**第1讲 绪论**

1、（1）对一位去医院做疾病诊断的病人而言，诊断技术的查全率重要还是查准率更重要？给出理由。（2）对不断向你推荐商品的网站，该网站的商品推荐系统的查全率还是查准率你更关注？给出理由。

（1）对一位去医院做疾病诊断的病人而言，诊断技术的查全率（召回率）更重要。理由如下：

查全率（召回率）是指在所有实际患有某种疾病的病人中，被诊断技术正确识别出来的比例。在疾病诊断中，尤其是对于严重或致命的疾病，漏诊（即没有被检测出来的病人）可能会带来严重的后果。即使误诊（即把健康人误认为病人）也会带来一些心理和经济上的负担，但总体来说，漏诊的风险和后果往往比误诊更严重。因此，确保所有患病的病人都能被诊断出来（即高查全率）是至关重要的。

（2）对不断向你推荐商品的网站而言，该网站的商品推荐系统的查准率（精确率）更重要。理由如下：

查准率（精确率）是指在所有被推荐的商品中，真正符合用户兴趣和需求的商品的比例。对于一个商品推荐系统来说，用户希望看到的推荐是与自己的兴趣高度相关的商品。如果推荐的商品大多数都不是用户感兴趣的，那么即使有很多推荐（高查全率），用户体验也会很差，可能导致用户对网站失去信任和兴趣。因此，提高推荐商品的相关性和精确度（即高查准率）是网站推荐系统的关键目标。

1. （1）简述数据模型的泛化能力与过拟合现象这两个概念；说明两者之间存在的一般性关系。（2）借助模型的过拟合和泛化能力背后的原理，分析个人的专业学习与职业发展前景之间的关系（可以结合自己的经验）。

**泛化能力**： 泛化能力是指一个数据模型在遇到新的、未见过的数据时，仍然能表现出良好的预测或分类能力。这意味着模型不仅在训练数据上表现良好，在测试数据或实际应用中的新数据上也能准确预测或分类。

**过拟合现象**： 过拟合是指一个数据模型在训练数据上表现过于优秀，但在新数据上表现较差的现象。这通常发生在模型过于复杂时，模型不仅学习了数据中的规律，还学习了数据中的噪音和偶然性，从而失去了泛化能力。

**一般性关系**： 过拟合和泛化能力之间存在一种负相关的关系。过拟合模型通常泛化能力较差，因为它对训练数据的拟合过于精细，无法很好地适应新的数据。相反，一个具有良好泛化能力的模型通常能够避免过拟合，能够捕捉到数据中的一般规律，而不是噪音和偶然性。因此，训练模型时需要平衡两者，既要有足够的复杂度来捕捉数据中的规律，又不能过于复杂以至于陷入过拟合。

3、试说明：在下面两种情况下，应该建立分类模型还是回归模型？人们最感兴趣的是推断还是预测？建模用的数据个数和特征个数分别是多少？各特征分别是什么？

（1）我们收集了世界500强公司的数据。每个公司都记录了利润、员工数、产业类型和CE0的薪酬。我们希望了解什么因素影响CEO的薪酬。

（2）我们考虑研发一个新产品，希望知道它能否成功。为此收集了先前研发的20个相近的产品数据，对每个产品数据都记录了它是成功的还是失败的，以及产品价格、市场预算、竞品价格和其他10个变量。

4、比较参数模型与非参数模型之间的不同。相对于非参数模型，参数模型的优点与缺点分别是什么？

参数模型和非参数模型是机器学习中两类不同的模型类型，它们在结构、复杂度、灵活性等方面存在显著差异。以下是它们的主要区别及参数模型的优缺点：

**参数模型**

**定义**：参数模型是指具有固定数量参数的模型。这些参数的数量不会随着数据量的增加而改变。模型假设数据符合某种特定的分布形式，通过调整这些固定参数来拟合数据。

**常见示例**：线性回归、逻辑回归、支持向量机（SVM）等。

**优点：**

1. **计算效率高**：由于参数固定，计算量通常较小，训练和预测速度较快。
2. **可解释性强**：参数模型通常比较简单，容易理解和解释。例如，线性回归模型中的系数可以直接表示每个特征对目标变量的影响。
3. **数据需求少**：通常参数模型在数据较少时也能表现良好，不需要大量数据来训练。
4. **泛化能力好**：在数据较少的情况下，参数模型往往能避免过拟合，泛化能力较好。

**缺点：**

1. **灵活性差**：由于假设数据符合某种特定分布，参数模型在面对复杂数据时表现可能不佳，难以捕捉复杂模式。
2. **假设依赖强**：模型依赖于对数据分布的假设，如果假设不符合实际数据分布，模型性能会显著下降。

**非参数模型**

**定义**：非参数模型不对数据分布做特定假设，参数数量随数据量的增加而增加，模型的复杂度可以随数据复杂度增加而增加。

**常见示例**：k-近邻算法（KNN）、决策树、随机森林、支持向量机（非线性核）、神经网络等。

**优点：**

1. **灵活性强**：非参数模型可以适应各种数据分布，能够捕捉复杂的数据模式。
2. **性能优越**：在处理大规模和复杂数据集时，非参数模型通常能提供更高的精度。

**缺点：**

1. **计算效率低**：由于参数数量随数据量增加而增加，训练和预测速度较慢，计算资源消耗大。
2. **可解释性差**：非参数模型往往比较复杂，难以直观理解和解释。例如，深度神经网络中的权重难以解释。
3. **数据需求大**：非参数模型通常需要大量数据才能表现良好，数据量不足时可能导致过拟合或欠拟合。
4. **过拟合风险高**：由于其高灵活性，非参数模型在数据量较少时容易发生过拟合。

**第2讲 线性判别分类**

1、线性回归与逻辑回归在模型设定、目标函数和求解算法上，有何相似和不同之处？

**相似之处**

1. **线性模型**：两者都假设输入特征与输出变量之间存在线性关系（在不同的尺度上）。
2. **目标**：两者都使用参数 w\mathbf{w}w 和 bbb 来进行预测。
3. **梯度下降**：两者都可以使用梯度下降方法来进行参数优化。

**不同之处**

1. **模型设定**：线性回归用于回归问题，逻辑回归用于分类问题。
2. **目标函数**：线性回归最小化均方误差（MSE），逻辑回归最大化对数似然。
3. **求解方法**：线性回归有解析解，而逻辑回归通常需要数值优化方法来求解参数。

2、如果你想将图片分类为户外／室内以及白天／黑夜。你应该实现两个逻辑回归分类器还是一个Softmax 回归分类器？

**建议实现两个逻辑回归分类器**。理由如下：

1. **任务独立性**：户外/室内和白天/黑夜是两个独立的任务，各自有明确的分类目标，使用独立的逻辑回归分类器可以避免一个任务的错误影响另一个任务。
2. **模型简单**：逻辑回归分类器简单易用，适合解决二分类问题，训练和调试相对容易。
3. **特征提取**：在实际应用中，可以先用一个特征提取器（例如卷积神经网络）提取图像特征，然后将这些特征同时输入两个逻辑回归分类器，从而实现特征共享，进一步提高效率。

3、（1）逻辑回归实际上是用样本点的回归函数值的logistic变换值来拟合样本点属于正类的概率。简要说明这种建模方法的合理性和不合理之处。（2）能否用正态随机变量的累积分布函数来替换logistic函数，作为逻辑回归中的概率变换函数？为什么？还能想到其他的可用于逻辑回归的变换函数吗？

#### 合理性：

1. **概率输出**：逻辑回归通过logistic函数（也称为sigmoid函数）将线性回归的输出（实数值）转换为概率值（0到1之间），这非常适合二分类问题中的概率估计。
2. **单调性和可微性**：logistic函数是单调可微的，这意味着输出值会随着输入值的增加而单调变化，且可以方便地进行梯度计算，有助于优化算法（如梯度下降）收敛。
3. **简洁性和解析性**：logistic函数具有简单的数学形式，便于推导和计算，同时在统计学中有广泛的应用和理论支持。
4. **对数几率线性关系**：逻辑回归假设对数几率（log-odds）与特征的线性组合之间存在线性关系，这在许多实际问题中是合理的假设。

