1. 结合课程学习，给出关于人工智能、机器学习的定义？

人工智能：

1. 中国《人工智能标准化白皮书（2018）》关于人工智能的定义：

人工智能是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

1. 谭铁牛院士在2019《求是》定义“人工智能”：

人工智能是研究开发能够模拟、延伸和扩展人工智能的理论、方法、技术及应用系统的一门技术科学。研究目的是促使智能机器：

会听（语音识别、机器翻译等）、

会看（图像识别、文字识别等）、

会说（语音合成、人机对话等）、

会思考（人机对弈、定理证明等）、

会学习（机器学习、知识表示等）

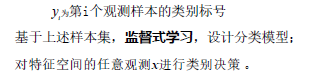
会行动（机器人、自动驾驶汽车等）

机器学习：

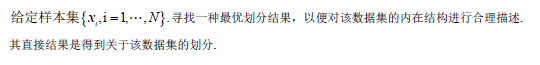
1. 机器学习是人工智能的一个分支，是一门科学学科，涉及算法的开发与设计，该算法的开发与设计，该算法以经验数据为输入，并产生（被认为生成数据的潜在机制特征的）模式或预测。
2. 机器学习（Machine Learning）是一门涉及统计学、系统辨识、逼近理论、神经网络、优化理论、计算机科学、脑科学等诸多领域的交叉学科，研究计算机怎样模拟户或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，是人工智能技术的核心。
3. 什么是分类？什么是聚类？请给出二者的区别与联系。什么是回归？

分类：





聚类：



区别：

分类是基于已知答案的数据集，监督式学习一种划分模型，以便对特征空间进行划分

聚类是得到关于数据集的直接划分结果，是无监督的学习的结果

联系：

可以借助聚类实现数据集的划分，以便实现自动式的标注；进而，基于这种标注结果，学习分类模型，以实现关于整个特惠总能空间的划分。

回归：

越来越接近期望值的过程，回归于事物的本质

1. 什么是监督式学习、非监督式学习？举例说明

监督学习：

用已知标签的数据指导着机器进行学习

非监督式学习：

只用到输入部分的数据，没有用到其他数据指导着机器学习

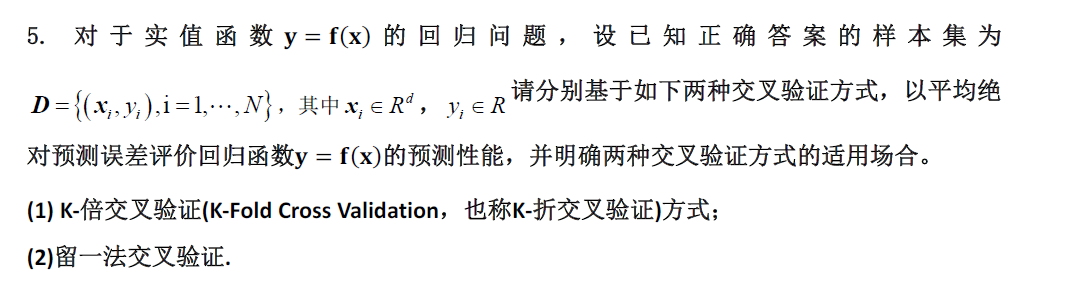
1. 以分类或回归任务为例，结合模型的学习，能够区分：训练集、测试集、验证集、估计集的作用？

