9.ID3算法：**

**算法5.2（ID3算法）**

输入：训练数据集，特征集阈值；

输出：决策树。

（1）若中所有实例属于同一类，则为单结点树，并将类作为该结点的类标记，返回；

（2）若，则为单结点树，并将中实例数最大的类作为该结点的类标记，返回；

（3）否则，按算法5.1计算中各特征对的信息增益，选择信息增益最大的特征；

（4）如果的信息增益小于值，则置为单结点树，并将中实例数最大的类作为该结点的类标记，返回；

（5）否则，对的每一可能值，依将分割为若干非空子集，将中实例数最大的类作为标记，构建子结点，由结点及其子结点构成树，返回；

（6）对第个子结点，以为训练集，以为特征集，递归地调用步（1）步(5)，得到子树，返回。

**10.例题5.3运用ID3算法建立决策树：**

**例5.3** 对表5.1的训练数据集，利用ID3算法建立决策树。

**解** 利用例5.2的结果，由于特征（有自己的房子）的信息增益值最大，所以选择特征作为根结点的特征。它将训练数据集分为两个子集（取值为“是”）和（取值为“否”）。由于只有同一类的样本点，所以它成为一个叶结点，结点的类标记为“是”。

对则需从特征（年龄），（有工作）和（信贷情况）中选择新的特征。计算各个特征的信息增益：

选择信息增益最大的特征（有工作）作为结点的特征。由于有两个可能取值，从这一结点引出两个子结点：一个对应“是”（有工作）的子结点，包含3个样本，它们属于同一类，所以这是一个叶结点，类标记为“是”；另一个是对应“否”（无工作）的子结点，包含6个样本，它们也属于同一类，所以这也是一个叶结点，类标记为“否”。

这样生成一棵如图5.5所示的决策树。该决策树只用了两个特征（有两个内部结点）。

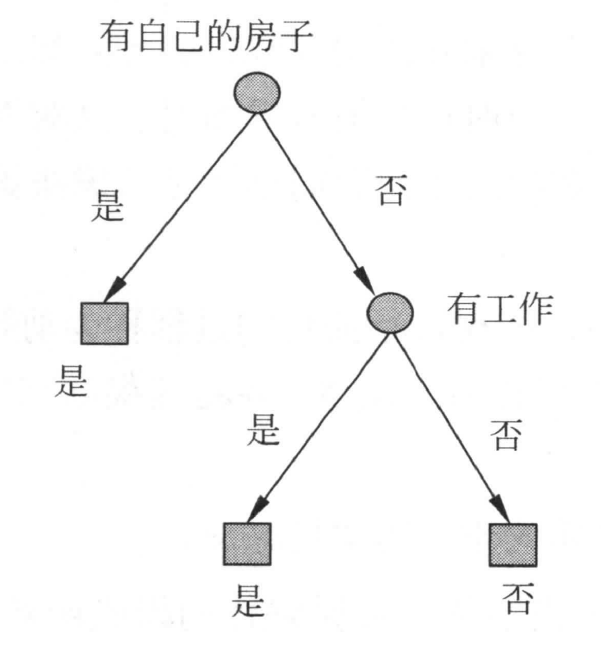


图5.5 决策树的生成

**11.基尼指数：**

分类树用基尼指数选择最优特征，同时决定该特征的最优二值切分点。

**定义5.4（基尼指数）** 分类问题中，假设有个类，样本点属于第类的概率为，则概率分布的基尼指数定义为

对于二类分类问题，若样本点属于第1个类的概率，则概率分布的基尼指数为

对于给定的样本集合，其基尼指数为

这里，是中属于第类的样本子集，是类的个数。

如果样本集合根据特征是否取某一可能值被分成和两部分，即

则在特征的条件下，集合D的基尼指数定义为

基尼指数表示集合的不确定性，基尼指数表示经分割后集合D的不确定性。基尼指数值越大，样本集合的不确定性也就越大，这一点与熵相似。

第六章逻辑斯蒂回归与最大熵模型

**1.逻辑斯蒂分布：**

**定义6.1（逻辑斯蒂分布）** 设是连续随机变量，服从逻辑斯蒂分布是指具有下列分布函数和密度函数：

式中，为位置参数，为形状参数。

**2.逻辑斯蒂回归模型：**

二项逻辑斯谛回归模型（binomial logistic regression model）是一种分类模型，由条件概率分布表示，形式为参数化的逻辑斯谛分布。这里，随机变量取值为实数，随机变量取值为1或0。我们通过监督学习的方法来估计模型参数。

