中南大学考试试卷

2019--2020 学年 下 学期　　时间 100 分钟　2020 年 7 月 7 日

机器学习 课程 48 学时 3 学分　考试形式： 开 卷

专业年级： 大数据 18 级　总分 100 分，占总评成绩 50%

注：此页不作答题纸，请将答案写在答题纸上

一、选择题（本题 30 分，每小题 2 分）

1. 关于 Bagging 算法，下列说法中正确的是(　)

A、如果我们使用每个叶子有一个采样点的决策树，那么 Bagging 算法会给出比一个普通决策树更低的训练误差

B、Bagging 算法对于逻辑回归是无效的，因为所有学习者都学习完全相同的决策边界

C、Bagging 算法的主要目的是减少学习算法的偏差

D、Bagging 算法的训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的

2. 关于 Boosting 算法，下列说法中正确的是(　)

A、每个样例在分类器中权值是根据上一轮的分类结果进行调整

B、使用均匀取样，每个样例的权重相等

C、训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的

D、各个预测函数可以并行生成

3.　以下哪一项是确定性算法的例子？(　)

A、PCA　　B、K-Means

C、以上都是　　D、以上都不是

4. 在神经网络中，非线性激活函数(如 sigmoid，tanh 和 ReLU)的主要作用是(　)

A、与线性单位相比，加速反向传播中的梯度计算

B、仅适用于输出单位

C、有助于学习非线性决策边界

D、始终输出介于 0 和 1 之间的值

5. 下列属于有监督学习算法的是：(　)

A. K-means 算法　　B. 主成分分析 PCA

C. 层次聚类　　D. SVM

6. 下列关于线性回归分析中的残差（Residuals）说法正确的是？（　）

A、残差均值总是为零

B、残差均值总是小于零

C、残差均值总是大于零

D、以上说法都不对

7. 训练一个线性回归模型时，判断下面两句话的是否正确？（　）

1） 如果数据量较少，容易发生过拟合。
2） 如果假设空间较小，容易发生过拟合。

A、 1 和 2 都错误

B、 1 正确，2 错误

C、 1 错误，2 正确

D、 1 和 2 都正确

8. 在其他条件不变的前提下，以下哪种做法容易引起机器学习中的过拟合问题 （ ）

A、 增加训练集量

B、 减少神经网络隐藏层节点数

C、 删除稀疏的特征 S

D、 SVM 算法中使用高斯核/RBF 核代替线性核

9. 假如我们使用非线性可分的 SVM 目标函数作为最优化对象, 我们怎么保证模型线性可分 : （ ）

A、 设 C=1

B、 设 C=0

C、 设 C=无穷大

D、 以上都不对

10. 在训练神经网络时，损失函数(loss)在最初的几个 epochs 时没有下降，可能的原因是？（ ）

A、 学习率(learning rate)太低

B、 正则参数太高

C、 陷入局部最小值

D、 以上都有可能

11. 关于支持向量机 SVM,下列说法错误的是（ ）

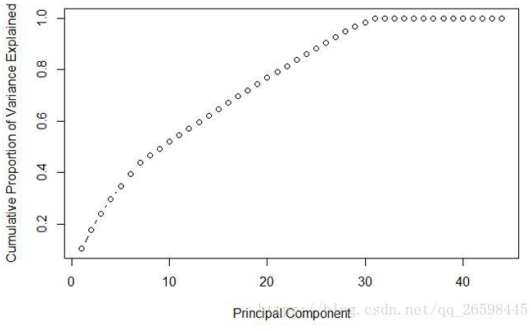
A、 L2 正则项，作用是最大化分类间隔，使得分类器拥有更强的泛化能力

B、 Hinge 损失函数，作用是最小化经验分类错误

C、 分类间隔为 1/||w||，||w||代表向量的模

D、当参数 C 越小时，分类间隔越大，分类错误越多，趋于欠学习

12. 对于右图, 最好的主成分选择是多少 ?: （ ）



A. 7

B. 30

C. 35

D. Can’t Say

13. 梯度下降算法的正确步骤是什么？(）

1. 计算预测值和真实值之间的误差
2. 重复迭代，直至得到网络权重的最佳值
3. 把输入传入网络，得到输出值
4. 用随机值初始化权重和偏差
5. 对每一个产生误差的神经元，调整相应的（权重）值以减小误差 （ ）

