

## 一、选择题

1.逻辑回归模型解决（ B ）

- A.回归问题      B.分类问题      C.聚类问题      D.推理问题

2.逻辑回归不能实现（ D ）

- A.二分类      B.多分类      C.分类预测      D.非线性回归

3.关于机器学习中的决策树学习，说法错误的是（ B A ）

- A.受生物进化启发      B.属于归纳推理  
C.用于分类和预测      D.自顶向下递推

4.在构建决策树时，需要计算每个用来分裂的属性得分值，选择分值最高（或最低）的特征，以下不能作为得分的是（ D ）

- A.信息增益      B.基尼系数      C.信息增益率      D.信息熵

5.在决策树学习过程中，（ D ）可能会导致问题数据（特征相同但是标签不同）

- A.数据噪音      B.现有特征不足以区分或决策  
C.数据错误      D.以上都是

6.根据信息增益来构造决策树的算法是（ A ）

- A.ID3 决策树      B.递归      C.归约      D.FIFO

7.决策树构成顺序是（ A ）

- A.特征选择、决策树生成、决策树剪枝  
B.决策树剪枝、特征选择、决策树生成  
C.决策树生成、决策树剪枝、特征选择  
D.特征选择、决策树剪枝、决策树生成

8.支持向量指的是（ C B ）

- A.对原始数据进行采样得到的样本点 B.决定分类面可以平移的范围的数据点  
C.位于分类面上的点 D.能够被正确分类的数据点

9.下面关于支持向量机(SVM)的描述错误的是 ( D )

- A.是一种监督式学习的方法 B.可用于多分类的问题  
C.支持非线性的核函数 D.是一种生成式模型

10.下面关于支持向量机(SVM)的描述错误的是 ( D )

- A.对于分类问题,支持向量机需要找到与边缘点距离最大的分界线,从而确定支持向量  
B.支持向量机的核函数负责输入变量与分类变量之间的映射  
C.支持向量机可根据主题对新闻进行分类  
D.支持向量机不能处理分界线为曲线的多分类问题

12.支持向量机中 margin 指 ( C )

- A.盈利率 B.损失误差 C.间隔 D.保证金

13.选择 margin 最大的分类器的原因是 ( D )

- A.所需的支持向量个数最少 B.计算复杂度最低  
C.训练误差最低 D.有望获得较低的测试误差

## 二、简答题

1.ID3 和 CART 算法有什么区别?

1. ID3 算法使用信息增益作为划分属性的准则。每属性  $a$  在  $D$  中的信息增益为  $Gain(D, a) = Ent(D) - \sum_{v \in V} \frac{|D_v|}{|D|} Ent(D_v)$ 。信息增益越大,用该属性划分所获得的纯度提升越大,越愿意选择该属性作为划分。但 ID3 只能处理离散属性,并且更倾向选择属性值较多的属性。  
CART 算法使用 Gini 不纯度作为划分标准,构造二叉树划分属性,还可用于回归。  
综上所述,有以下几点区别:  
① 划分准则不同: ID3 使用信息增益, CART 使用 Gini 系数  
② 决策树形状不同: ID3 构造的是普通树, CART 构造二叉树  
③ 计算量不同: CART 算法计算 Gini 时计算量较小  
④ 应用场景不同: ID3 只能处理离散值分类问题, CART 可以处理分类与回归

2. 过拟合和欠拟合会导致什么后果，应该怎样避免？

2. 过拟合导致模型泛化能力差，欠拟合导致模型准确性低  
过拟合可以采用提前停止训练，添加正则项等方式解决  
欠拟合可以采用增加训练次数，扩大训练集，多项式回归等方式解决

3. 什么是正则化，正则化有什么意义？

3. 正则化就是在误差目标函数中增加一项来防止模型过拟合，如：  
 $L_1$ 正则： $L(\theta) = \lambda L_0(\theta) + (1-\lambda) \sum ||w||$   
 $L_2$ 正则： $L(\theta) = \lambda L_0(\theta) + (1-\lambda) \sum ||w||^2$   
正则化实际是对损失函数的惩罚，对训练中一些参数加以限制，防止训练过拟合。如  $L_1$  正则一般用于特征选择， $L_2$  正则用于防止过拟合（ $L_1$  也可）