**定义6.2（逻辑斯蒂回归模型）** 二项逻辑斯蒂回归模型是如下的条件概率分布：

这里，是输入，是输出，和**是参数，**称为权值向量，称为偏置，为和的内积。

**3.逻辑斯蒂回归模型参数估计：**

逻辑斯蒂回归模型学习时，对于给定的训练数据集，其中，，，可以应用极大似然估计法估计模型参数，从而得到逻辑斯蒂回归模型。

设：

似然函数为

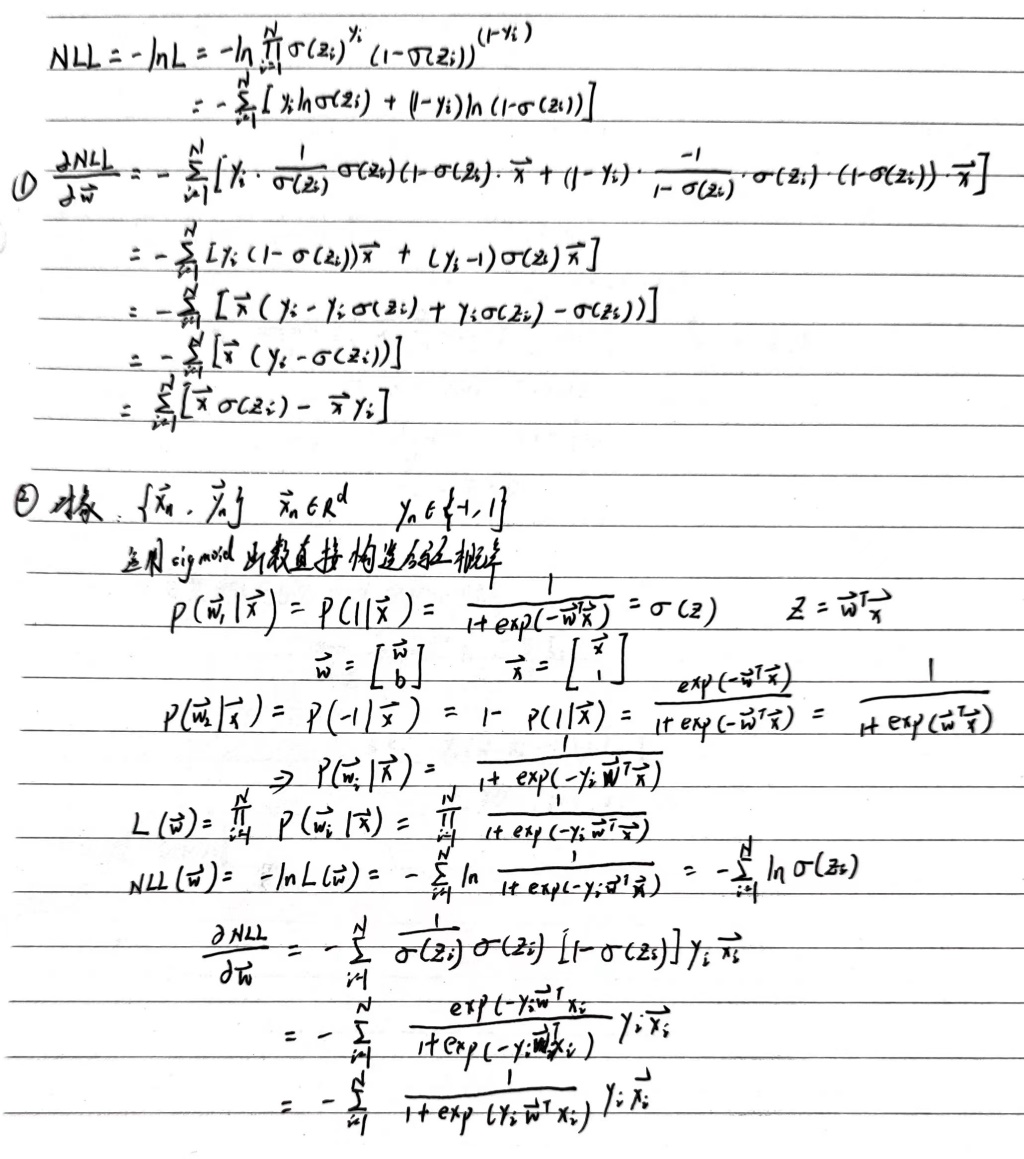
对数似然函数为

对求极大值，得到的估计值。

**4.Logistic Regression：**

文本, 信件

描述已自动生成



**5.课后习题6.2：**

写出逻辑斯谛回归模型学习的梯度下降算法。

**logistic回归模型**

模型：

因此：

可将以上两式进行如下组合：

**极大似然估计**

两边求对数得：

**求解梯度（求导）**

将代入上式得：

因为，所以上式可化简为：

所以：

**梯度下降：**

即为学习率，即为梯度。因此对进行随机梯度下降：

第七章 支持向量机

**1.支持向量机：**

支持向量机（support vector machines，SVM）是一种二类分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划（convex quadratic programming）的问题，也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

**2.线性可分支持向量机：**

**定义7.1（线性可分支持向量机）** 给定线性可分训练数据集，通过间隔最大化或等价地求解相应的凸二次规划问题学习得到的分离超平面为

以及相应的分类决策函数

称为线性可分支持向量机。

**3.几何间隔：**

**定义7.3（几何间隔）** 对于给定的训练数据集和超平面，定义超平面关于样本点的几何间隔为

定义超平面关于训练数据集的几何间隔为超平面关于中所有样本点的几何间隔之最小值，即

**4.线性可分支持向量机的优化方法：**

下面考虑如何求得一个几何间隔最大的分离超平面，即最大间隔分离超平面。具体地，这个问题可以表示为下面的约束最优化问题：

考虑几何间隔和函数间隔的关系式(7.8)，可将这个问题改写为

函数间隔的取值并不影响最优化问题的解。事实上，假设将和按比例改变为和，这时函数间隔成为.函数间隔的这一改变对上面最优化问题的不等式约束没有影响，对目标函数的优化也没有影响，也就是说，它产生一个等价的最优化问题。这样，就可以取。将代入上面的最优化问题，注意到最大化和最小化是等价的，于是就得到下面的线性可分支持向量机学习的最优化问题：

这是一个凸二次规划（convex quadratic programming）问题。

**5.算法7.1：**

**算法7.1（线性可分支持向量机学习算法——最大间隔法）**

输入：线性可分训练数据集，其中，，；

输出：最大间隔分离超平面和分类决策函数。

（1）构造并求解约束最优化问题：

求得最优解，。

（2）由此得到分离超平面：

分类决策函数

**6.学习的对偶算法：**

为了求解线性可分支持向量机的最优化问题(7.13)(7.14)，将它作为原始最优化问题，应用拉格朗日对偶性（参阅附录C），通过求解对偶问题（dual problem）得到原始问题（primal problem）的最优解，这就是线性可分支持向量机的对偶算法（dual algorithm）。这样做的优点，一是对偶问题往往更容易求解；二是自然引入核函数，进而推广到非线性分类问题。

首先构建拉格朗日函数（Lagrange function）。为此，对每一个不等式约束(7.14)引进拉格朗日乘子（Lagrange multiplier），定义拉格朗日函数：

其中，为拉格朗日乘子向量。

根据拉格朗日对偶性，原始问题的对偶问题是极大极小问题：

所以，为了得到对偶问题的解，需要先求对，的极小再求对的极大。

（1）求

将拉格朗日函数分别对，求偏导数并令其等于0。

得

将式(7.19)代入拉格朗日函数(7.18)，并利用式(7.20)，即得

即

（2）求对的极大，即是对偶问题

将式(7.21)的目标函数由求极大转换成求极小，就得到下面与之等价的对偶最优化问题：

**7.算法7.2：**

**算法7.2（线性可分支持向量机学习算法）**

输入：线性可分训练数据集，其中，，；

输出：分离超平面和分类决策函数。

（1）构造并求解约束最优化问题

求得最优解。

（2）计算

并选择的一个正分量，计算

（3）求得分离超平面

分类决策函数

**8.软间隔最大化：**

线性不可分意味着某些样本点不能满足函数间隔大于等于1的约束条件(7.14)。为了解决这个问题，可以对每个样本点引进一个松驰变量，使函数间隔加上松弛变量大于等于1。这样，约束条件变为