训练集：用学习，训练得到最终的分类模型所用的数据集

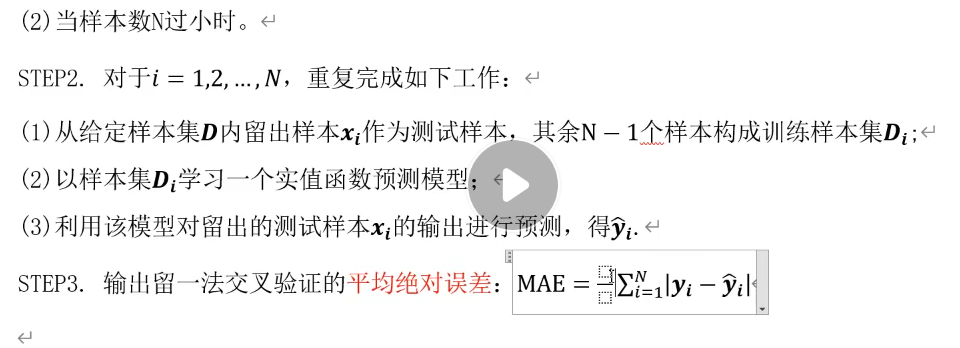
测试集：用来评价最终模型的数据集

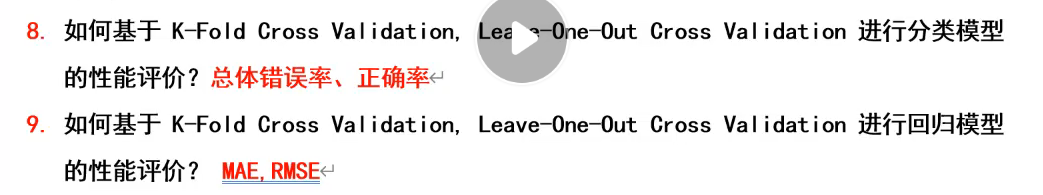
验证集：评价所选择的模型

估计集：用于来选择合适的模型

5、

留一法交叉验证：





MAE 平均绝对误差 RMSE 均方根误差

1. KNN是什么的简称？此处的K值是什么意思？

K近邻算法 （K-Nearest Neighbor algorithm）

在训练数据集中找到与该实例最邻近的K个实例（也就是上面所说的K个邻居），这K个实例的多数属于某个类，就把该输入实例分类到这个类中。

1. 掌握基于KNN近邻进行分类或回归的实现步骤

网上：

1. 计算测试样本和训练样本中每个样本点的距离（欧氏距离、马氏距离等）
2. 对上面所有的距离进行排序
3. 选前K个最小距离的样本
4. 根据这K个样本标签进行投票，得到最后的分类类别
5. KNN近邻决策时，那些因素会影响决策结果
   1. 样本的规范化预处理
   2. 典型的距离度量方式
   3. 超参数K值
6. 如何面向分类或回归任务，采用m-折交叉验证的方式进行K值优选？

你是如何评价每个备选K值的

1）K值较小，则模型复杂度较高，容易发生过拟合，学习的估计误差会增大，预测结果对近邻的实例点非常敏感。

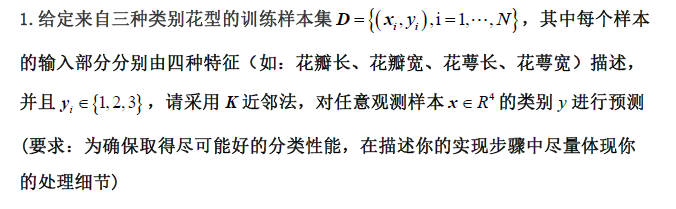
2）K值较大可以减少学习的估计误差，但是学习的近似误差会增大，与输入实例较远的训练实例也会对预测起作用，使预测发生错误，k值增大模型的复杂度会下降。