#### 不合理之处：

1. **线性可分性假设**：逻辑回归假设特征空间中的类别是线性可分的，尽管通过logistic函数进行非线性变换，但其基本假设仍是线性的。这在某些复杂数据集上可能不成立，影响模型性能。
2. **尾部行为**：logistic函数的尾部变化较慢，这意味着当输入值极大或极小时，概率变化会非常缓慢，可能导致在某些极端情况下模型对这些样本点的判别能力下降。
3. **不适用于多类分类**：标准逻辑回归只能处理二分类问题，对于多类分类问题需要扩展，如使用多项逻辑回归（Softmax回归）。

#### 使用正态随机变量的累积分布函数

正态随机变量的累积分布函数（CDF），通常称为Probit函数，可以用于替换logistic函数。Probit回归与逻辑回归非常相似，只是使用了不同的变换函数。**合理性**：

1. **概率输出**：Probit函数也是单调递增的，输出值范围在0到1之间，适合概率估计。
2. **对数几率线性关系**：Probit模型也假设对数几率与特征的线性组合之间存在线性关系，与逻辑回归类似。

**差异**：

1. **尾部行为**：Probit函数在尾部的变化比logistic函数更快，可能在极端值情况下表现更好。
2. **计算复杂性**：Probit函数的数学形式复杂，需要计算正态分布的CDF，计算复杂度较高，尤其在大规模数据集上。

#### 其他可用于逻辑回归的变换函数

1. **Softmax函数**：用于多类分类问题，Softmax函数将输入向量转换为一组非负数，这些数值和为1，适合用于多类概率估计。
2. **Log-Log函数**：用于极端值处理，适合建模某些特定分布的二分类问题。
3. **Complementary Log-Log函数**：用于处理生存分析中的时间到事件数据，也适用于某些分类问题。

4、一个三类问题，其判别函数为

， ， 

* 1. 设这些函数是在多类情况1条件下确定的，绘出判别界面及每一模式类别的区域。
  2. 设为多类情况2，并使，，，绘出判别界面及每一模式类别的区域。

设，和是在多类情况3的条件下确定的，绘出其判别界面及每一模式类别的区域。

5、简要说明线性回归模型的两种正则化方法（岭回归，Lasso回归）在正则化目标和实际效果等方面的异同。

线性回归模型的正则化方法通过引入额外的约束项来减少模型的复杂度，防止过拟合。岭回归（Ridge Regression）和Lasso回归（Least Absolute Shrinkage and Selection Operator）是两种常见的正则化方法。它们在正则化目标和实际效果等方面存在一些相同点和不同点。

### 岭回归（Ridge Regression）

**正则化目标**：

* 岭回归在损失函数中加入了权重参数的平方和作为惩罚项，其目标函数为： J(w)=∑i=1n(yi−wTxi)2+λ∑j=1pwj2J(\mathbf{w}) = \sum\_{i=1}^{n} (y\_i - \mathbf{w}^T \mathbf{x}\_i)^2 + \lambda \sum\_{j=1}^{p} w\_j^2J(w)=i=1∑n​(yi​−wTxi​)2+λj=1∑p​wj2​ 其中，λ\lambdaλ 是正则化参数，控制惩罚项的权重，nnn 是样本数量，ppp 是特征数量，w\mathbf{w}w 是回归系数。

**实际效果**：

* **缩小系数值**：通过引入系数的平方和惩罚项，迫使某些系数变小，但不会使它们变为零。
* **稳定性**：对于多重共线性问题，岭回归能够提高模型的稳定性和预测能力。
* **保持所有特征**：因为不会将任何系数完全压缩到零，岭回归保持了所有输入特征。

### Lasso回归（Lasso Regression）

**正则化目标**：

* Lasso回归在损失函数中加入了权重参数的绝对值和作为惩罚项，其目标函数为： J(w)=∑i=1n(yi−wTxi)2+λ∑j=1p∣wj∣J(\mathbf{w}) = \sum\_{i=1}^{n} (y\_i - \mathbf{w}^T \mathbf{x}\_i)^2 + \lambda \sum\_{j=1}^{p} |w\_j|J(w)=i=1∑n​(yi​−wTxi​)2+λj=1∑p​∣wj​∣ 其中，λ\lambdaλ 是正则化参数，控制惩罚项的权重。

**实际效果**：

* **系数稀疏性**：通过引入系数的绝对值和惩罚项，Lasso回归可以将一些系数缩小到零，从而实现特征选择。
* **特征选择**：由于Lasso回归能够将不重要的特征的系数压缩到零，导致在模型中只保留那些重要的特征，有利于解释性和简化模型。
* **适用于高维数据**：对于高维数据，Lasso回归能有效处理，并自动选择特征，适合用于特征数量大于样本数量的情况。

### 相同点

* **目的**：两者都旨在通过增加惩罚项来减少模型的复杂度，防止过拟合。
* **超参数**：两者都引入了正则化参数 λ\lambdaλ，需要通过交叉验证等方法进行选择。
* **改进稳健性**：在多重共线性情况下，两者都可以改进模型的稳健性和预测性能。

### 不同点

* **正则化项**：
  + 岭回归使用的是二次惩罚（L2正则化），即权重参数的平方和。
  + Lasso回归使用的是一阶惩罚（L1正则化），即权重参数的绝对值和。
* **系数稀疏性**：
  + 岭回归通常不会将系数压缩到零，而是让系数变小。
  + Lasso回归可以将某些系数缩小到零，实现特征选择。
* **特征选择**：
  + 岭回归不执行特征选择，保留所有特征。
  + Lasso回归能够进行特征选择，通过零化不重要的系数简化模型。

### 总结

岭回归和Lasso回归都是常用的正则化方法，用于防止模型过拟合。岭回归更适合在存在多重共线性的情况下使用，而Lasso回归则通过特征选择来简化模型。具体选择哪种方法应根据数据特性和具体需求来决定。

6、设回归变量和响应变量均为标量，拟使用三个样本，训练一个简单线性回归模型 。请利用正则方程计算模型参数。



**第3讲 概率分类**

1、就分类而言，在模型训练前，对一个样本所来自的总体的类别变量分布的先验知识，与在模型训练后对该类别变量分布的后验知识，主要区别是什么？在什么意义下，两种是一致的？

### 先验知识（Prior Knowledge）

先验知识是在进行任何观察或模型训练之前，关于类别变量分布的预先假定或已知信息。先验知识通常来源于领域专家的经验、历史数据或其他先验假设。数学上，先验知识表示为先验概率分布 P(C)P(C)P(C)，其中 CCC 是类别变量。

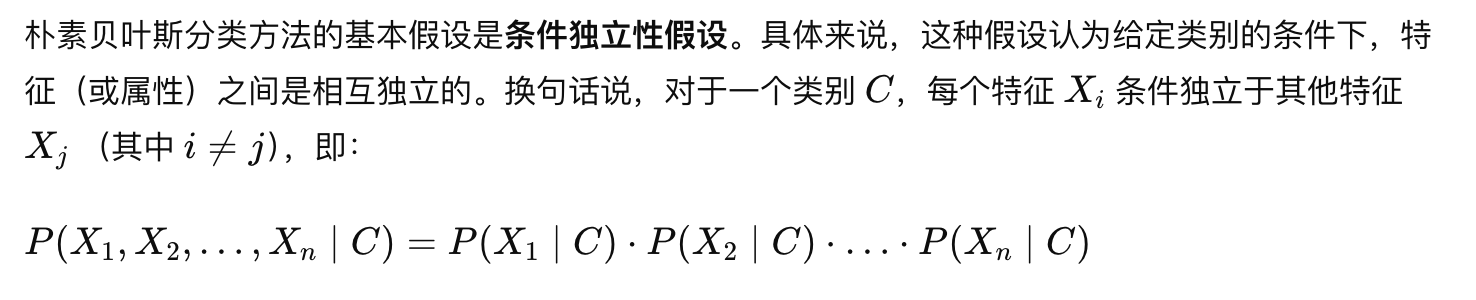
### 后验知识（Posterior Knowledge）

后验知识是在对数据进行观察和模型训练之后，更新的关于类别变量分布的信息。后验知识结合了先验知识和新观察到的数据。数学上，后验知识表示为后验概率分布 P(C∣D)P(C \mid D)P(C∣D)，其中 DDD 是观察到的数据。