同时，对每个松弛变量，支付一个代价。目标函数由原来的变成

这里，称为惩罚参数，一般由应用问题决定，值大时对误分类的惩罚增大，值小时对误分类的惩罚减小。最小化目标函数(7.31)包含两层含义：使尽量小即间隔尽量大，同时使误分类点的个数尽量小，是调和二者的系数。

线性不可分的线性支持向量机的学习问题变成如下凸二次规划（convex quadratic programming）问题（原始问题）：

**9.线性支持向量机学习的对偶算法：**

原始问题(7.32)(7.34)的对偶问题是

原始最优化问题(7.32)(7.34)的拉格朗日函数是

其中，，。

对偶问题是拉格朗日函数的极大极小问题。首先求对，，的极小，由

得

将式(7.41)(7.43)代入式(7.40)，得

再对求的极大，即得对偶问题：

将对偶最优化问题(7.44)(7.48)进行变换：用等式约束(7.46)消去，从而只留下变量，并将约束(7.46)(7.48)写成

再将对目标函数求极大转换为求极小，于是得到对偶问题(7.37)(7.39)。

**10.算法7.3：**

**算法7.3（线性支持向量机学习算法）**

输入：训练数据集，其中，，；

输出：分离超平面和分类决策函数。

（1）选择惩罚参数，构造并求解凸二次规划问题

求得最优解。

（2）计算

选择的一个分量适合条件，计算

（3）求得分离超平面

分类决策函数：

**11.合页损失函数：**

对于线性支持向量机学习来说，其模型为分离超平面及决策函数，其学习策略为软间隔最大化，学习算法为凸二次规划。

线性支持向量机学习还有另外一种解释，就是最小化以下目标函数：

目标函数的第1项是经验损失或经验风险，函数

称为合页损失函数（hinge loss function）。下标“”表示以下取正值的函数。

**12.正定核的充要条件：**

**定理7.5（正定核的充要条件）** 设是对称函数，则为正定核函数的充要条件是对任意，对应的Gram矩阵：

是半正定矩阵。

**13.常用核函数：**

1.多项式核函数（polynomial kernel function）

对应的支持向量机是一个次多项式分类器。在此情形下，分类决策函数成为

2.高斯核函数（Gaussian kernel function）

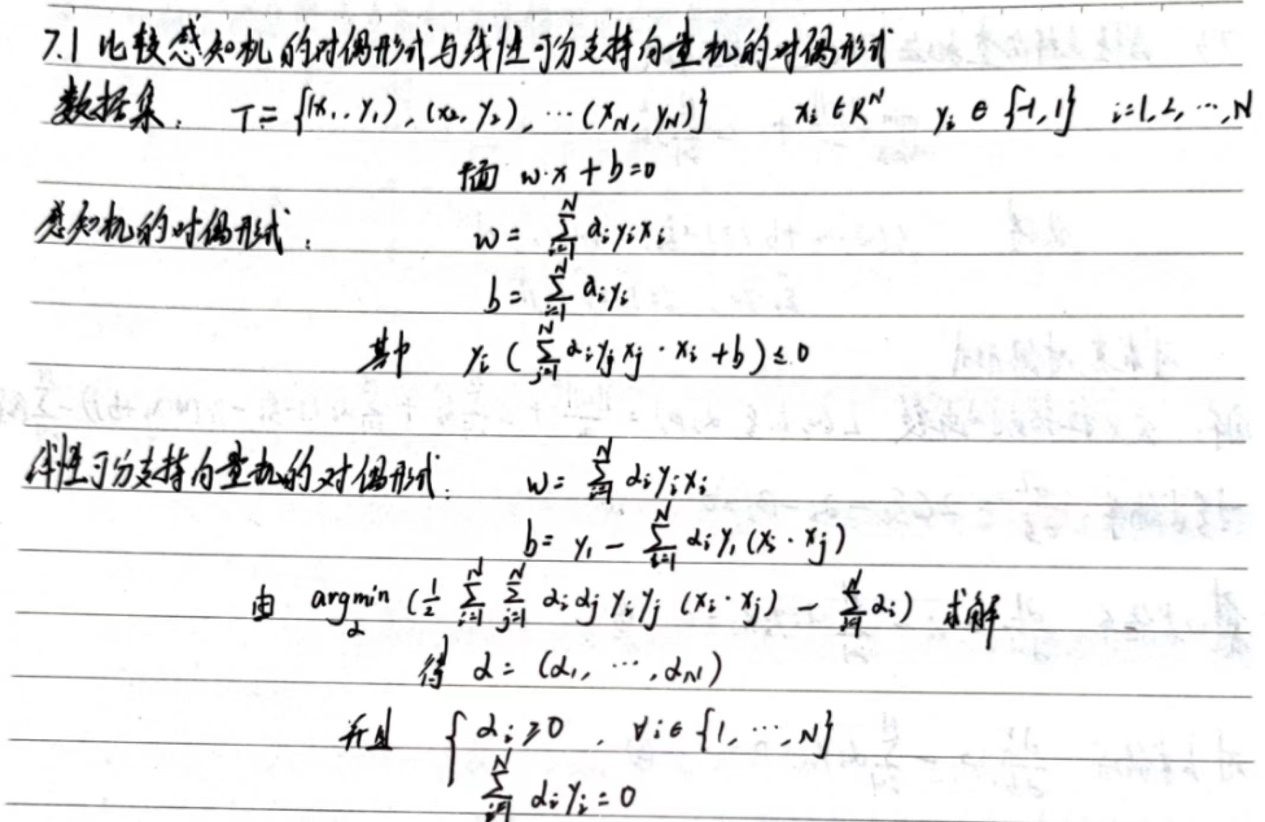
对应的支持向量机是高斯径向基函数（radial basis function）分类器。在此情形下分类决策函数成为

3.字符串核函数（string kernel function）

核函数不仅可以定义在欧氏空间上，还可以定义在离散数据的集合上。比如，字符串核是定义在字符串集合上的核函数。字符串核函数在文本分类、信息检索、生物信息学等方面都有应用。

