第一题

假设我们目前有包含m个样本的数据集D，为了满足训练和测试的需求，我们对D进行适当的处理，从中产生出训练集S和测试集T。

简述从数据集D中产生训练集S和测试集T的方法及优缺点（最少两种方法）

1.留出法

直接将数据集D划分为两个互斥的集合，其中一个集合作为训练集S，另一个作为测试T。

优点：步骤相对简单

缺点：训练集S与测试集T的划分比例不好确定，评估结果不够稳定准确

2.交叉验证法

先将数据集D划分为k个大小相似的互斥子集。然后，每次用k-1个子集的并集作为训练集，余下的那个子集作为测试集

特例：数据集D中包含m个样本k=m时为留一法

优点：不受随机样本划分方式的影响。评估结果较为准确。

缺点：数据集较大时 训练开销大。

3.自助法

对D进行采样产生数据集D'：每次随机从D中挑选一个样本，并将其拷贝放入D'中，然后再将该样本放回数据集D中，使得该样本在下次采样时仍有可能被采到；

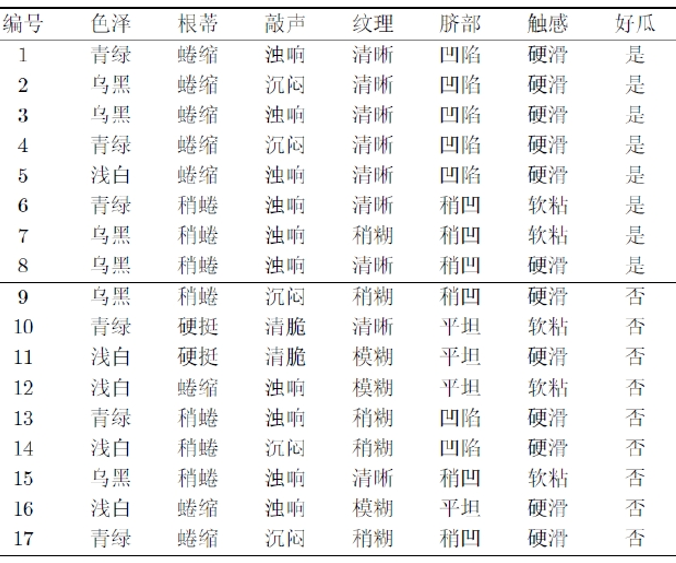
这个过程重复执行m次后，得到了包含m个样本的数据集D'。将D'作为训练集，将D-D‘（集合减法）作为测试集。

优点：数据集小、难以划分训练/测试集时很有用

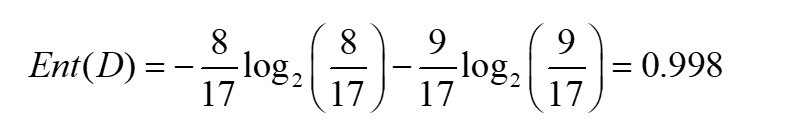
缺点：改变初始数据集分布，引入估计偏差。

第二题

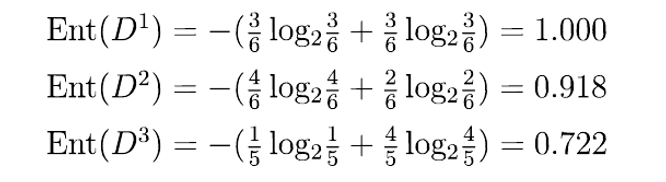
现有西瓜数据如下表所示，计算构建决策树第一次分支时“色泽”属性的信息增益，并根据最佳属性划分“纹理”画出第一次分支后的分类情况。