先验知识和后验知识在以下意义下是一致的：

1. **相同数据下的逐渐收敛**：如果使用相同的先验知识并连续进行多次观察和更新，在数据量非常大的情况下，后验分布通常会逐渐收敛，且对初始先验的依赖性会减小。此时，后验分布主要由数据本身决定，而先验影响变得微不足道。
2. **充分数据**：在数据充分的情况下，不同的合理先验知识会得到相似的后验分布。换句话说，随着数据量的增加，先验对后验的影响会变小，后验更多地反映数据本身的信息。

2、朴素贝叶斯分类方法的基本假设是什么？试举一例说明该假设的合理性。



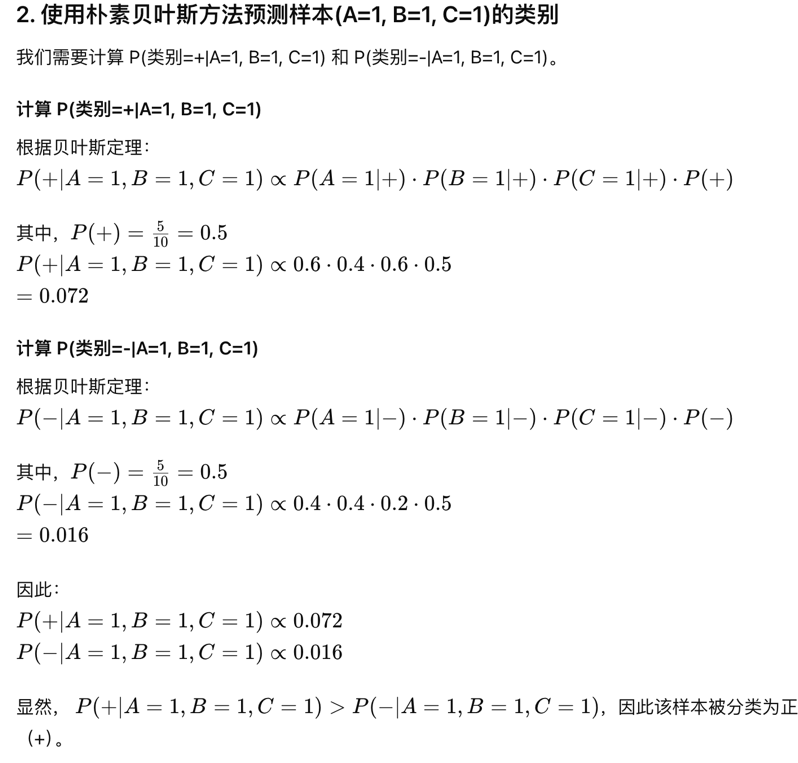
3、假设在某个地区的疾病普查中，异常细胞（）和正常细胞（）的先验概率分别为，。现有一待识别细胞，其观察值为*X*，从类概率密度分布曲线上查得，试对该细胞利用最小错误率贝叶斯决策规则进行分类。

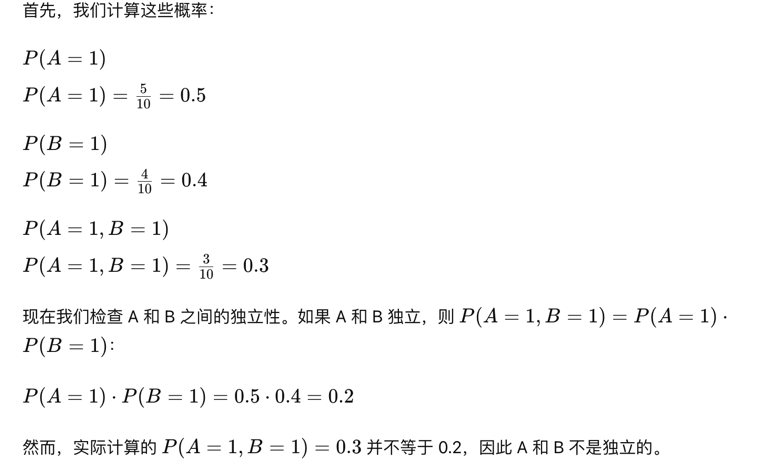
4、对前一题中两类细胞的分类问题（异常细胞，正常细胞），除已知的数据外，若损失函数的值分别为，，，，试用最小风险贝叶斯决策规则对细胞进行分类。

5、考虑下表中的数据集。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 样本序号 | A | B | C | 类别 |
| 1 | 0 | 0 | 1 | - |
| 2 | 1 | 0 | 1 | + |
| 3 | 0 | 1 | 0 | - |
| 4 | 1 | 0 | 0 | - |
| 5 | 1 | 0 | 1 | + |
| 6 | 0 | 0 | 1 | + |
| 7 | 1 | 1 | 0 | - |
| 8 | 0 | 0 | 0 | - |
| 9 | 0 | 1 | 0 | + |
| 10 | 1 | 1 | 1 | + |

（1）估计以下条件概率

（2）根据估计的条件概率，使用朴素贝叶斯方法预测样本的类别。

（3）比较陈述变量*A、B*之间的统计关系。

（4）比较，给定类＋，变量*A、B*条件独立吗？

**第4讲 支持向量机**

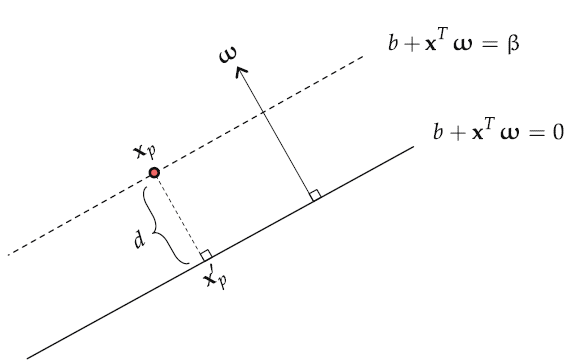
1、简要说明支持向量机技术背后的基本思想，并说明软间隔的具体含义。

支持向量机的基本思想是拟合类别之间可能存在的、最宽的 缓冲区，其目的是使决策边界之间的间隔最大化，从而分隔出两个 类别的训练实例。SVM 执行软间隔分类时，实际上是在完美分类和 拟合最宽的街道之间进行妥协(也就是允许少数实例最终还是落在 街道上)。