**14.课后习题7.1：**

比较感知机的对偶形式与线性可分支持向量机的对偶形式。



对于感知机的学习，是基于梯度下降算法的，每次只选取一个数据点在满足条件的情况下进行参数更新，迭代更新。而支持向量机的学习是基于条件约束规划，通过对偶问题来更新，在数据集不大时，可以一次性求得模型的参数的具体表达。支持向量机可以在理论上求得解析解。

**15.课后习题7.3：**

线性支持向量机也可以定义为以下形式：

试求出其对偶形式。

我们先引入拉格朗日函数，其中，

那么原始问题就是

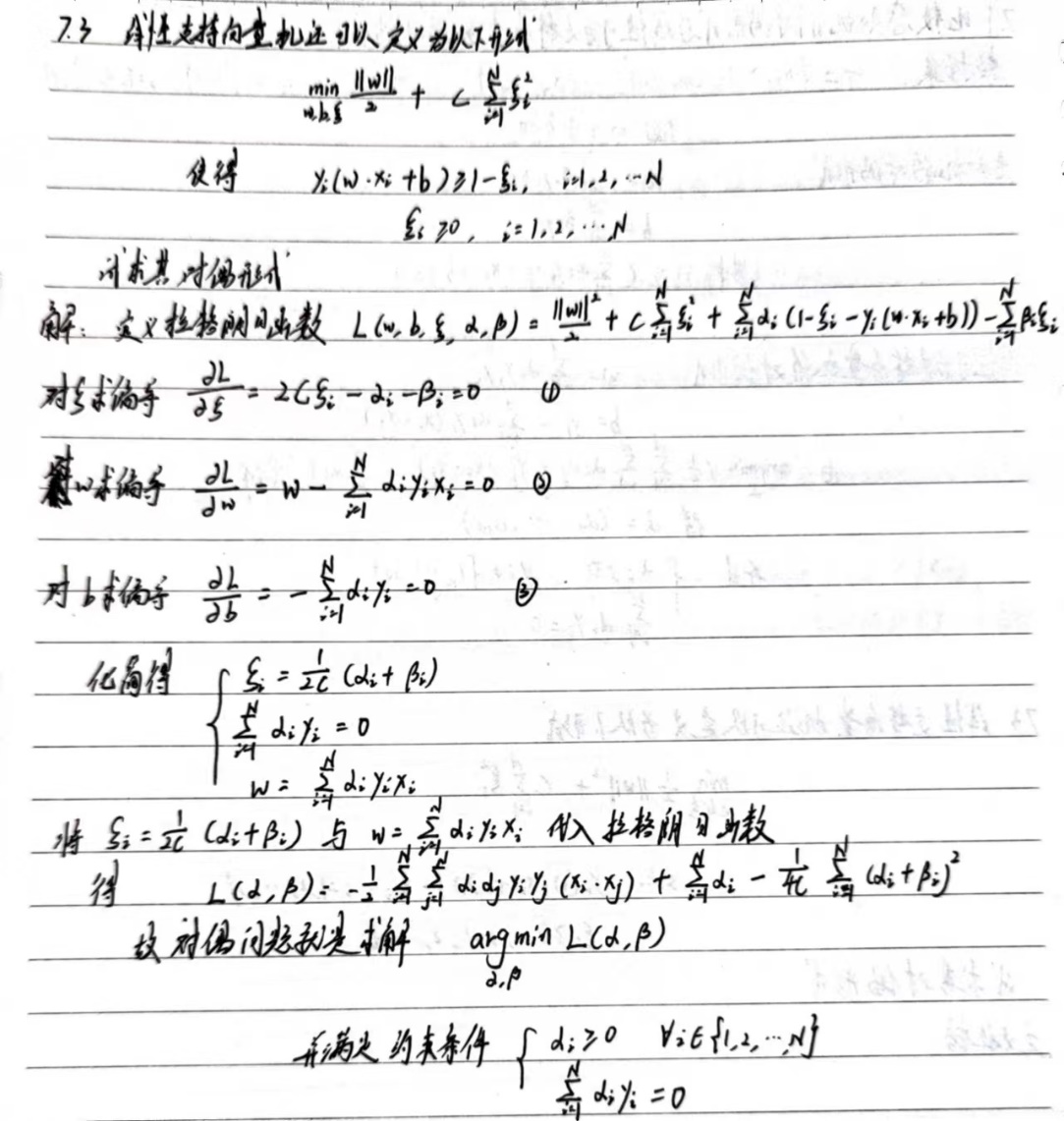
原始问题是先关于，极大化，这是因为，假如有一个数据点不满足条件，也就是，因为我们是先极大化，那么在极大化的时候，就是让对应的趋近于，只要让其他的，均为0，此时的目标函数的值就是。此时

如果是数据点满足约束条件，因为是极大化，就是让所有的，均为0，此时

所以有

因此再对进行一次极小化调整参数，使所有的数据满足约束条件，就可以得到

那么对偶问题就是调整极大化和极小化的顺序，也就是



第八章 提升方法

**1.AdaBoost：**

现在叙述AdaBoost算法。假设给定一个二类分类的训练数据集

其中，每个样本点由实例与标记组成。实例，标记，是实例空间，是标记集合。AdaBoost利用以下算法，从训练数据中学习一系列弱分类器或基本分类器，并将这些弱分类器线性组合成为一个强分类器。

**算法8.1（AdaBoost）**

输入：训练数据集，其中，；弱学习算法；

输出：最终分类器。

（1）初始化训练数据的权值分布

（2）对

（a）使用具有权值分布的训练数据集学习，得到基本分类器

（b）计算在训练数据集上的分类误差率

（c）计算的系数

这里的对数是自然对数。

（d）更新训练数据集的权值分布

这里，是规范化因子

它使成为一个概率分布。

（3）构建基本分类器的线性组合

得到最终分类器

**2.前向分布算法与AdaBoost：**

**定理8.3** AdaBoost算法是前向分步加法算法的特例。这时，模型是由基本分类器组成的加法模型，损失函数是指数函数。

**证明** 前向分步算法学习的是加法模型，当基函数为基本分类器时，该加法模型等价于AdaBoost的最终分类器

由基本分类器及其系数组成,。前向分步算法逐一学习基函数，这一过程与AdaBoost算法逐一学习基本分类器的过程一致。下面证明前向分步算法的损失函数是指数损失函数（exponential loss function）

