统计与数据科学学院本科生 2021——2022 学年第 2 学期数据挖掘和机器学习课程期末考试试卷（A 卷）

草 稿 区

专业：_____年级：_____学号：_____姓名：_____成绩:_____

得 分

一 、选择题（本题共 20 分，每小题 2 分）

1. 下列哪一项不是大数据的特点：（ ）

A. 数据量大 B. 数据多样性强 C. 数据产生速度快 D. 数据价值高

2. 下列哪种策略有可能降低模型泛化误差的上界：（ ）

A. 提高模型在训练集上的精度 B. 减小训练集样本容量

C. 增大模型规模至过拟合 D. 在测试集上选择合适的超参数

3. 根据偏差-方差分解理论，在模型复杂度较高的情况下，下列描述符合事实的是：（ ）

A. 模型可能发生欠拟合 B. 训练数据的轻微扰动可能会对结果产生明显影响

C. 模型的训练偏差在总误差中占主导地位 D. 增加模型参数或可提升效果

4. 下列关于分类器训练的说法正确的是（ ）

A. 模型的超参数可以通过模型在测试集上的准确率来确定，可选择测试集上效果最优的一组超参数作为模型超参数

B. K 近邻算法是一种典型的无监督学习方法

C. 训练过程中，我们通常以损失函数在测试集上的误差为优化目标，希望该误差尽可能小

D. 用某高校所有男生的网购数据训练模型来预测该校女生的网购习惯，这种做法的不合理性主要来自于无法满足独立同分布假设

5. 下列关于模型的评估说法错误的是（ ）

A. 泛化误差是指模型对未知测试数据的预测误差

B. 若希望尽可能避免正样本被错分为负样本，则应当努力提高模型的精度 Precision

C. 可以根据 ROC 曲线下的面积评估模型的性能

D. PR 曲线是一条单调递减的曲线，而 ROC 曲线是单调递增的曲线

6. 下列关于主成分分析法的描述正确的是：（ ）

A. 是一种特征选择方法

B. 是一种监督降维方法

C. 希望投影后的数据类内方差更小

D. 经过线性组合（正交）变换，得到一组按“重要性”（方差）从大到小排列的特征

7. 下列哪种算子是非凸的：（ ）

A. L0 B. L1 C. L2 D. 都不是

8. 下列线性判别准则中，具有闭式解的是：（ ）

A. 感知准则 B. 最小错分样本数准则

C. 最小平方误差准则 D. 线性支持向量机

9. 下列特征常用于直线提取的是：（ ）

A. 直方图 B. Hough 变化 C. 灰度共生矩 D. 凹凸性

草 稿 区

10. 决策树的最优分裂的度量通常是根据分裂后子女节点不纯性的程度。下列准则不能用于度量不纯度的是：（ ）

A. Entropy　　B. Gini Index　　C. Rank　　D. Misclassification error

得 分

二 、判断题（本题共 20 分，每小题 2 分）

1. “没有免费午餐”定理表明，针对某具体问题，所有学习算法的期望性能是相同的。（ ）

2. 结构风险最小化策略通常是在经验风险最小化的基础上，增加对模型规模的正则化项。（ ）

3. 在模型训练过程中，验证集的标签是已知的。（ ）

4. h1=<Sunny, ?, ?, Strong, ?, ?>，h2=<Sunny, ?, ?, ?, ?, ?>，则 h2 比 h1 更一般。（ ）

5. 泛化误差的上界与假设空间中所包含的函数数量正相关。在其他条件相同的情况下，假设空间越大，泛化误差上界越大。（ ）

6. 垂直平分分类器不需要假设样本集线性可分。（ ）

7. 已知剪辑法的步骤是（1）预分类（2）剪掉错分样本（3）反复剪辑。则剪辑法的目的是剪掉分类边界上的难分样本。（ ）

8. SVM 的核函数可以将低维特征映射至高维，使样本在高维空间具有显式的特征表达。（ ）

9. 硬间隔最大化 SVM 和软间隔最大化 SVM 的区别在于，后者通过引入核函数的方式解决了样本线性不可分问题。（ ）

10. BP 算法不能优化 4 层及以上的前馈网络。（ ）

草 稿 区

得 分

三 、简答题（本题共 60 分）

1.（本题 12 分）

(1)简述 Fisher 投影准则(6 分)

(2)使用(1)中准则对下表所示的二维数据进行降维，计算最佳单位投影向量(6 分)

A 类		B 类	
x	y	x	y
1	-4	1	0
2	6	-1	-4
-2	-2	-2	-6
-1	0	2	2
0	5	0	-2

草 稿 区

2.(本题 10 分)

(1) 针对频率学派与贝叶斯学派，请分别写出一种对应的机器学习方法的名称 (4 分).

(2) 依据(1)中的例子，简述两种学派观点上的异同(6 分).

草 稿 区

3. (本题 10 分)

(1) 现有如下表所示的数据及初始类中心 C1,C2,C3，请采用欧氏距离作为度量，分别计算前两次循环结束后的类中心（7 分）

(2)简述初始类中心的选取对聚类结果的影响。（3 分）

X	Y	
2	10	C1
2	5	
8	4	
5	8	C2
7	5	
6	4	
1	2	C3
4	9	

4.（本题 12 分）

现给定如下数据及相应定义，请基于信息增益准则，构建决策树.（12 分）

信息熵：$I(p) = -p\ log(p)$

信息增益：$GAIN = I(p) - \sum_{i=1}^{k} \frac{n_i}{n} I(i)$

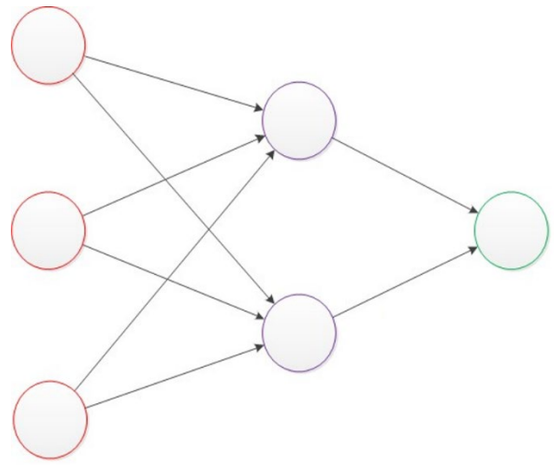
性别	车型	类
男	运动	C0
男	运动	C0
男	豪华	C0
男	运动	C0
男	运动	C0
男	运动	C0
女	运动	C0
女	运动	C0
女	豪华	C0
女	豪华	C0
男	运动	C1
男	运动	C1
男	运动	C1
男	豪华	C1
女	豪华	C1
女	豪华	C1
女	豪华	C1
女	豪华	C1
女	豪华	C1
女	豪华	C1

草　稿　区

5.（本题 16 分）

反向传播法则（BP）是一种常见的神经网络优化算法。现有如下结构的人工神经网络，

其激活函数为 Sigmoid 函数.现给出如下定义:



x_{ji} **:单元** j **的第** i**个输入**

w_{ji} **:与单元** j **相关联的权值**

net_j **:单元** j**的输入的加权和**

o_j **:单元** j**计算出的输出**

t_j **:单元** j**的目标输出**

σ **: sigmoid 函数**

(1) 对于训练样例 d**，试写出关于这个样例的误差函数** E_d**。（4 分）**

(2) 假设单元 j**是网络的一个隐藏单元，单元** j**的输出所能到达单元的集合为** $Down(j)$**，**

根据误差函数 E_d**推导关于样例** d**的梯度修改权值** $\triangle w_{ij}$**（12 分）**

草　稿　区