2、简要说明核函数在非线性支持向量机中的作用。

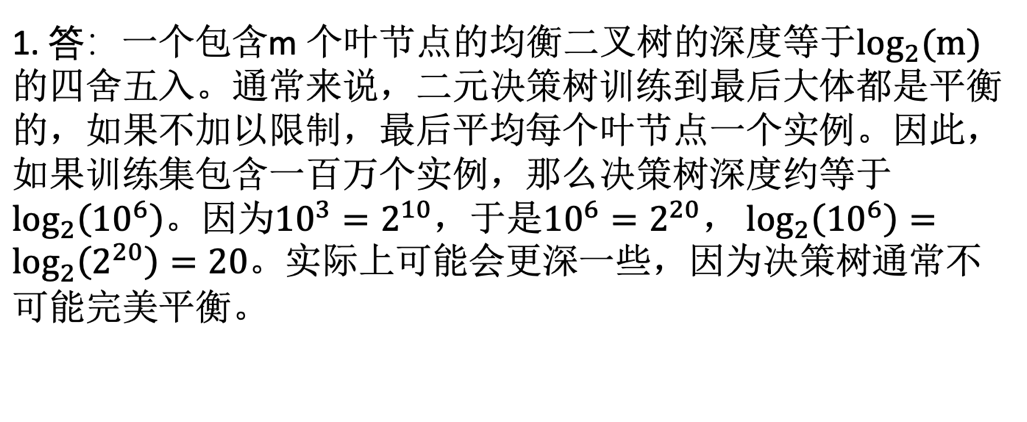
为了解决原始特征空间中的非线性可分问题，需要通过非线性特征变换，把原始特征映射到高维空间，使得在高维空间中近似 线性可分。但是直接计算变换后的高维特征矢量的内积，计算量太 大，因此通过核函数，可以巧妙地解决这个问题:核函数能通过低 计算量求值操作，来近似代替高维矢量的内积。因此，通过核函数 隐含的非线性特征变换，结合线性支持向量机算法，就能把原空间 中非线性可分的数据进行很好的分类。

3、试证：位于超平面中的点到超平面的代数距离为。

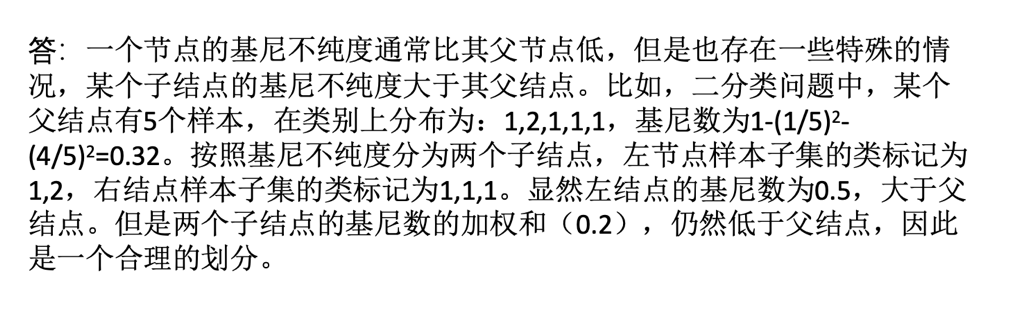


**第5讲 决策树与随机森林**

1、如果训练集有100 万个实例，对叶节点样本数和数的深度均不加约束，则训练一个特征均为二值变量、平衡的二叉决策树，树的大致深度是多少？



2、CART算法按照基尼不纯度（基尼指数）进行结点划分，通常来说，子节点的基尼不纯度是高于还是低于其父节点？有没有可能，某个子结点的基尼不纯度高于其父结点的？给出算例说明你的结论。

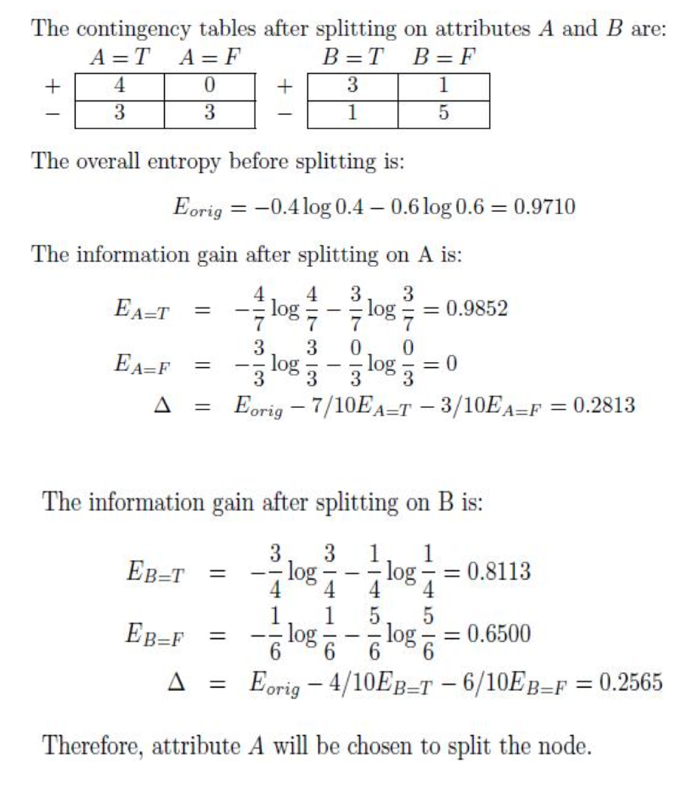


3、如果已经在完全相同的训练集上训练了五个不同的学习器，并且它们都达到了95%的准确率，是否还有机会通过结合这些学习器来获得更好的结果？如果可以，该怎么做？如果不行，为什么？

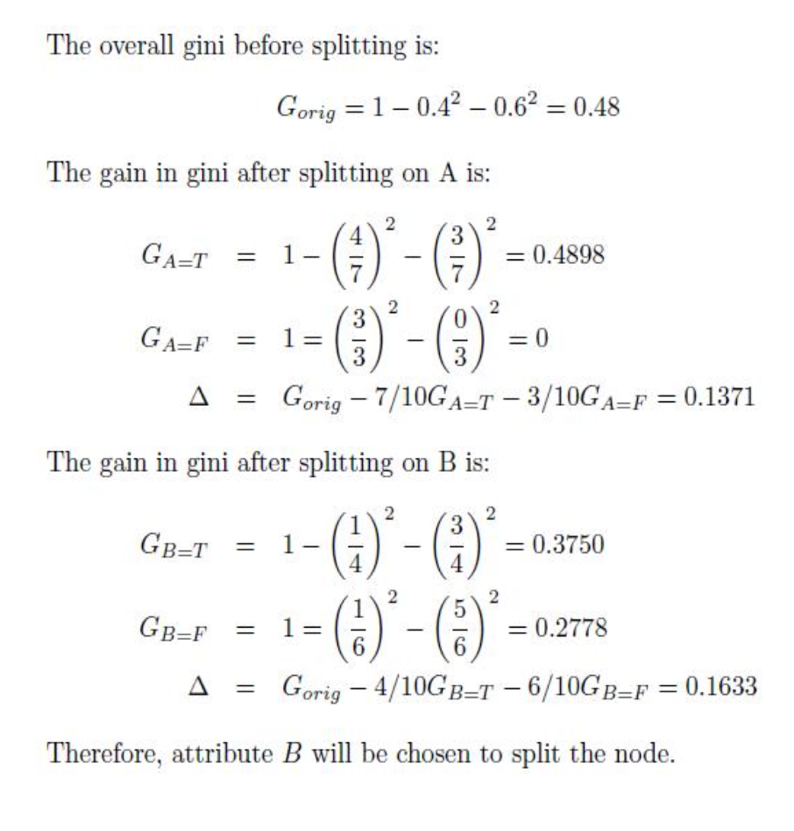
如果你已经训练了五个不同的学习器，并且都达到了95% 的精度，你可以尝试将它们组合成一个投票集成，这通常会带 来更好的结果。如果学习器模型之间非常不同(例如，一个 SVM 分类器， 一个决策树分类器，以及一个Logistic 回归分类 器等)，则效果更优。如果它们是在不同的训练样本集(这是 bagging 集成的关键点)上完成训练，那就更好了;如果不是， 只要学习器模型非常不同，这个集成仍然有效

4、考虑如下二分类问题的数据集。

|  |  |  |  |
| --- | --- | --- | --- |
| 样本序号 | A | B | 类别 |
| 1 | T | F | + |
| 2 | T | T | + |
| 3 | T | T | + |
| 4 | T | F | - |
| 5 | T | T | + |
| 6 | F | F | - |
| 7 | F | F | - |
| 8 | F | F | - |
| 9 | T | T | - |
| 10 | T | F | - |