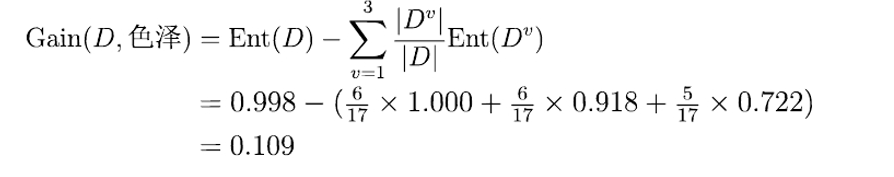
解：信息熵



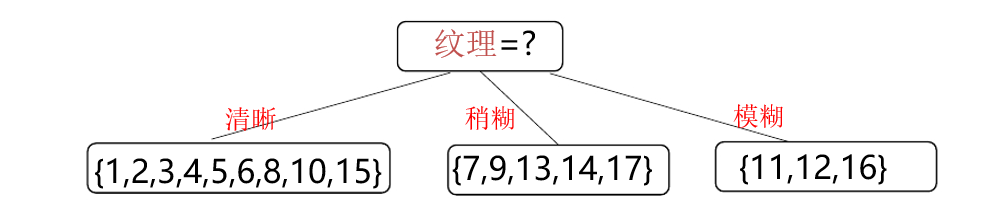
属性“色泽”的3个数据子集是D1(色泽=青绿)，D2(色泽=乌黑)，D3(色泽=浅白)



色泽的信息增益为



分类后



第三题

现有一决策树(设每个叶子节点只有一个数据)如下所示，根据验证集画出经过后剪枝操作后优化的决策树。

3根蒂

1脐部

4色泽

2色泽

平坦

稍凹

凹陷

硬挺

蜷缩

稍蜷

青绿

浅白

乌黑

浅白

青绿

乌黑

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 色泽 | 根蒂 | 脐部 | 好瓜 |
| 1 | 青绿 | 蜷缩 | 凹陷 | 是 |
| 2 | 浅白 | 蜷缩 | 凹陷 | 是 |
| 3 | 乌黑 | 稍蜷 | 稍凹 | 是 |
| 4 | 浅白 | 稍蜷 | 稍凹 | 是 |
| 5 | 浅白 | 硬挺 | 平坦 | 否 |
| 6 | 浅白 | 蜷缩 | 平坦 | 否 |
| 7 | 青绿 | 稍蜷 | 凹陷 | 否 |
| 8 | 青绿 | 蜷缩 | 稍凹 | 否 |

解：第一步，考虑结点4，剪枝前精度有5/8，剪枝后精度6/8，将结点4剪枝。

第二步，考虑结点3，剪枝前精度6/8，剪枝后精度5/8，不剪枝。

第三步，考虑结点2，剪枝前精度6/8，剪枝后精度7/8，将结点2剪枝。

第四步，考虑结点1，剪枝前精度7/8，剪枝后精度1/2，不剪枝。

剪枝后的决策树如下：

硬挺

蜷缩

稍蜷

凹陷

稍凹

平坦

1脐部

3根蒂

第四题

试述将线性函数用作神经元激活函数的缺陷，试写出常用的激活函数（至少三种）并说明其特点。

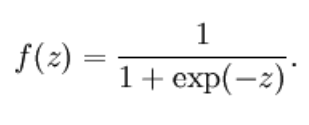
答：

使用线性函数作为激活函数时，无论是在隐藏层还是在输出层（无论传递几层），其单元值都还是输入x的线性组合，这个时候的神经网络等价于逻辑回归，若输出层也使用线性函数作为激活函数，那么就等价于线性回归 。

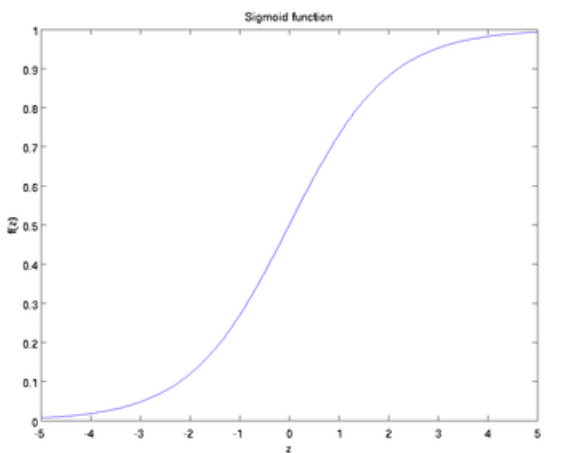
常用的激活函数：

1）Sigmoid函数

公式：



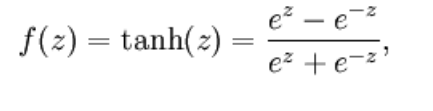
图像：



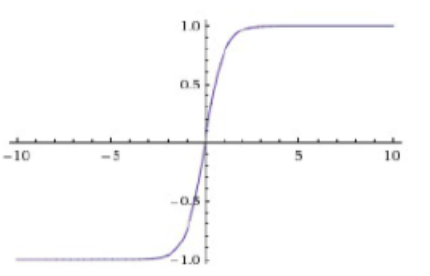
sigmoid激活函数计算量大，求导涉及除法，反向传播时，很容易就会出现梯度消失的情况，从而无法完成深层网络的训练