时，其学习的具体操作等价于 AdaBoost 算法学习的具体操作。

**3.习题8.2：**

比较支持向量机、AdaBoost、逻辑斯谛回归模型的学习策略与算法。

**支持向量机的学习策略与算法：**

学习策略：结构风险极小化

算法：最大间隔方法、对偶算法、软间隔算法

**AdaBoost模型的策略和算法：**

学习策略：经验风险极小化

学习算法：提升树算法等

**逻辑斯谛回归的学习策略和算法：**

学习策略：经验风险极小化

算法：极大似然估计法

第九章 EM算法及其推广

**1.三硬币模型：**

**例9.1（三硬币模型）** 假设有3枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是，和。进行如下掷硬币试验：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后掷选出的硬币，掷硬币的结果，出现正面记作1，出现反面记作0；独立地重复次试验（这里，）观测结果如下：

假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数。

**解** 三银币模型可以写作

这里，随机变量是观测变量，表示一次试验观测的结果是1或0；随机变量是隐变量，表示未观测到的掷硬币A的结果；是模型参数。这一模型是以上数据的生成模型。注意，随机变量的数据可以观测，随机变量的数据不可观测。

将观测数据表示为，未观测数据表示为则观测数据的似然函数为

即

考虑求模型参数的极大似然估计，即

这个问题没有解析解，只有通过迭代的方法求解。EM算法就是可以用于求解这个问题的一种迭代算法。下面给出针对以上问题的EM算法，其推导过程省略。

EM算法首先选取参数的初值，记作，然后通过下面的步骤迭代计算参数的估计值，直至收敛为止。第次迭代参数的估计值为。EM算法的第次选代如下。

E步：计算在模型参数，，下观测数据来自掷硬币B的概率

M步：计算模型参数的新估计值

进行数值计算。假设模型参数的初值取为

由式(9.5)，对与均有。

利用迭代公式(9.6)公式(9.8)，得到

由式(9.5)，

继续迭代，得

于是得到模型参数的极大似然估计：

表示硬币A是均匀的，这一结果容易理解。

如果取初值，，，那得到的模型参数的极大似然估计是，，。这就是说，EM算法与初值的选择有关，选择不同的初值可能得到不同的参数估计值。

**2.算法9.1EM算法：**

**算法9.1（EM算法）**

输入：观测变量数据，隐变量数据，联合分布，条件分布；

输出：模型参数。

（1）选择参数的初值，开始迭代；

（2）E步：记为第次迭代参数的估计值，在第次迭代的E步，计算

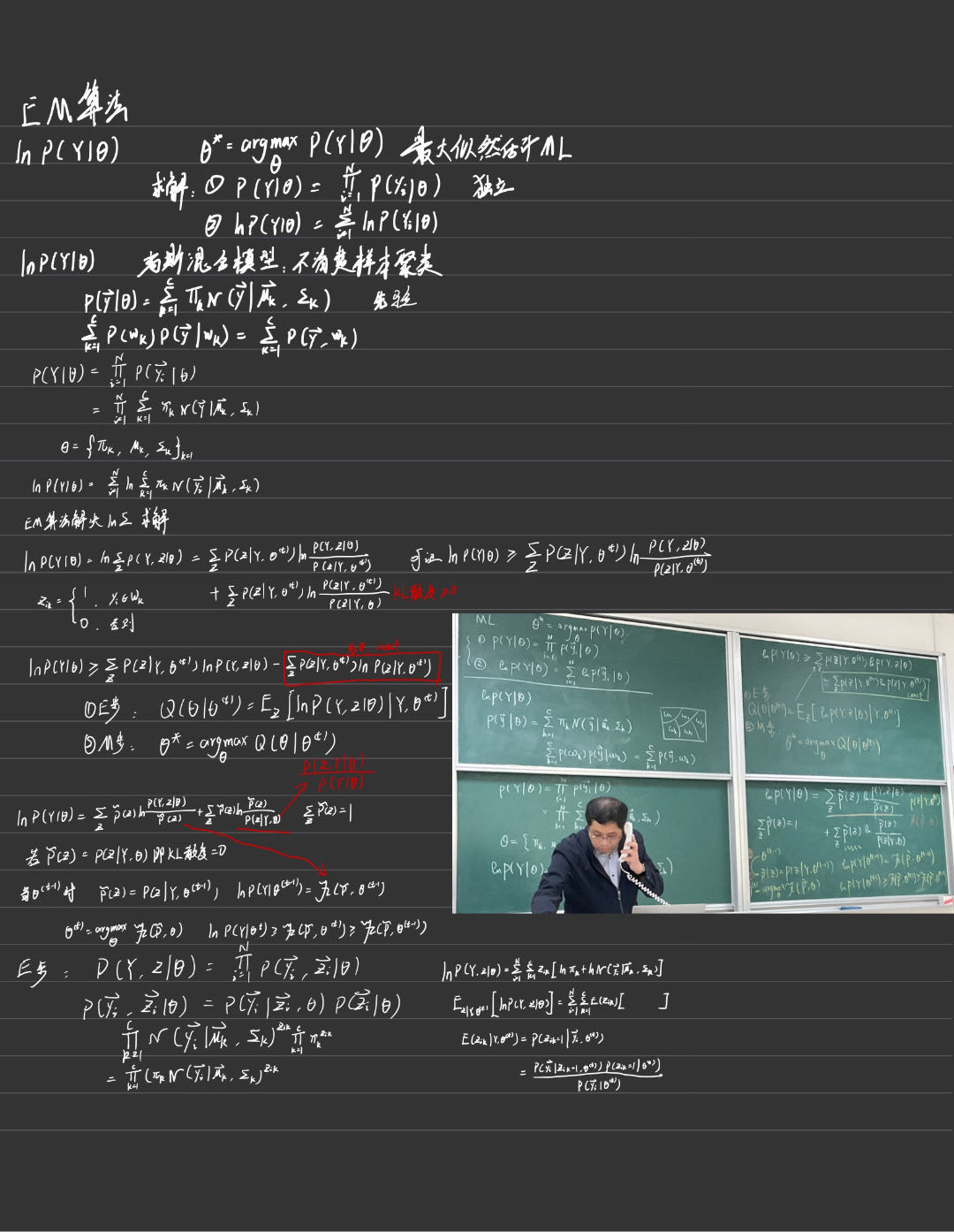
这里，是在给定观测数据和当前的参数估计下隐变量数据的条件概率分布；

（3）M步：求使极大化的，确定第次迭代的参数的估计值

（4）重复第（2）步和第（3）步，直到收敛。

**3.EM算法推导：**

EM算法推导，证明分解



第十四章 聚类方法

**1.闵可夫斯基距离：**

在聚类中，可以将样本集合看作是向量空间中点的集合，以该空间的距离表示样本之间的相似度。常用的距离有闵可夫斯基距离，特别是欧氏距离。闵可夫斯基距离越大相似度越小，距离越小相似度越大。