（1）分别计算按照属性A和属性B划分时的信息增益。使用信息增益准则的决策树分类算法应该使用哪个属性？

（2）分别计算按照属性A和属性B划分时的基尼指数。使用基尼指数的决策树分类算法应该使用哪个属性？

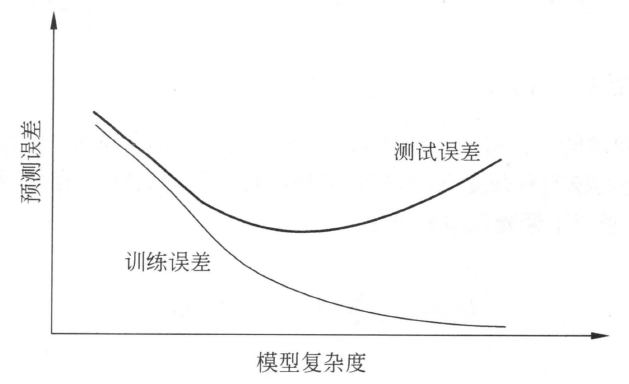


**第6讲 模型选择**

1、如果用测试集来调优超参数，会出现什么情况？

如果多次使用测试集来调整模型的超参数，最终选择表现最好的超参数，可能会导致模型在测试集上过拟合。因为测试集在模型选择的过程中已经被使用过多次，最终选择的模型可能会在测试集上表现比实际应用场景中更好，从而无法准确评估模型的泛化能力。

2、指出下图验证曲线中模型的欠拟合区、过拟合区。如何找到最佳复杂度模型？



3、考虑一个包含50个正例和50个负例的样本集集，样本点特征取值是完全随机的，不包含与类别标签相关的有效信息。因此，在该数据上学习的任何分类模型的泛化错误率预计均为0.5。考虑一个分类器，其分类判决规则是：将训练集的多数类标签判给任何测试样本，不管测试样本的特征值如何；正类作为判决的缺省类（即：如果训练集中两类标签的样本数相同，则判为正标签）。这种方法叫作“多数类诱导的分类器”。使用以下方法确定该分类器的验证错误率。

（1）留一法。

（2）2折分层交叉验证法（分层，指每折中各个类的样本比例与原样本集的相同）。

* 1. 从前面两问的结果中，关于这两种验证方案，可以得出什么结论？

### (1) 留一法

留一法是一种交叉验证方法，适用于小样本数据集。对于你描述的样本集（50个正例和50个负例），留一法将每个样本依次作为验证集，其余样本作为训练集。对于多数类诱导的分类器来说，它的分类判决规则是将训练集的多数类标签（即50个负例）判给任何测试样本，因此无论验证集中的样本是正例还是负例，分类器都会将其预测为负例。

因此，在留一法中，每次验证的错误率都为50%：

* 对于每个正例样本来说，分类器将其错误地预测为负例，错误率为50%。
* 对于每个负例样本来说，分类器正确地预测为负例，错误率为0%。

总的验证错误率为 50%×50+0%×50100=25%\frac{50\% \times 50 + 0\% \times 50}{100} = 25\%10050%×50+0%×50​=25%。

### (2) 2折分层交叉验证法

2折分层交叉验证将数据集分为两部分，每部分中正例和负例的比例与原样本集相同。每一次验证时，其中一部分作为验证集，另一部分作为训练集。由于数据集是完全随机的，每部分中的正例和负例数量相等。

在第一次验证中，训练集中有50个正例和50个负例，多数类诱导的分类器将所有样本判为负例。因此，验证错误率为50%。 在第二次验证中，情况相同，验证错误率也为50%。

总的验证错误率为 50%+50%2=50%\frac{50\% + 50\%}{2} = 50\%250%+50%​=50%。

从上述两种验证方法的结果可以得出以下结论：

* **验证错误率高：** 无论是留一法还是2折交叉验证，由于多数类诱导的分类器的特性（总是预测为多数类，即负例），其验证错误率都较高。这是因为对于每个正例样本，分类器都会错误地将其预测为负例，导致错误率较高。
* **分类器特性影响验证结果：** 验证错误率的高低直接受到分类器特性的影响，即使在不同的验证方法下（留一法和2折交叉验证），由于分类器的判决规则不变，验证错误率的结果也是一致的。

综上所述，多数类诱导的分类器在这种随机数据集上，无论采用什么样的验证方法，其验证错误率都会非常高，这反映了分类器无法从数据中学到任何有效信息的情况。

**第7讲 特征工程**

1、（1）为什么在机器学习与模式识别应用中，需要尽可能进行特征提取？（2）能用特征提取方法进行有效的数据降维，需要数据有什么样的特点？

(1)减少特征数量，有利于降低模型复杂度;对特征之间存 在共线性的数据，通过有效的特征提取，能降低模型求解 的困难 (改善求解的病态条件)。

(2)特征之间存在较大相关性 (多重共线性)，是特征提取得 以有效进行的基础。

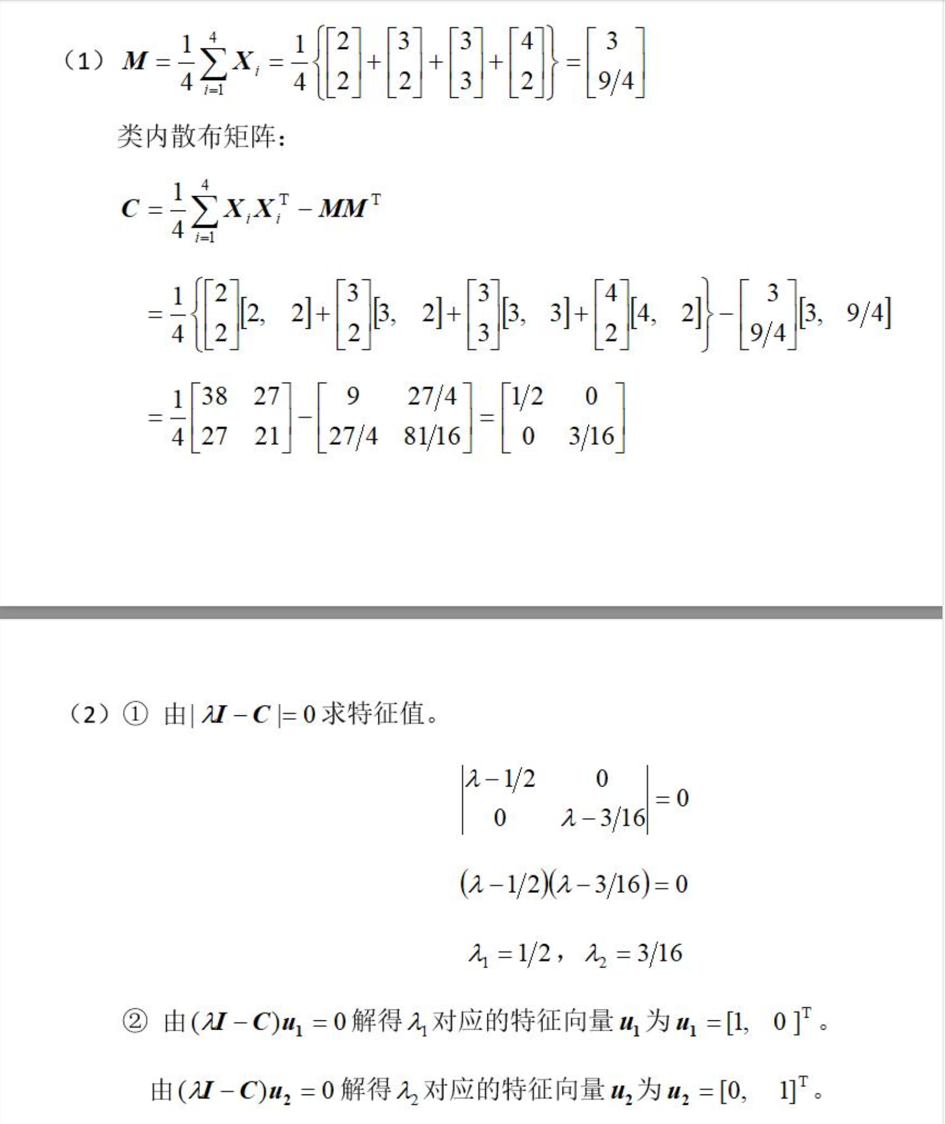
2、简要说明KL变换（PCA）背后的基本思想，并简要说明实现KL变换的大致流程。