2）tanh函数

公式：



图像：



与 sigmoid 的区别是，tanh 是 0 均值的，因此实际应用中 tanh 会比 sigmoid 更好。

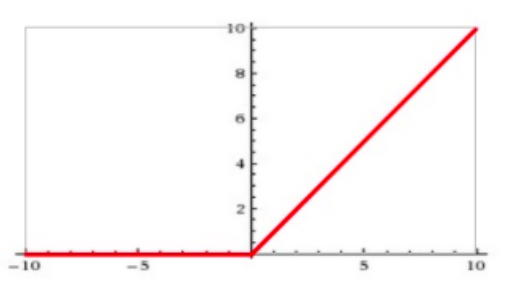
缺点是也容易饱和从而使得梯度消失。

3）ReLU函数

公式：



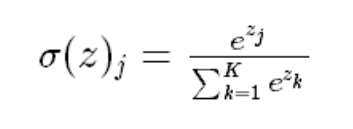
图像：



对随机梯度下降的收敛有巨大的加速作用，而且非常容易计算。

4） Softmax函数

公式：

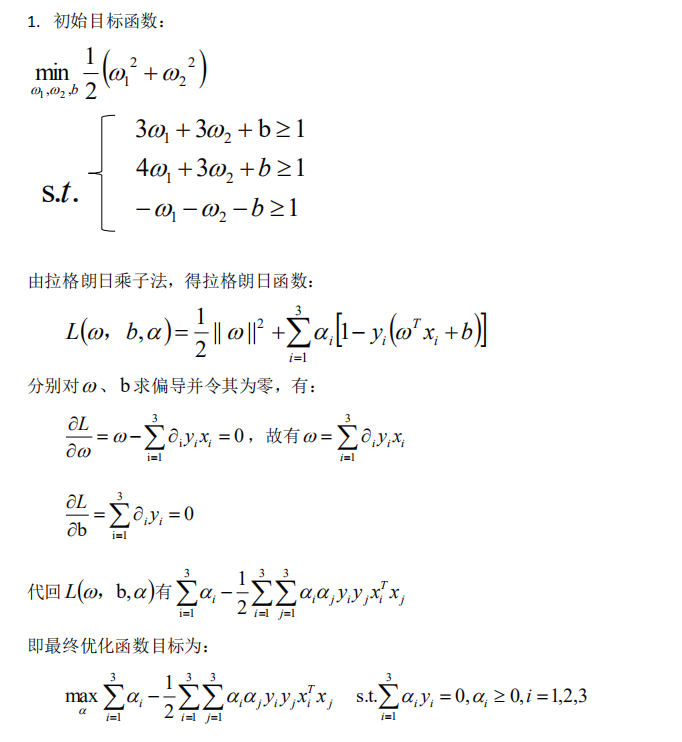


主要应用是多分类。

第五题

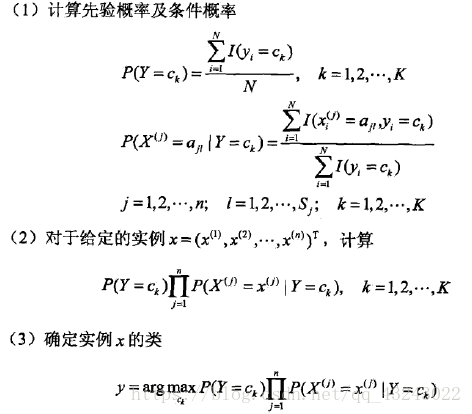
现有正实例点,负实例点。写出优化目标函数与约束条件。根据拉格朗日乘子法及其对偶问题的转化，写出最终优化函数与其对应的约束条件。

**解答：**



第六题

1. 简述朴素贝叶斯分类步骤



第七题

根据个体学习器的生成方式，目前集成学习方法大致可以分为两大类：Boosting和Bagging。Boosting中的典型代表是AdaBoost算法。试从采样方法、适用情况、个体学习器间的依赖关系以及偏差、方差的角度阐述Adaboost与Bagging各自的特点。

