9．常用的聚类距离计算采用（）。

A.余弦距离

B.闵可夫斯基

C.马氏距离

D.流形距离

10．缓解维数灾难的操作是（）

A.降维

B.升维

C.低维嵌入

D.流型处理

二、判断题（每小题1分，共10分）

1．机器学习的目标就是获得与训练集一致的假设。 ( )

2．过拟合不可以彻底避免。 ( )

3．决策树学习算法中，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于不同类别。 ( )

4．基尼值可以度量样本集合的纯度。 ( )

5．只拥有一层功能神经元（能进行激活函数处理）的感知机学习能力依然很强，能解决异或这样的非线性可分问题。 ( )

6．引入软间隔是为了允许支持向量机在一些样本上出错。 ( )

7．核函数的引入是将原始空间中不可分的训练样本映射到高维的特征空间后变得可分。 ( )

8．极大似然估计是试图在θ。所有可能的取值中，找到一个能使数据出现的“可能性”的最大的值。 ( )

9．聚类性能度量大致有两类指标：外部指标和内部指标。 ( )

10．在高维情形下出现的数据样本稀疏、距离计算困难等问题是所有机器学期方法共同面临的严重障碍，被称为维数灾难。 ( )

此，线性判别分析也常被视为一种经典的（）。

A.监督降维技术

B.分类技术

C.监督分类技术

D.降维技术

5．在属性划分中，信息增益越大，结点的纯度（）。

A.不变

B.变为零

C.提升越大

D.降低越快

6．BP算法基于（）策略对参数进行调整

A.梯度下降

B.梯度上升

C.最小化误差

D.误差逆传播

7．当训练样本线性不可分时可采用（）来缓解和解决

A.增加正例样本

B.核函数

C.训练误差最小

D.测试误差最小

8．对分类任务来说，在所有相关概率都已知的理想情况下，（）考虑如何基于这些概率和误判损失来选择最优的类别标记。

A.支持向量机

B.贝叶斯决策论

C.决策树

D.聚类

课程名称：机器学习

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 题号 | - | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 | 十 | 总分 |
| 得分 |  |  |  |  |  |  |  |  |  |  |  |
| 签字 |  |  |  |  |  |  |  |  |  |  |  |

注意事项：所有的答案都必须写在答题纸（答题卡）上，答在试卷上一律无效。

-、选择题（每小题2分，共20分）

．对西瓜的成熟度进行预测得到结果为0.51，这属于（）学习任务。

A.分类

B.回归

C.聚类

D.其余选项都不是

2．在学习过程中，X表示数据集，Y是所有标记的集合，也称为（）。

A.输出空间

B.函数

C.属性集合

D.样本集合

3．下列关于查全率和查准率的说法哪种正确（）。

A.查全率和查准率成反比

B.查全率和查准率成正比

C.查全率和查准率存在着互逆关系

D．好的模型可以做到查全率和查准率都达到100％

4．线性判别分析可通过该投影减小样本点的维数，且投影过程中使用了类别信息。因

三、填空（每空2分，共10分）

1．模型的评估包括交叉验证法、 、自助法。

2．线性回归模型中，联系函数g（）为对数几率函数In（y／（l-y））时，该线性模型成为＿。

3．决策树的核心技术包含划分选择和 。

4．集成学习中的集成主要分为同质集成和 两种。

5．神经网络的输出，是通过 的处理得到的。

四、简答题（每小题6分，共24分）

1．线性判别分析LDA是如何对新样本进行分类的？

2．写出支持向量机的KKT条件并阐述支持向量机解的稀疏性。

3．简述欠拟合和过拟合的含义。

4．简述最大重构性和最大可分性的含义。

五、计算题（3小题，共36分）

1．关于线性回归算法，请回答以下问题： (12)

（1）给出单一属性的线性回归目标函数．

（2）若用最小二乘法求解线性回归模型，给出最小二乘法的闭式解及其推导过程。

2．设在某个班级男女生识别中男生ω2和女生ω1两类的先验概率分别为：

男生： 女生：

现有一待识别的学生，其观察值为x，从类条件概率密度分布曲线上查得

女生：

请回答以下问题：

(12)

（1）给出贝叶斯判定准则的具体式子。

（2）使用贝叶斯决策对该学生x进行男女性别分类（给出具体计算过程及计算结果）

3．选择决策树、支持向量机、随机森林以及贝叶斯网络四种算法中的一种算法，给出算法原理，并举例说明该算法在日常生活中的应用。(12)