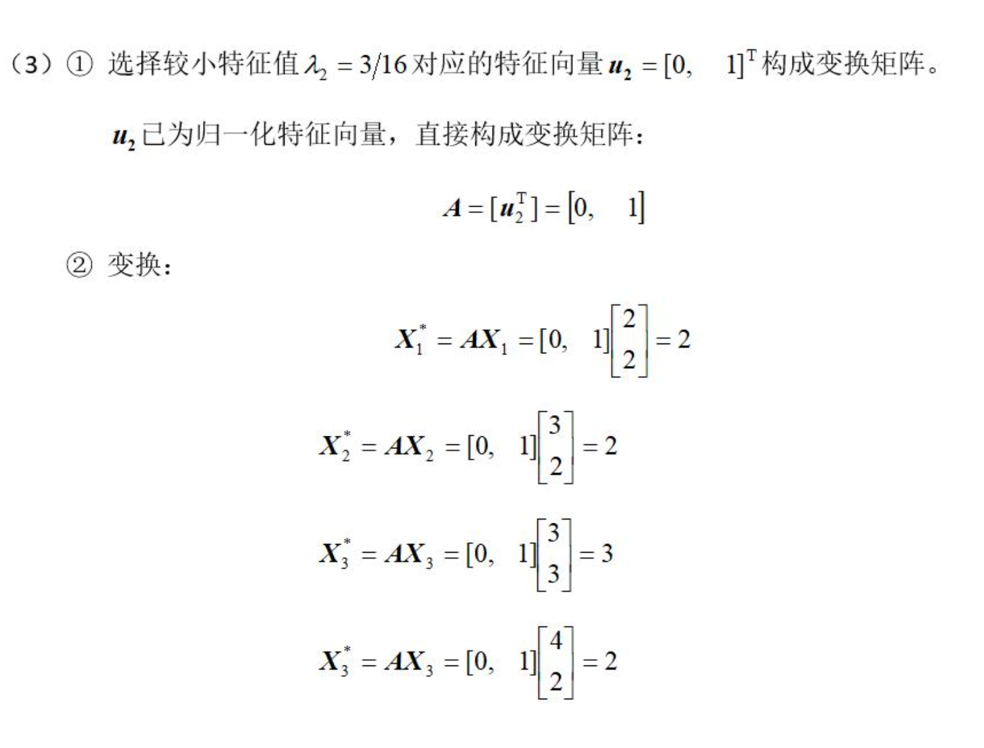
3、假定类的样本集为，它们分别为

， ， ，

1. 求类内散布矩阵；
2. 求类内散布矩阵的特征值和对应的特征向量；

求变换矩阵*A*，将二维模式变换为一维模式。





**第8讲 聚类分析**

1、简述K-均值聚类方法的优缺点。查阅资料，谈一谈可以从哪些方面对K均值进行改进。

### 优点

1. **简单易懂**：K-均值算法直观且易于理解，实现和应用都比较简单。
2. **效率高**：对于大规模数据集，K-均值算法相对较快，尤其是小规模的 K 值。
3. **可扩展性好**：适用于大数据集和高维数据，可以通过并行计算加速。
4. **收敛性好**：在大多数情况下，算法能够快速收敛到局部最优解。

### 缺点

1. **需要预先指定K值**：必须在运行算法前指定聚类数K，这对于未知数据分布是个挑战。
2. **对初始值敏感**：初始中心点的选择会显著影响结果和收敛速度，可能会导致不同的聚类结果。
3. **容易陷入局部最优**：算法可能会陷入局部最优，而不是全局最优。
4. **对噪声和异常值敏感**：噪声和异常值会显著影响聚类结果，因为它们可能会形成独立的簇。
5. **只适用于凸形簇**：K-均值假设簇是球状或凸形的，因此不适用于非凸形状的簇结构。
6. **不适合类别大小差异大的情况**：算法假设每个簇的大小差不多，对于簇大小差异较大的情况效果不佳。

 **改进初始化方法**：

* **K-means++**：改进初始化点选择方法，通过一种概率分布选择初始中心，减少K-均值对初始值的敏感性和局部最优问题。

 **自动确定K值**：

* **肘部法（Elbow Method）**：通过计算不同K值下的总内聚度（簇内距离平方和）并绘制曲线，根据“肘部”位置确定最优K值。
* **轮廓系数（Silhouette Coefficient）**：通过计算不同K值下的平均轮廓系数来评估聚类质量，从而选择最优K值。

 **处理异常值**：

* **使用预处理方法**：在聚类之前进行数据预处理，去除或减少异常值对聚类结果的影响。
* **改进算法**：例如K-medoids算法，用样本点作为中心点，减少异常值对聚类中心的影响。

 **提高收敛质量**：

* **多次运行K-均值**：通过多次运行算法并选择最优的聚类结果来减少局部最优的影响。
* **混合模型**：结合其他算法，如结合高斯混合模型（GMM），可以提高结果的稳定性和准确性。

