重庆理工大学本科生课程考试试卷

2022 ～ 2023 学年第 一 学期

开课学院 理学院 课程名称 大数据算法基础 考核方式 闭卷

考试时间 120 分钟 A 卷（A/B/C……） 第 1 页 共 3 页

考生姓名 考生班级 考生学号

一、 简答题（本大题共 20 小题，每小题 2 分，总计 40 分）

1. 请说说你对大数据四个维度当中的“速度”以及“多样性”的理解？
2. 在对大数据进行具体分析之前，我们需要对数据进行清洗，那么请问在做数据清洗时，需要清理哪些数据样本（至少回答两种）？
3. 请分别举出结构化数据和非结构化数据的一个例子。
4. 监督学习要解决的两大类问题是什么？
5. 机器学习模型的三要素是什么？
6. 数据集一般要分为训练集和测试集，那么请问给定一个数据集，该如何合理地分配训练集和测试集中样本的个数？
7. 请说出下边几种情况分别属于什么学习方式？（备选：无监督，监督，强化）（1）某银行要对其客户群体进行聚类分析 （2）永辉超市的打价机器通过物品识别的方式快速的找到相应单价 （3）AlphaZero 学习如何快速地进行矩阵运算。
8. 为什么说线性回归模型有较好的可解释性？
9. 请画出 sigmoid 函数的图形？
10. 请解释随机森林算法中“随机”的含义？
11. 朴素贝叶斯分类器最重要的前提假设是什么？
12. 贝叶斯信念网络的结构$\boldsymbol{G}$和参数$\boldsymbol{\theta}$分别指的是什么？
13. 支持向量机的优化目标是什么？
14. 请分别说出非线性降维和线性降维的一个方法。
15. 马尔可夫决策过程的核心思想是什么？
16. 神经网络模型中 dropout 的作用是什么？
17. 实验课我们利用很多种模型对不同的数据进行分析，那么请简述利用机器学习算法（或大数据分析方法）分析数据的一般流程是什么？
18. 如果数据比较复杂，而模型比较简单的情况下，训练数据会容易发生什么现象？
19. 遗传算法中，为了获得基因更好的个体，对种群中的个体做了哪些操作？
20. 请解释模糊逻辑中隶属度的的概念？

重庆理工大学本科生课程考试试卷

2022 ～ 2023 学年第 一 学期

开课学院 理学院 课程名称 大数据算法基础 考核方式 闭卷

考试时间 120 分钟 A 卷（A/B/C……） 第 2 页 共 3 页

考生姓名 考生班级 考生学号

二、判断题，请判断下面各小题对错，如果错误，请阐述原因。（本大题共 10 小题，每小题 2 分，总计 20 分）

1. 非结构化数据比结构化数据在数量上多出很多。
2. 线形回归模型是用来解决回归问题的。
3. k-means 算法是惰性学习的典型代表。
4. SVM（支持向量机）是一种无监督学习的模型。
5. Bootstrap 抽样中，对于每一个通过抽样获得的子样本集中，不存在重复的样本。
6. 贝叶斯（信念）网络能够成立的前提是特征之间相互独立。
7. 集成学习的 Boosting 方法中，个体学习器是相互独立的，并行生成的。
8. 在 Apriori 算法中，如果一个项集不是频繁项集，那么它的子集也不会是频繁项集。
9. 聚类模型期望的聚类结果是“簇内相似度低，簇间相似度高”。
10. 通过降维可以实现数据可视化。

三、计算题（本大题共 5 小题，每小题 8 分，总计 40 分）

1．已知一个训练数据集，其正例点是$x_1 = (2,2)^T$，$x_2 = (3,2)^T$，负例点是$x_3 = (0,0)^T$,请写出利用支持向量机模型所构造的优化目标以及约束条件。并绘图说明哪些点是支持向量。

2．给定 5 个样本的集合

$$X = \begin{bmatrix} 1 & 0 & 0 & 3 & 2 \\ 0 & 0 & 1 & 3 & 3 \end{bmatrix}$$

试用k均值聚类算法将样本聚到 2 个类中。

3. 已知一个数据集如下表所示,它有三个类别的水果(Banana, Orange, Other)，三个特征（long, Sweet, Yellow）以及 1000 个样本。现给定一个新的样本，并已知它的特征是：Long=True, Sweet=True, Yellow=True,请使用朴素贝叶斯分类模型判断这个样本是属于三个类别（Banana, Orange, Other）当中的哪一类？

重庆理工大学本科生课程考试试卷

2022 ～ 2023 学年第 一 学期

开课学院 理学院 课程名称 大数据算法基础 考核方式 闭卷

考试时间 120 分钟 A 卷（A/B/C……） 第 3 页 共 3 页

考生姓名 考生班级 考生学号

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	400	100	300	200	500
Orange	100	200	100	200	200	100	300
Other	100	100	200	0	100	100	200
Total	600	400	700	300	600	400	1000

贝叶斯公式：$p(C_k \mid \mathbf{x}) = \dfrac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$

4. 给定 5 个样本的集合

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 & 3 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}$$

试用主元分析方法（PCA）找出该数据集的第一主元以及第二主元。

5. 下图是某数据集中的变量 D，I，G，S，L 所构成的贝叶斯（信念）网络，试根据该网络求出变量 D，I，G，S，L 的联合概率分布，并计算出（1）D=False，I=True，G=True， S=False， L=True 发生的概率；（2）如果当 D=True，I=True， G=False 时，L=True 的概率是多少。