**定义14.1** 给定样本集合，是维实数向量空间中点的集合，其中，，，样本与样本的闵可夫斯基距离（Minkowski distance）定义为

这里。当时称为欧氏距离（Euclidean distance），即

当时称为曼哈顿距离（Manhattan distance），即

当时称为切比雪夫距离（Chebyshev distance），取各个坐标数值差的绝对值的最大值，即

**2.马哈拉诺比斯距离：**

马哈拉诺比斯距离（Mahalanobis distance），简称马氏距离，也是另一种常用的相似度，考虑各个分量（特征）之间的相关性并与各个分量的尺度无关。马哈拉诺比斯距离越大相似度越小，距离越小相似度越大。

**定义14.2** 给定一个样本集合，，其协方差矩阵记作。样本与样本之间的马哈拉诺比斯距离定义为

其中

当为单位矩阵时，即样本数据的各个分量互相独立且各个分量的方差为1时，由式(14.6) 知马氏距离就是欧氏距离，所以马氏距离是欧氏距离的推广。

**3.相关系数：**

样本之间的相似度也可以用相关系数（correlation coefhcient）来表示。相关系数的绝对值越接近于1，表示样本越相似；越接近于0，表示样本越不相似。

**定义14.3** 样本与样本之间的相关系数定义为

其中

**4.夹角余弦：**

样本之间的相似度也可以用夹角余弦（cosine）来表示。夹角余弦越接近于1，表示样本越相似；越接近于0，表示样本越不相似。

**定义14.4** 样本与样本之间的夹角余弦定义为

**5.例题14.1：**

**例14.1** 给定5个样本的集合，样本之间的欧氏距离由如下矩阵表示：

其中表示第个样本与第个样本之间的欧氏距离。显然为对称矩阵。应用聚合层次聚类法对这5个样本进行聚类。

**解** （1）首先用5个样本构建5个类，，这样，样本之间的距离也就变成类之间的距离，所以5个类直接的距离矩阵亦为。

（2）由矩阵可以看出，为最小，所以把和合并为一个新类，记作。

（3）计算与，，之间的最短距离，有

又注意到其余两类之间的距离是

显然，最小，所以将与合并成一个新类，记作。

（4）计算与，之间的最短距离，

又注意到

显然，其中最小，所以将与合并成一新类，记作。

（5）将与合并成一个新类，记作，即将全部样本聚成1类，聚类终止。

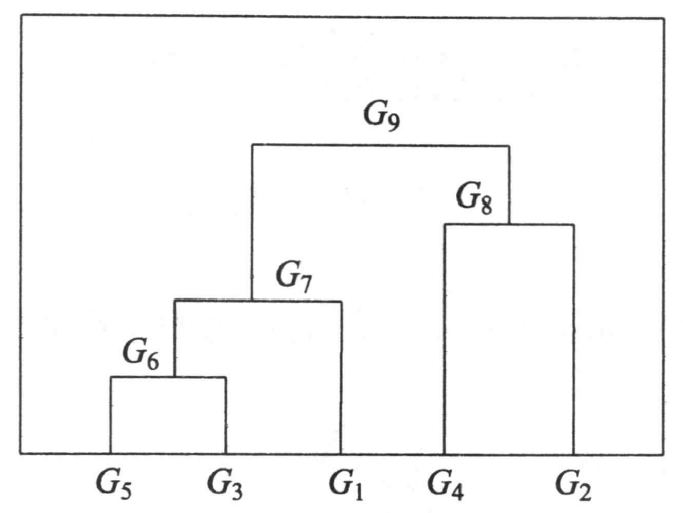


图14.2 层次聚类图

**6.算法14.2（均值聚类算法）：**

输入：个样本的集合；

输出：样本集合的聚类。

（1）初始化。令，随机选择个样本点作为初始聚类中心。

（2）对样本进行聚类。对固定的类中心，其中为类的中想，计算每个样本到类中心的距离，将每个样本指派到与其最近的中心的类中，构成聚类结果。

（3）计算新的类中心。对聚类结果，计算当前各个类中的样本的均值，作为新的类中心。

（4）如果迭代收敛或符合停止条件，输出。

否则，令，返回步（2）。