2、设有10个二维模式样本，如图所示。若，试用最大最小距离算法对该样本集进行聚类分析。按该聚类算法的步骤，详细给出各个中间结果。

1

3

5

7

9

1

3

5

7

***X***1

***X***4

***X***3

***X***5

***X***8

***X***9

***X***7

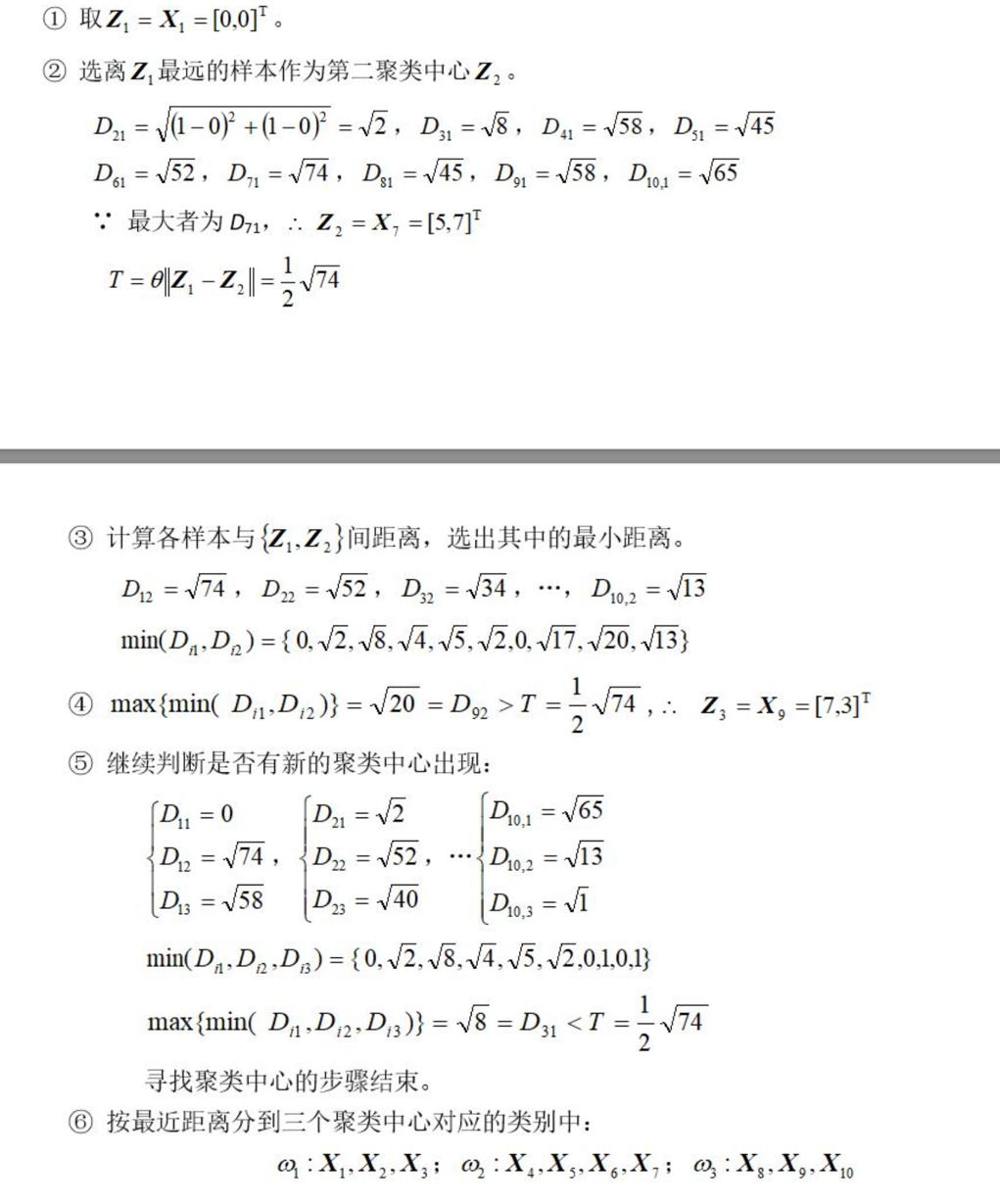
***X***10

***X***2

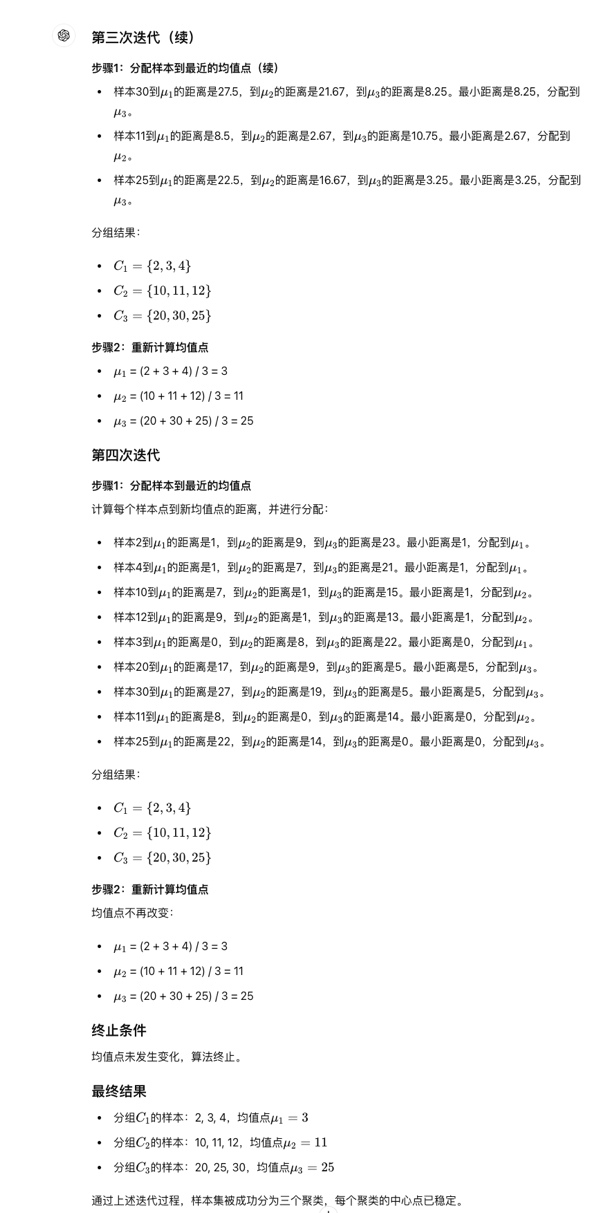
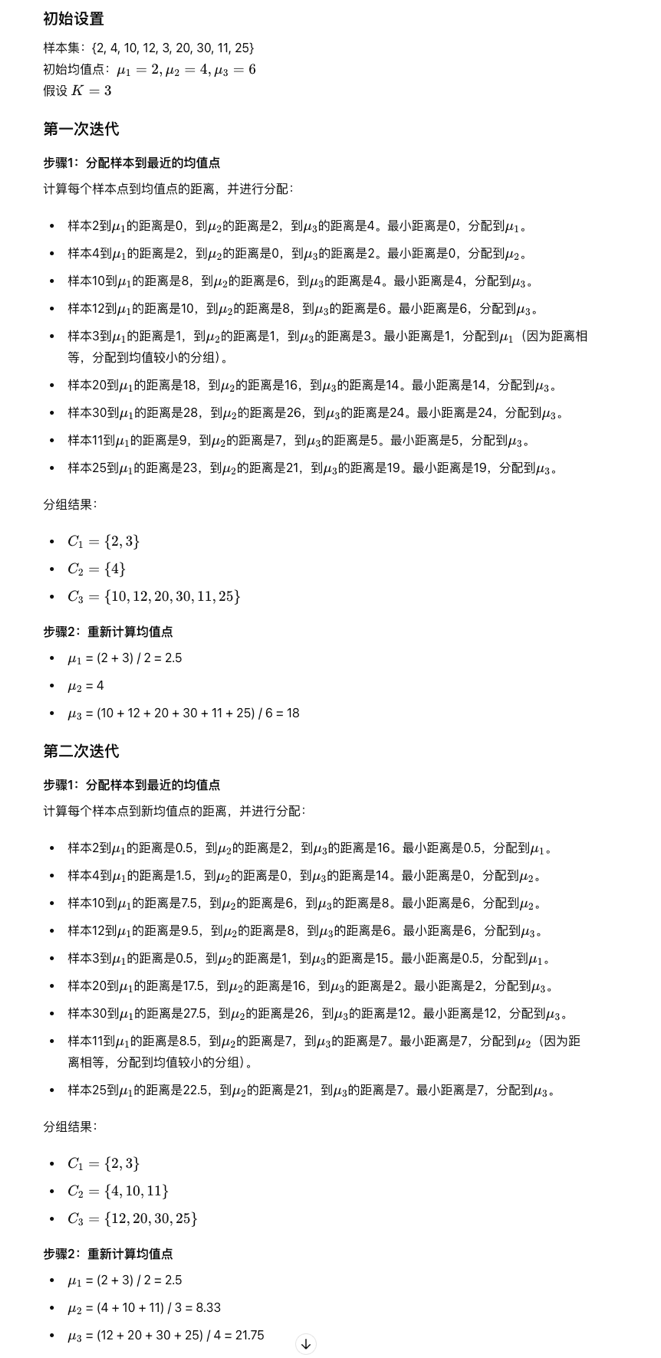
***X***6

