2018 -- 2019 学年____学期　　时间 100 分钟　2019 年 6 月 29 日

机器学习 课程 48 学时 3 学分 考试形式：开 卷

专业年级：信安、大数据 16 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 10 分，每小题 2 分）

1. 关于 Bagging 算法，下列说法中正确的是(B

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

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.以下哪种策略不能帮助减少决策树中的过度拟合？ B)

A、修剪 ✓

B、确保每个叶节点都是一个纯类

C、在叶节点中强制使用最少数量的样本 ✓

D、强制树的最大深度 ✓

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(C

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. 以下哪项有助于减少 SVM 分类中的过拟合？(A)

A、使用松弛变量 ✓

B、规范化数据

C、高次多项式特征

D、设置非常低的学习率

二、决策树（10 分）。假设拟根据学生以往的 GPA（高、中、低）和重修与否(Studied)这两个属性来预测本次机器学习课程考试的通过(Passed)情况，收集得到的数据集如下表所示

GPA	Studied	Passed
低	否	否
低	是	是
中	否	否
中	是	是
高	否	是
高	是	是

（1）求训练集的信息熵；

（2）基于信息增益，生成对应的决策树。

三、神经网络(15 分)。假设有如下结构的神经网络结构(如图 1 所示)：其中每个神经元为 logistic unit，其权重初始化为

$\Theta_{10}^{(1)} = -0.4,\ \Theta_{11}^{(1)} = 0.2,\ \Theta_{12}^{(1)} = 0.1,\ \Theta_{20}^{(1)} = -0.2,\ \Theta_{21}^{(1)} = 0.4,\ \Theta_{22}^{(1)} = -0.1$

$\Theta_{10}^{(2)} = 0.1,\ \Theta_{11}^{(2)} = -0.2,\ \Theta_{12}^{(2)} = 0.1,\ \Theta_{20}^{(2)} = 0.4,\ \Theta_{21}^{(2)} = -0.1,\ \Theta_{22}^{(2)} = 0.1$

（1）若输入为 $x = [1, 0]^T$，试给出所有隐藏层和输出层神经元的输入；

（2）若 x 对应的 ground-truth 为 $y = [0.9, 0.1]^T$，采用带有 Momentum 的梯度下降法对根据该样本采用反向传播算法更新参数，学习率为 0.1, Momentum $\rho = 0.9$，试给出第一次更新后的参数的值。

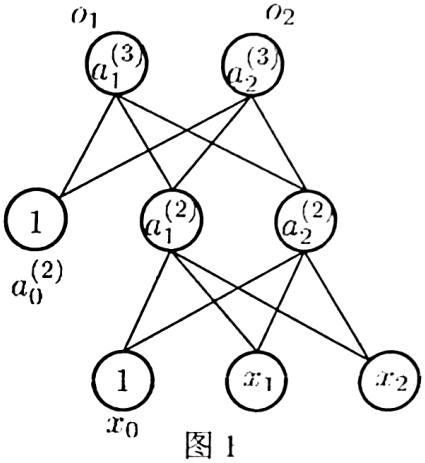

图 1

四、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下图所示

（1）试根据图 2 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；

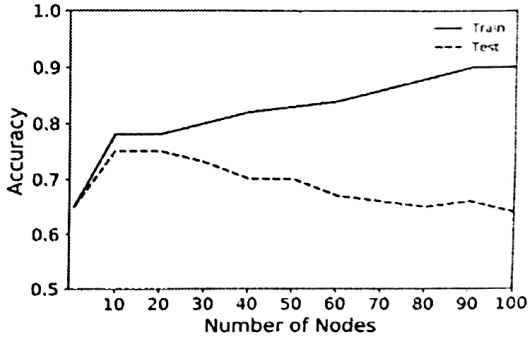

图 2

五、Adaboost（10 分）。给定训练集如图 3 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；

1.297

（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？

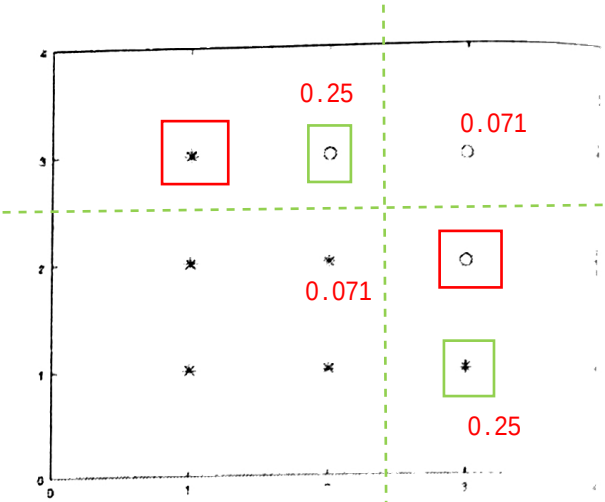

图 3

六、k-means（10 分）。给定训练集 $x_1 = [0, 2]^T$，$x_2 = [5, 0]^T$，$x_3 = [3, 0]^T$，$x_4 = [0, 0]^T$，若假定初始的类均值 $\mu_1 = [0, 0]^T$，$\mu_2 = [2, 0]^T$。请手动给出采用 k-means 聚类的两次迭代结果。

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

八、SVM（15 分）。假设 $x_1 = [-1, 1]^T$，$x_2 = [2, 0]^T$，$x_3 = [1, -1]^T$，$x_4 = [0, 2]^T$ 为确定 hyperplane

$$w^T x + b = 0$$

的四个支持向量，对应的拉格朗日乘子系数分别为 $\alpha_1 = -0.5$，$\alpha_2 = 0.5$，$\alpha_3 = -0.5$，$\alpha_4 = 0.5$（$\alpha < 0$ 表示负样本），分类面的 bias $b = -1$.

（1）在示意图上画出所有支持向量、hyperplane 和 margin；

（2）试用该分类器对 $x = [1, 1]^T$ 进行分类，并在示意图上画出（请详细给出计算过程）；

（3）请给出 hyperplane 的法向量，并写出 hyperplane 的方程。

九、精度-召回率(10 分)。假设一个训练好的 SVM 分类器在验证集上的预测结果如下表所示

No.	预测得分	实际类别
1	7	正
2	4	正
3	2	负
4	1	负
5	-1	负
6	-4	正
7	-5	负
8	-6	负

试计算出若阈值分别取 5, 3, 1, -3, -6 时对应的精度，召回率和 False positive rate.