均值聚类算法的复杂度是，其中是样本维数，是样本个数，是类别个数。

**7.例题14.2：**

**例14.2** 给定含义5个样本的集合

试用均值聚类算法将样本聚到2个类中。

（1）选择两个样本点作为类的中心。假设选择，。

（2）以，为类，的中心，计算，，与，的欧式距离平方。

对，，，将分到类。

对，，，将分到类。

对，，，将分到类。

（3）得到新的类，，计算类的中心，：

（4）重复步骤（2）和步骤（3）。

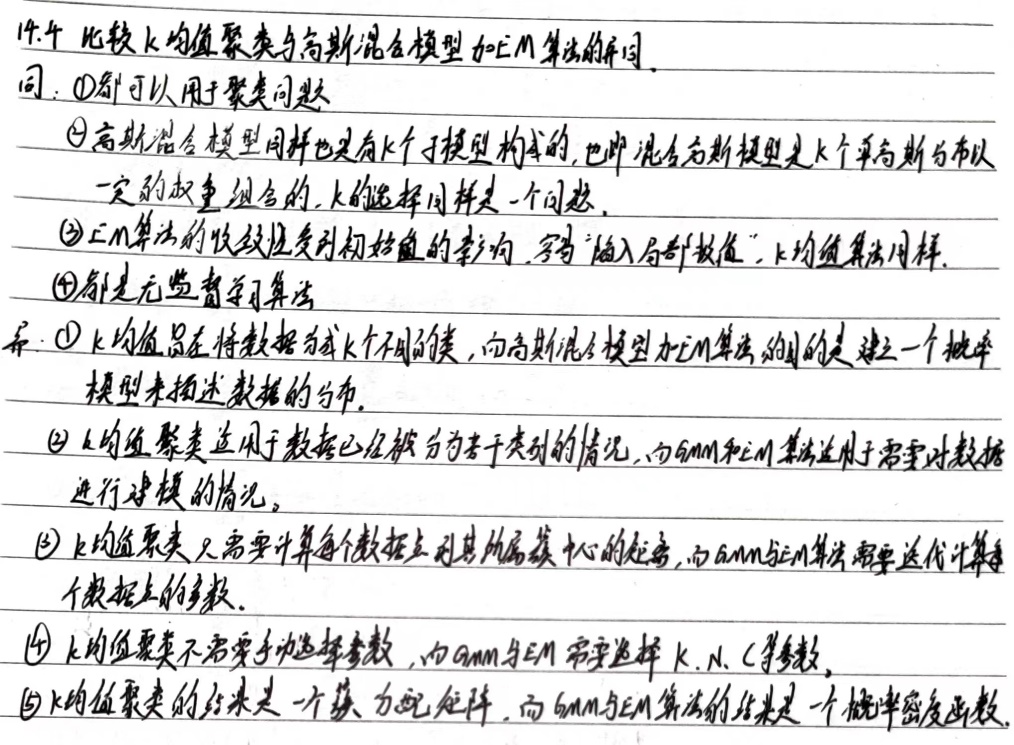
将分到类，将分到类，分到类，分到类，分到类。

得到新的类，。

由于得到的新的类没有改变，聚类停止。得到聚类结果：

**8.课后习题14.4：**

比较均值聚类与高斯混合模型加EM算法的异同。



第十五章 奇异值分解

**1.奇异值分解基本定理：**

**定理15.1（奇异值分解基本定理）** 若为一实矩阵，，则的奇异值分解存在

其中是阶正交矩阵，是阶正交矩阵，是矩形对角矩阵，其对角线元素非负，且按降序排列。

**2.紧奇异值分解与截断奇异值分解：**

1.紧奇异值分解

**定义15.2** 设有实矩阵，其秩为，则称为的紧奇异值分解（compact singular value decomposition），即

其中是矩阵，是矩阵，是阶对角矩阵；矩阵由完全奇异值分解中的前列、矩阵由的前列、矩阵由的前个对角线元素得到。紧奇异值分解的对角矩阵的秩与原始矩阵的秩相等。

2.截断奇异值分解

在矩阵的奇异值分解中，只取最大的个奇异值（，为矩阵的秩）对应的部分，就得到矩阵的截断奇异值分解。实际应用中提到矩阵的奇异值分解时，通常指截断奇异值分解。

**定义15.3** 设为实矩阵，其秩，且，则称为矩阵的截断奇异值分解（truncated singular value decomposition）

其中是矩阵，是矩阵，是阶对角矩阵；矩阵由完全奇异值分解中的前列、矩阵由的前列、矩阵由的前个对角线元素得到。对角矩阵的秩比原始矩阵的秩低。

**3.矩阵的最优近似：**

奇异值分解是在平方损失（弗罗贝尼乌斯范数）意义下对矩阵的最优近似，即数据压缩。

**定理15.2** 设矩阵，矩阵的秩，并设为中所有秩不超过的矩阵集合，，则存在一个秩为的矩阵，使得

称矩阵为矩阵在弗罗贝尼乌斯范数意义下的最优近似。

本书不证明这一定理，将应用这个结果，通过矩阵的奇异值分解求出近似矩阵。

**定理15.3** 设矩阵，矩阵的秩，有奇异值分解，并设为中所有秩不超过的矩阵的集合，，若秩为的矩阵满足

则

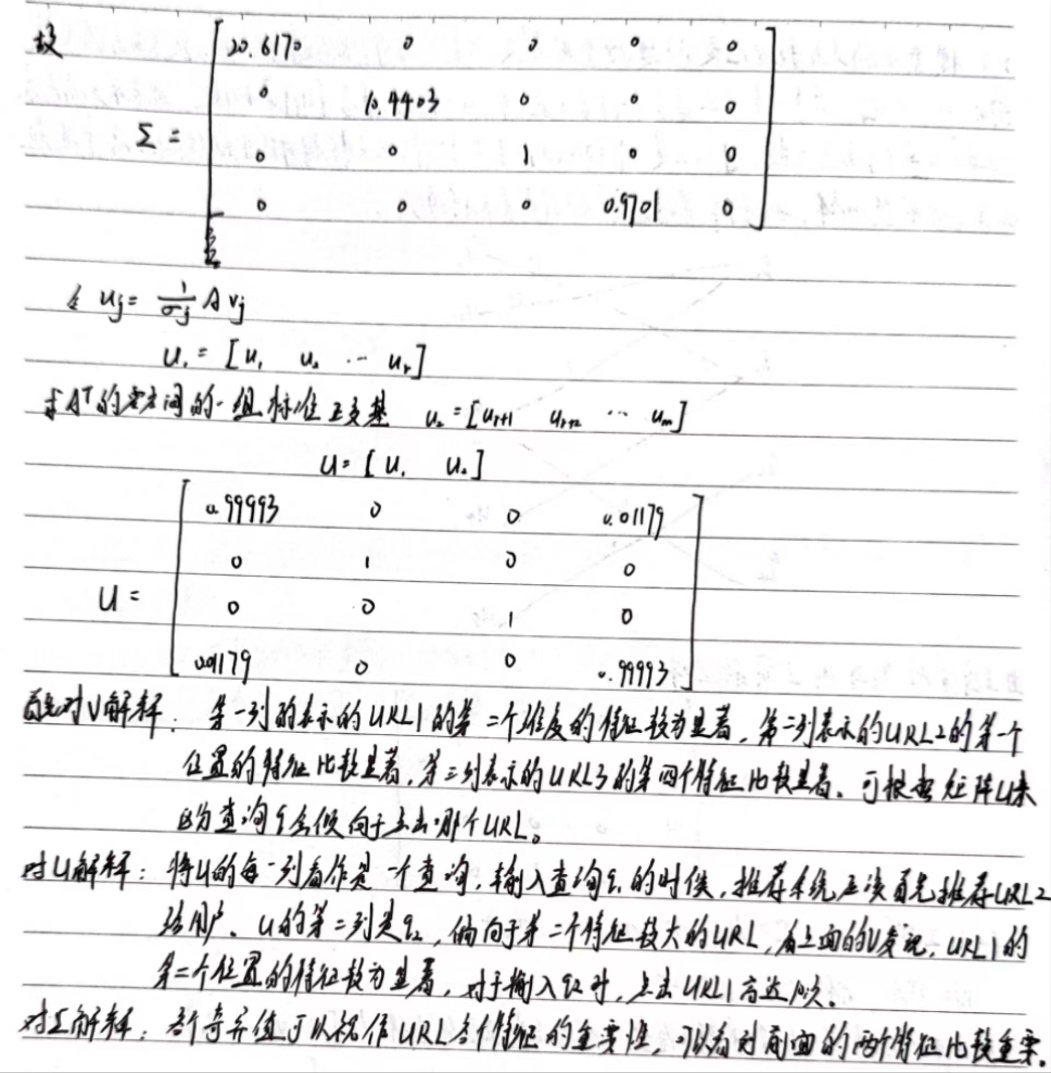
特别地，若，其中

则

**4.课后习题15.5：**

对点击数据进行奇异值分解，并且解释得到的三个矩阵所表示的内容。





第十六章 主成分分析

**1.定理16.1：**

**定理16.1** 设是维随机变量，是的协方差矩阵，的特征值分别是，特征值对应的单位特征向量分别是，则的第主成分是

的第主成分的方差是