*x*1

*x*2



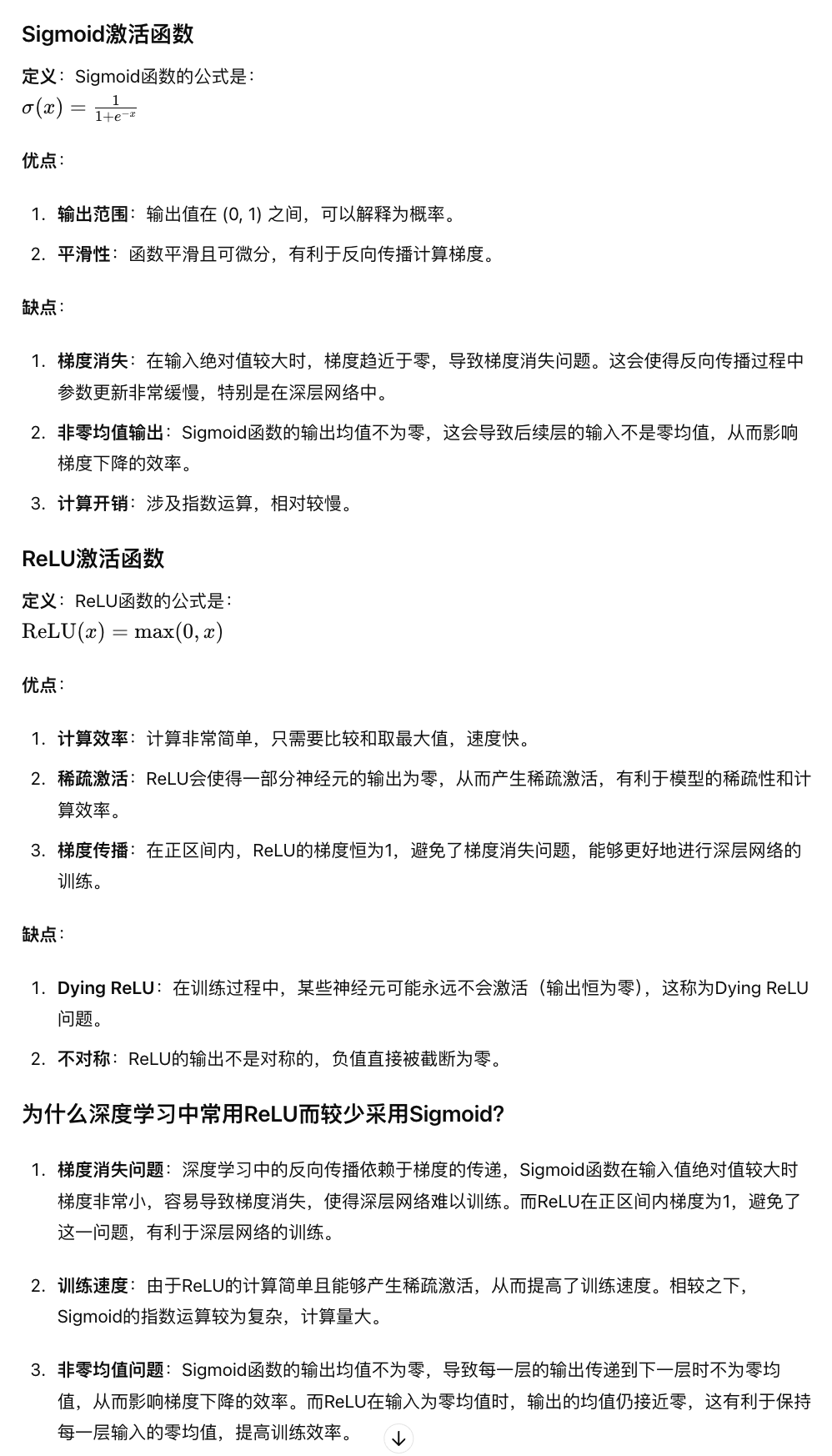
3、给定以下一维样本集：2、4、10、12、3、20、30、11、25，假设K=3，且随机选择初始均值点为。写出使用K-均值算法对样本集进行聚类分析的完整过程（要求写出每次迭代后各分组的样本号以及各分组的均值点）。说明：如果出现与两个均值点距离相等的情况，则样本点被指派给均值较小的分组。



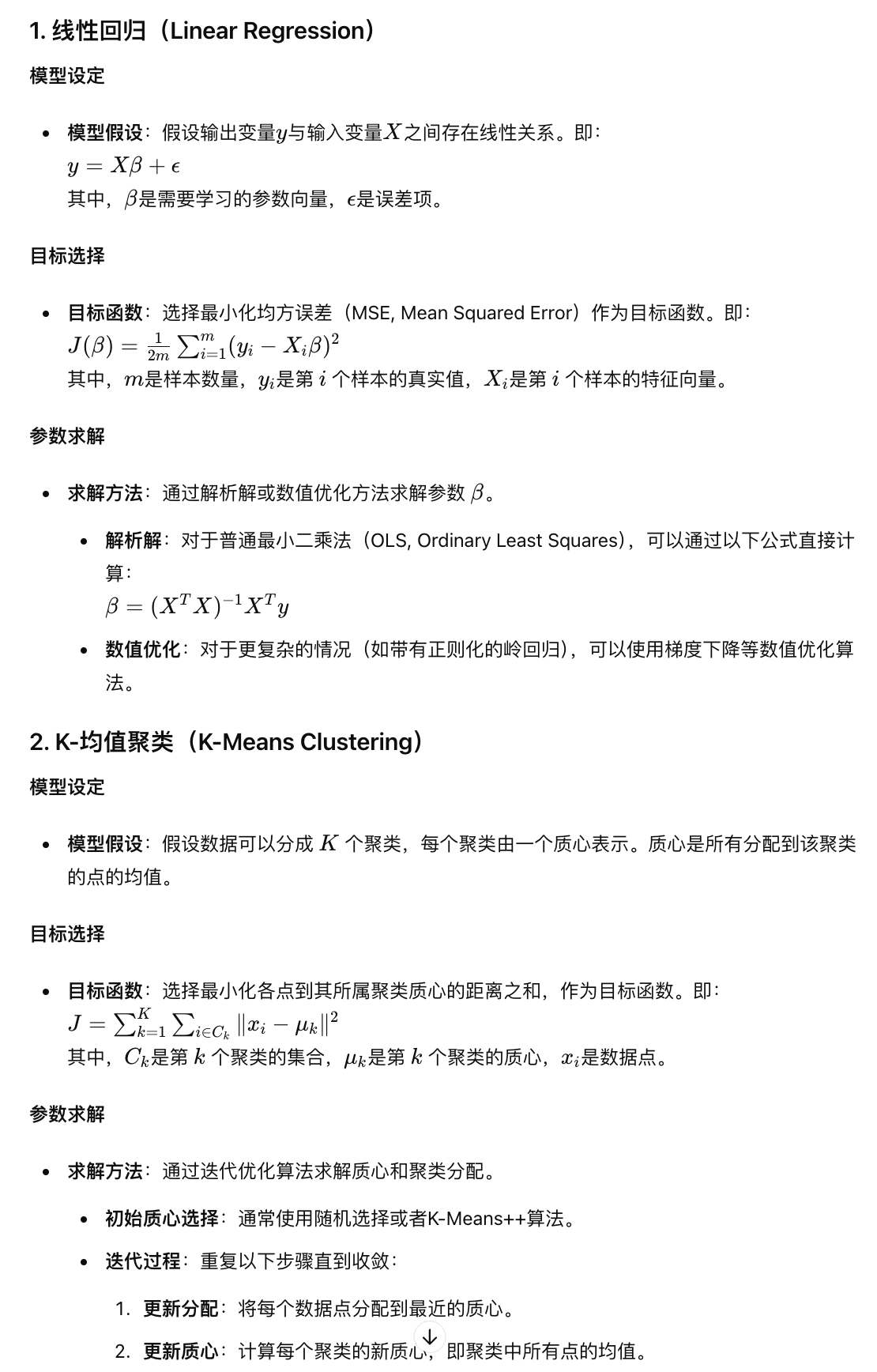
**第9讲 神经网络与深度学习**

1、人工神经元模型从哪些方面对生物神经元模型进行了模拟？请写出人工神经元模型的数学表达形式，并解释各个变量的含义。

2、Sigmoid激活函数与ReLu激活函数相比各有什么优缺点？为什么深度学习中常用ReLu作为激励函数而较少采用Sigmoid激活函数？



3、机器学习方法的三个主要步骤是：模型设定，目标选择和参数求解。试用这一分析框架，对学过的至少两种机器学习方法进行分析。



4、已知两类训练样本为

：

：

设，用感知器算法求解判别函数，并绘出判别界面。