**答案:**

1.采样方法：

AdaBoost采用序列采样的方法，先从初始训练集训练出一个基学习器，再根据基学习器的表现以重赋权的方式对训练样本分布进行调整，使得先前基学习器做错的训练样本在后续受到更多关注，然后基于调整后的样本分布来训练下一个基学习器。

Bagging采用自助采样方法，根据每个采样集训练出一个基学习器，再将这些基学习器进行结合。

2.适用情况：

AdaBoost：适用于个体学习器较弱，正确率略高于50%，没有办法很好的拟合训练样本的情况。

Bagging适用于复杂的、容易过拟合的模型（例如决策树）

3.依赖关系:

AdaBoost的个体学习器存在较强的依赖关系，是串行生成的。

Bagging的个体学习器不存在强依赖关系，是并行生成的。

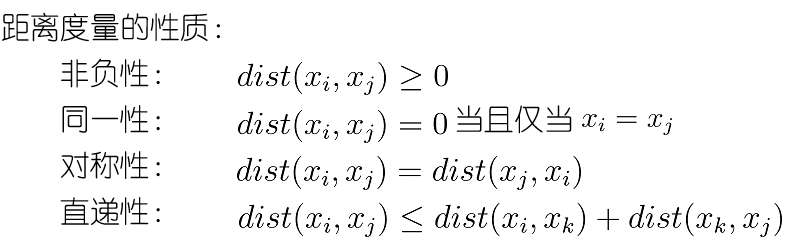
4.偏差方差：

AdaBoost降低偏差，可对泛化性能相当弱的学习器构造出很强的集成。

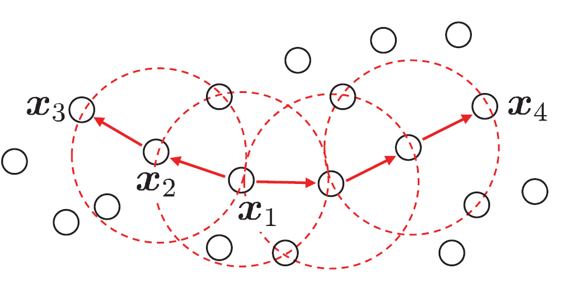
Bagging主要关注降低方差 ，在不剪枝的决策树、神经网络等易受样本影响的学习器上效果更好。

第八题

（1）请写出距离度量的性质



（2）如图



虚线显示出邻域，为核心对象，则

由\_\_\_\_\_\_\_

由\_\_\_\_\_\_\_

由\_\_\_\_\_\_\_

若, 则与\_\_\_\_\_\_\_

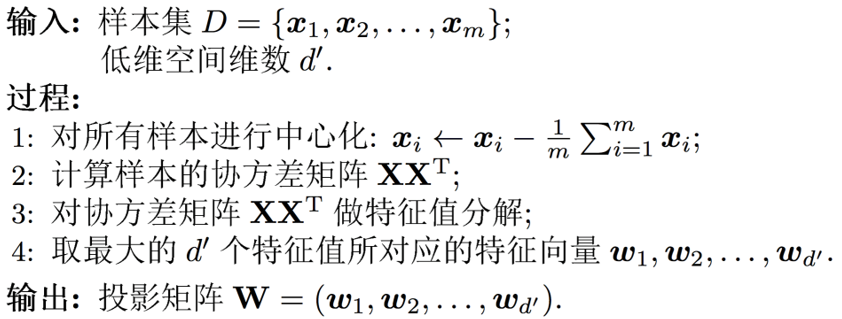
a)密度可达 b) 密度相连 c) 密度直达

答案: cabb

第九题

请写出PCA算法的流程

答案：



第十题

简述字典学习与压缩感知。

字典学习也叫稀疏编码,指的是为普通稠密表达的样本找到合适的字典,将样本转化为合适的稀疏表达形式,从而使学习任务得以简化,模型复杂度得以降低的过程；

压缩感知关注的是利用信号本身的稀疏性,从部分观测样本中恢复原信号，分为感知测量和重构恢复两个阶段。感知测量是对原始信号进行处理以获得稀疏表示，重构恢复是基于稀疏性从少量的观测中恢复原信号。其中重构恢复比较重要，可利用矩阵补全等方法来解决推荐系统之类的协同过滤任务。