即协方差矩阵的第个特征值。

**2.例题16.1：**

**例16.1** 假设有个学生参加四门课程的考试，将学生们的考试成绩看作随机变量的取值，对考试成绩数据进行标准化处理，得到样本相关矩阵，列于表16.1。

表16.1 样本相关矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 课程 | 语文 | 外语 | 数学 | 物理 |
| 语文 |  |  |  |  |
| 外语 |  |  |  |  |
| 数学 |  |  |  |  |
| 物理 |  |  |  |  |

试对数据进行主成分分析。

**解** 设变量分别表示语文、外语、数学、物理的成绩。对样本相关矩阵进行特征值分解，得到相关矩阵的特征值，并按大小排序，

这些特征值就是各主成分的方差贡献率。假设要求主成分的累计方差贡献率大于75%，那么只需取前两个主成分即可，即，因为

求出对应于特征值，的单位特征向量，列于表16.2，表中最后一列为主成分的方差贡献率。

表16.2 单位特征向量和主成分的方差贡献率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 项目 |  |  |  |  | 方差贡献率 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

由此按照式(16.50)可得第一、第二主成分：

这就是主成分分析的结果。变量和表示第一、第二主成分。

接下来由特征值和单位特征向量求出第一、第二主成分的因子负荷量，以及第一、第二主成分对变量的贡献率，列于表16.3。

表16.3 主成分的因子负荷量和贡献率

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 项目 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| 对的贡献率 |  |  |  |  |

从表16.3中可以看出，第一主成分对应的因子负荷量，均为正数，表明各门课程成绩提高都可使提高，也就是说，第一主成分反映了学生的整体成绩。还可以看出，因子负荷量的数值相近，且的数值最大，这表明物理成绩在整体成绩中占最重要位置。

第二主成分对应的因子负荷量，有正有负，正的是语文和外语，负的是数学和物理，表明文科成绩提高都可使提高，而理科成绩提高都可使降低，也就是说，第二主成分反映了学生的文科成绩与理科成绩的关系。

图16.3将原变量，，，（分别表示语文、外语、数学、物理）和主成分，（分别表示整体成绩、文科对理科成绩）的因子负荷量在平面坐标系中表示。可以看出变量之间的关系。4个原变量聚成了两类；因子负荷量相近的语文、外语为一类，数学、物理为一类，前者反映文科课程成绩，后者反映理科课程成绩。

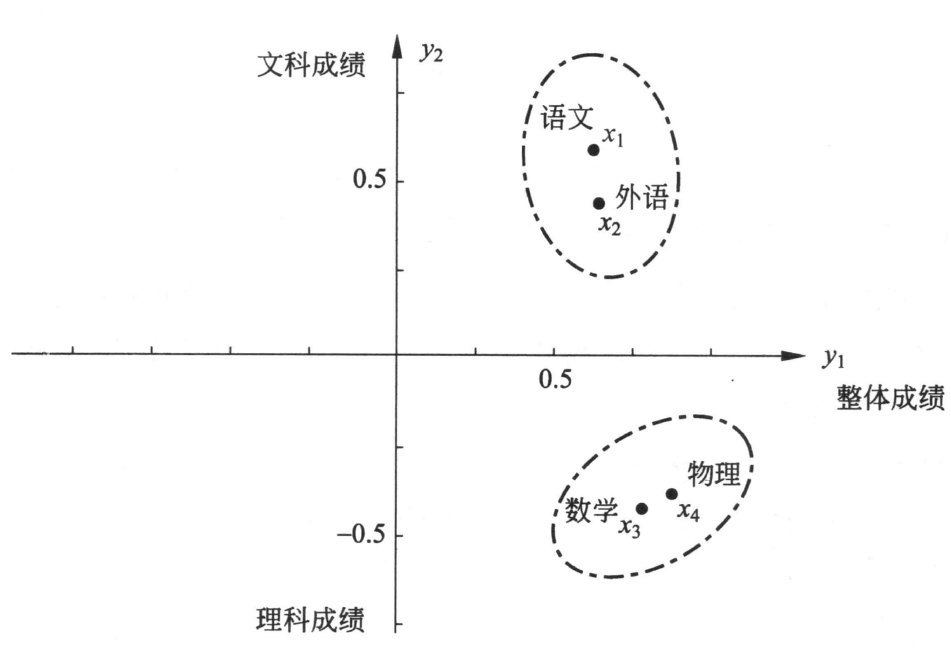


图16.3 因子负荷量的分布图