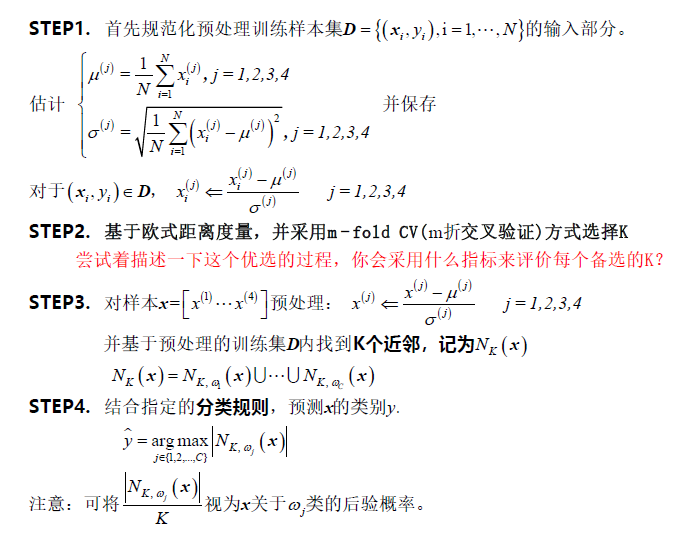
3）在应用中，k值一般取一个比较小的值，通常采用交叉验证法来来选取最优的K值。

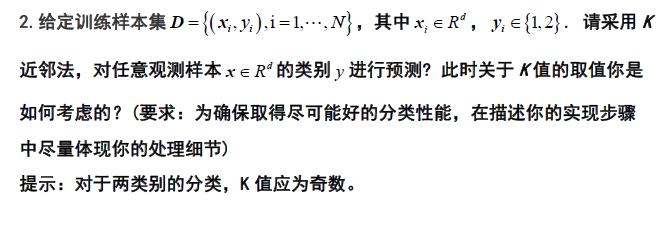
5、若采用KNN法进行两类别的分类，K值的设定会有那些考虑？

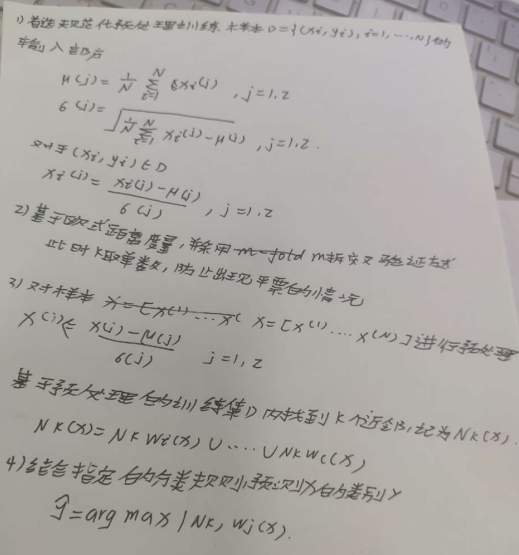
在二元（两类）分类问题中，选取*k*为奇数有助于避免两个分类平票的情形

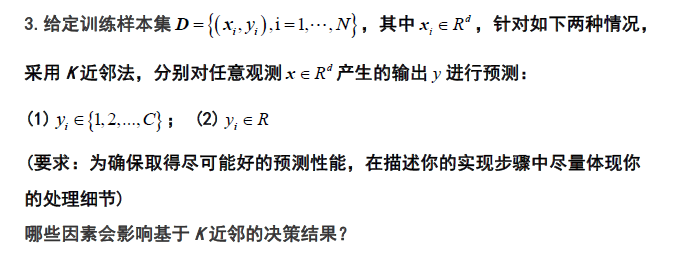
练习题：











那些因素会影响基于K近邻的结果：

数据的预处理

K值的选择

决策方式的选择

1. 什么是决策树？决策树可以完成那些可能的机器学习任务？

基于树形结构的决策模型——决策树

决策树是一种机器学习的方法。决策树的生成算法有ID3,C4.5和C5.0等。

决策树是一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。

大数据集上推荐神经网络

在项目时间较短的项目上，如果数据质量低(大量缺失值、噪音等)，优先使用集成树模型

在硬件条件有限及机器学习知识有限的前提下，优先选择树模型

2、什么是单结点数？什么是决策树的树桩？

？？

3、决策树与特征空间、训练样本集是什么关系？

决策树是一种基本的分类与回归方法，是定义在特征空间与类空间上的条件概率分布。

1. 决策树模型？ID3,C4.5,CART
2. 那些方式可以度量决策树中某个结点的不纯度？（三种方式）

熵不纯度、Gini不纯度、误差不纯度等

1. ID3（分类）、C4.5（分类）、CART(分类或回归)三种决策树模型在构建过程中，结点是如何进行特征选择的。
2. ID3决策树内每个非叶结点的特征选择，采用最大“绝对信息增益”准则，选特征。
3. C4.5决策树基于候选特征，估计“增益率”平均值，确定增益率高出平均水平，并具有最大增益率的特征。
4. Cart决策树（用于分类时）基于最小“划分后基尼指数”原则，进行结点特征选择。