第十一题

### 试证明Jensen不等式：对任意凸函数f(x)有

### 证明：显然，对任意凸函数f(x)，必然有



**取****所以：**

**以此类推得：**

第十二题

试解释主动学习、半监督学习、纯半监督学习、直推学习的区别。

答：

主动学习：在某些情况下，没有类标签的数据相当丰富而有类标签的数据相当稀少，并且人工对数据进行标记的成本又相当高昂。在这种情况下，我们可以让学习算法主动地提出要对哪些数据进行标注，之后我们要将这些数据送到专家那里让他们进行标注，再将这些数据加入到训练样本集中对算法进行训练。若每次提出对改善模型性能帮助大的数据，则只需要询问专家少量的数据就能构建出比较强的模型，从而降低标记成本，这一过程叫做主动学习。在利用未标注数据的时候，是从未标注数据中找到最容易判断错误的样例来交由专家进行标注，这个过程是一个筛选差数据的过程，也是一个互动交互的过程，引入了额外的专家的知识。  
半监督学习：指的是学习算法不需要人工干预，基于自身对未标记数据加以利用，来提高学习模型的泛化性能。半监督学习进一步分为纯半监督学习和直推学习。纯半监督学习假定训练数据中的未标记样本并非待预测的样本，希望学得模型能够适用于训练过程中未观察到的数据，即“开放世界”；直推学习假定学习过程中考虑的未标记样本恰是待预测数据，仅试图对学习过程中观察到的未标记数据进行预测，即“封闭世界”。

第十三题

试述**成对马尔科夫性、局部马尔科夫性和全局马尔科夫性。**

**成对马尔科夫性：**设u和v是无向图G中任意两个没有边连接的结点（也就是说两个之间没有依赖的关系），结点u和v分别对应随机变量Yu和Yv。其他所有结点为O，对应的随机变量组是YO。则成对马尔科夫性的表达式如下

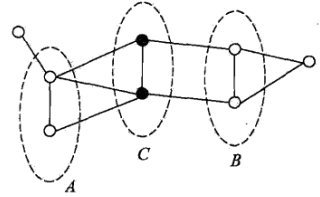
https://images2018.cnblogs.com/blog/1335117/201807/1335117-20180714144912112-355215415.png

即给定了随机变量组YO的条件下随机变量Yu和Yv是条件独立的。

**局部马尔科夫性：**设v是无向图G中的任意结点，W是与v有边连接的所有结点，O是v，W以外的所有结点（相当于W将v和O给隔开了）。则在给定W的条件下，v和O之间是相互独立的，表达式如下

https://images2018.cnblogs.com/blog/1335117/201807/1335117-20180714145301934-1839414950.png

**全局马尔科夫性：**设结点集A，B在无向图G中被结点集合C分开的任意结点的集合，具体如下图所示

（图就不画了吧，懂了就行^\_^）

则在给定了集合C的条件下结点集合A和B之间是相互独立的，具体表达式如下

https://images2018.cnblogs.com/blog/1335117/201807/1335117-20180714145544563-496453188.png

第十四题

马尔可夫决策问题（MDP）是一个四元组的问题，请解释各要素。

答：

E=<X,A,P,R>；

X为状态空间；状态是对环境的描述，在智能体做出动作后，状态会发生变化，且演变具有马尔可夫性质。MDP所有状态的集合是状态空间。状态空间可以是离散或连续的。

A为动作空间；动作是对智能体行为的描述，是智能体决策的结果。MDP所有可能动作的集合是动作空间。动作空间可以是离散或连续的。

P为转移函数，若某个动作a∈A作用在当前状态x上，则潜在的转移函数P将使得环境从当前状态按某种概率转移到另一个状态，其中R指定了状态转移概率；

R为奖赏函数：智能体给出动作后环境对智能体的反馈。是当前时刻状态、动作和下个时刻状态的标量函数。

第十五题

能否结合GMM谈谈EM算法的优化目标函数、符号意义和M-Step、E-Step的内容。最后，能否说说它和逻辑回归比较起来，优势和弱势在什么地方？

答案自己准备。