A、 1, 2, 3, 4, 5 B、 5, 4, 3, 2, 1

C、 3, 2, 1, 5, 4　　D、4, 3, 1, 5, 2

14. 假设我们要解决一个二类分类问题，我们已经建立好了模型，输出是 0 或 1，初始时设阈值为 0.5，超过 0.5 概率估计，就判别为 1，否则就判别为 0；如果我们现在用另一个大于 0.5 的阈值，那么现在关于模型说法，正确的是：（　　）

A. 模型分类的召回率会降低或不变

B. 模型分类的召回率会升高

C. 模型分类准确率会升高

D. 模型分类准确率会降低

15. 回归模型中存在多重共线性，解决这个问题错误的选择是？（　　）

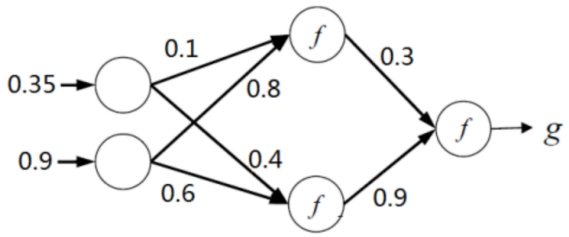
A. 去除共线性变量

B. 我们可以先去除一个共线性变量

C. 计算 VIF(方差膨胀因子)，采取相应措施

D. 为了避免损失信息，我们可以使用一些正则化方法，比如，岭回归和 lasso 回归.

二、神经网络(15 分)。给出如下一个三层的神经网络，并且假设 f（a）=a（即这个函数的导数是 1），损失函数为 $J(W,b)=\frac{1}{2}\|output-y\|^2$，目标值为 0.5，学习率 α =0.5。

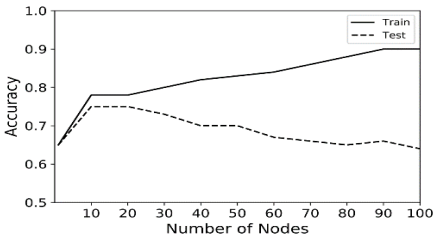


试利用反向传播算法来更新权重并验证是否有效果。

三、过拟合（10 分）。给定有限训练集训练决策树分类器，其在训练集和测试集的准确率随着决策树节点数的变化关系如右图所示

（1） 试根据右图阐述下过拟合和欠拟合的含义；

（2） 若训练集逐步增加到无穷大，试分析训练曲线和测试曲线的趋势；



四、聚类（10 分）。 假定我们对 A、B、C、D 四个样品分别测量两种特征（X1 , X2）和得到的结果见下表：

样品	特征	
	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

试将以上的样品聚成两类。

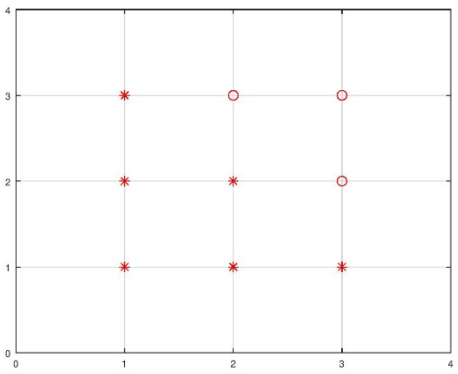
五、Adaboost（10 分）。给定训练集如下图所示，’*’和’o’分别表示正样本和负样本，采用 Adaboost 算法来学习分类器，弱分类器采用 decision stump

（1）画出 Adaboost 选择出的第一个弱分类器 h_1，用实线表示，并在决策边界画出对应类别；

（2）在图中圈出第一轮样本权重更新后权重最大的样本，并计算出样本权重更新后分类器的错误率；

（3）画出 Adaboost 选出的第二个弱分类器 h_2，用虚线表示，并在决策边界画出对应类别；

（4）画出这两个弱分类器组合形成的分类器$H = \text{sgn}(\alpha_1 h_1 + \alpha_2 h_2)$的决策边界，说明其是否可以对所有数据进行正确分类？



六、PCA（10 分）。给定训练集 x_1=[0, 1]^T，x_2=[-1, 0]^T，x_3=[0, 0]^T，x_4=[-1, -2]^T，x_5=[2, 1]^T 试求出对应的两个主元矢量。

七、SVM（15 分）。已知训练数据集 A={x1, x2 , x3, x4}，其正例点是 x2=[2，0]^T，x4=[0，2]^T，负例点是 x1=[-1，1]^T，x4=[1，-1]^T。

（1）试求最大间隔分离超平面。

（2）画图标出所有支持向量、超平面和间隔（margin）；