5. 分别面向分类/回归问题，掌握CART树的实现步骤，分类树的叶子结点预测值是什么？

回归树的叶子结点预测值是什么？它们与特征空间是什么对应关系？

分类——多棵决策树投票决策

回归——多棵决策树预测结果的平均

自己：

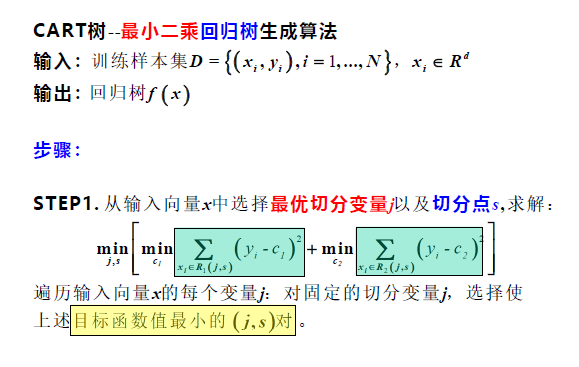
分类树的叶子结点预测值是预测的是该数据应该是那个类别

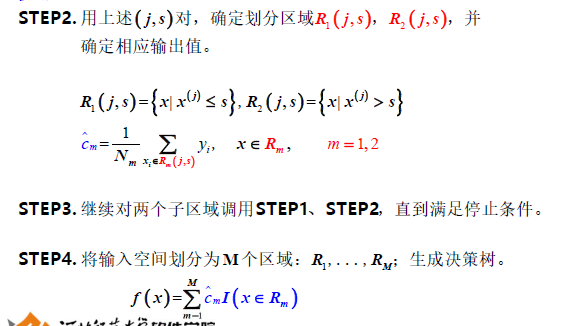
回归树的叶子结点预测值是在其临近的数据单元的一个划分

一个分类树对应这一个具体的类别。

一个回归树对应输入空间（或特征空间）的一个划分，以及在该划分单元上的输出值。

1. 给定已知标签的训练样本集，简述基于该样本集构造CART回归树（或CART分类树）树桩的实现步骤，并指出该树桩的叶子结点输出值是如何估计出来的。





1. 随机森林、Bagging两种模型的学习步骤，以及使用方式。

Bagging

1)、从原始样本集中抽取训练集

2）、每次使用一个训练集得到一个模型，k个训练集共得到K个模型。将上部得到K个模型采用投票的方式得到分类结果，对回归问题，计算上述模型的均值作为最后的结果。

随机森林

1. 设有N个样本，有放回地随机选择n个样本
2. 设每个样本有M个属性，在每个决策树进行分裂时，随机抽样m个属性，m远小于M，评价每个属性，选择最佳属性作为分裂属性
3. 对于每个节点而言都要按照第二步来走，一直到评价指标遍历所有抽样出来的属性都无法得到改进为止。每棵树都尽可能生长没有剪枝。
4. 重复前面三个步骤就变成了随机森林。
5. 面向两类别分类，理解AdaBoost集成模型的基本思想、算法的实现步骤

1）基学习器之间存在强依赖关系，一系列基学习器需要串行生成，代表算法是Boosting

2）Boosting指的是一类集成方法，其主要思想就是将弱的基学习器提升（boost）为强化学习器：

A、先用每个样本权重相等的训练集训练一个初始学习器

B、根据上轮得到的学习器对训练集的预测表现情况调整训练集中的样本权重(例如提高被错分类的样本的权重使之在下轮训练中得到更多的关注), 然后据此训练一个新的基学习器

