诚信应考,考试作弊将带来严重后果!

考试中心填写:

___年__月__日

考 试 用

湖南大学课程考试试卷

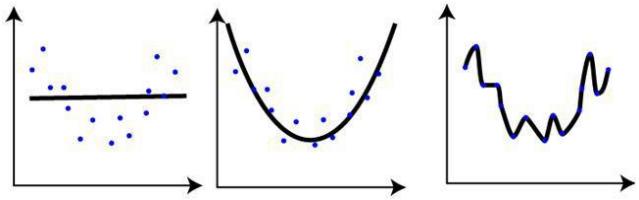
课程名称:**机器学习与大数据分析**;课程编码:CS06171 试卷编号: A ;

考试时间:120 分钟

题 号	一	二	三	四	五						总分
应得分	24	24	12	20	20						100
实得分											
评卷人											

一、 判断与选择题 (每小题 4 分,共 24 分)

1. 回归问题和分类问题都有可能发生过拟合。 (√)
2. 在决策树算法中,属性选择时使用信息增益率比信息增益更加有效。 (√)
3. 单层感知机无法处理异或的表达。 (√)
4. 下图分别给出了三个训练好的回归模型,关于这些模型描述下面说法正确的是( )



(a) (b) (c)

A、图(a)中出现了过拟合 B、图(b)中出现了过拟合

C、图(c)中出现了过拟合 D、无法判断,因为没有验证集或测试集

5. 下列关于bootstrap 描述中正确的是( )

A、有放回地从n 个特征中抽样n'个特征

B、无放回地从n 个特征中抽样n'个特征

C、有放回地从m 个样本中抽样m'个样本

D、无放回地从m 个样本中抽样m'个样本

6. 为了选择参数θ,分别选取不同的θ值进行了训练和验证,对应结果如下表所示。此时最优的选择是( )

参数 θ	训练错误	验证错误
θ = 1	100	90
θ = 2	150	85
θ = 3	100	85
θ = 4	80	120

A、θ=1 B、θ=2

C、θ=3 D、θ=4

湖南大学课程考试试卷

专业班级:

学号:

姓名:

装订线(题目不得超过此线)

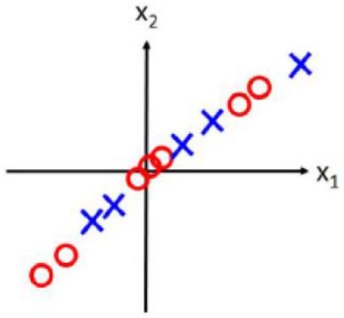
湖南大学教务处考试中心

二、简答题 （每小题12分，共24分）

1. 如何解决类别不平衡问题？在二分类问题中，一类数据有90%，而另一类只有10%。我们可以轻易的得到90%准确率的模型，但是它对第二类的预测值为0。请具体描述我们需要对这样的数据如何进行处理？ （12分）

2. 一个初学机器学习的同学对股市进行预测，他在一个N=1000个股票数据的数据集上匹配了一个有564个参数的模型，该模型能解释数据集上99%的变化，请问该模型能很好的预测其他股票的走势吗？请说明原因。（12分）

三、假设给定训练数据如右图所示，注意到该数据在原始空间中数据线性不可分。我们可以对其如何进行处理使其可用？ （12分）



四、 假定采用深度神经网络模型来诊断用户是否患有视网膜病变，医生的错误率为10%。（本题 20 分）

（1）若当前模型在训练集上的错误率为12%，在验证集上的错误率为30%。请给出下一步最可行的方案，并说明理由。过拟合

（2）若当前模型在训练集上的错误率为25%，在验证集上的错误率为30%。请给出下一步最可行的方案，并说明理由。

欠拟合

1）1. 增加正则化，限制模型复杂度，避免模型过度拟合训练集噪声
2.简化模型结构，可以减少神经元数量、神经网络层数
3.对训练数据进行更多变换，对视网膜图像进行旋转、翻转、亮度调整等，扩充数据集
4.提前终止训练

2）1.延长训练时间，增加训练轮数，让模型充分学习
2.调节模型学习率，优化参数
3.增加模型复杂度，

五、假设给定如右数据集，其中 A、B、C 为二值随机变量，y 为待预测的二值变量。对一个新的输入 A=0, B=0, C=1，朴素贝叶斯分类器将会怎样预测 y？（20 分）

A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1