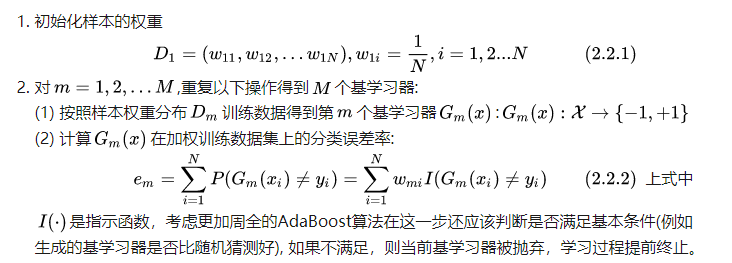
C、重复2直到得到M个基学习器，最终的集成结果是M个基学习器的组合。

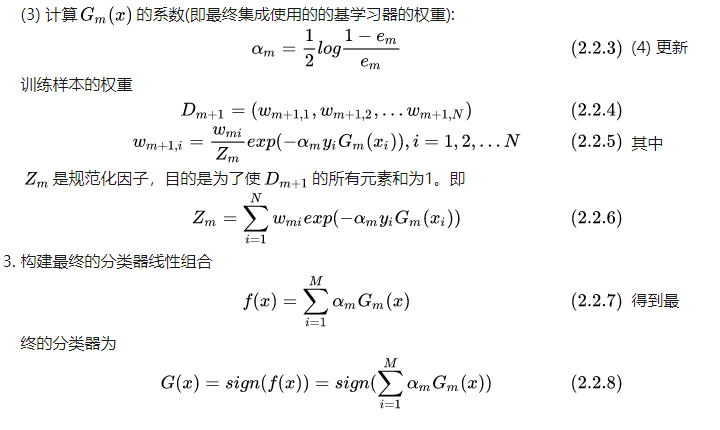
AdaBoost(Adaptive Boosting, 自适应增强)算法采取的方法是:

提高上一轮被错误分类的样本的权值，降低被正确分类的样本的权值；

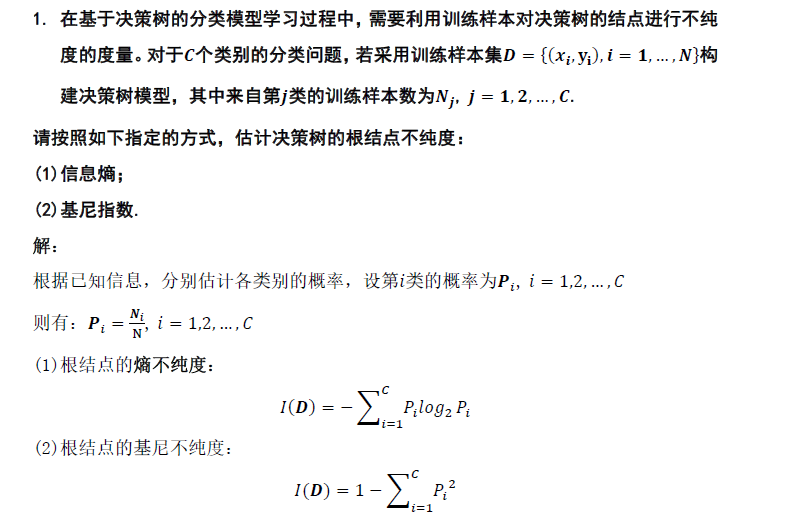
线性加权求和。误差率小的基学习器拥有较大的权值，误差率大的基学习器拥有较小的权值。

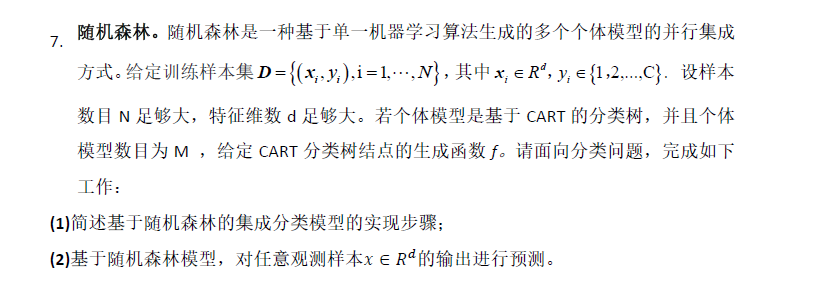
算法步骤：

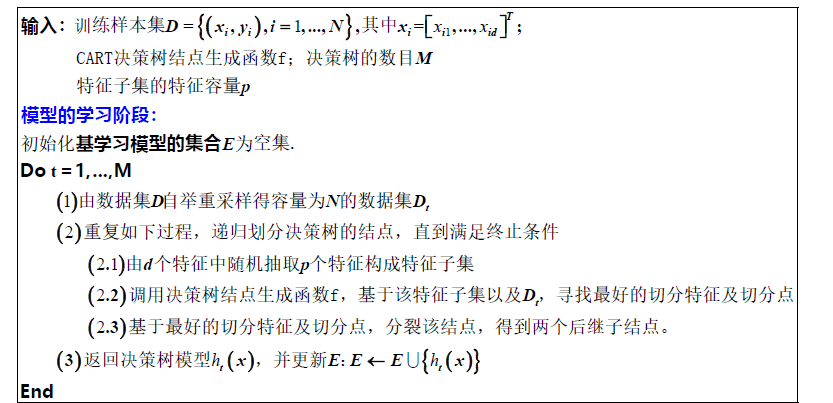


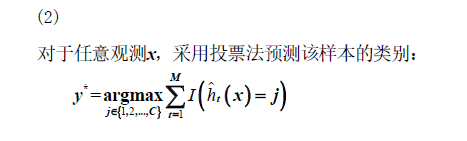


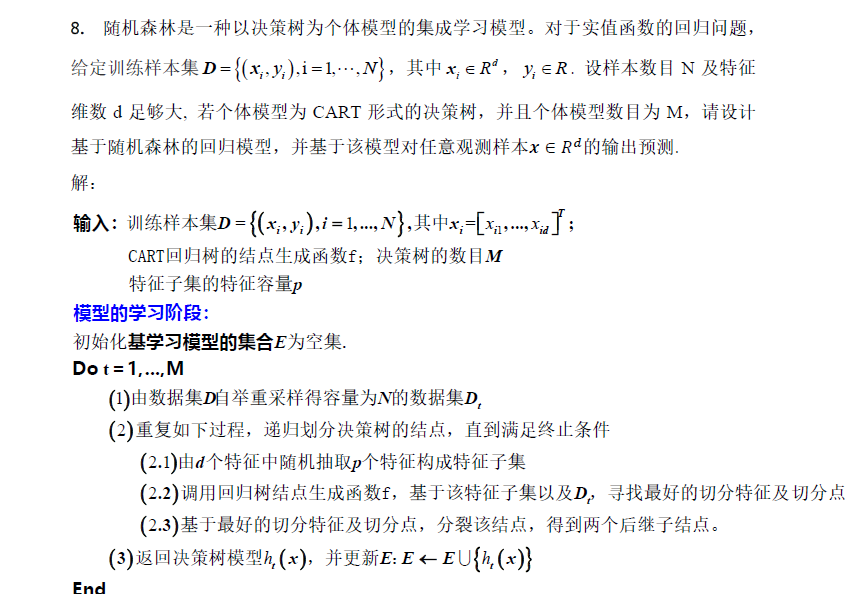
练习：











1. 掌握连续特征空间、离散特征空间两种情况下，基于最小错误率的贝叶斯分类决策规则。

