**平 时 作 业 报 告**

**课程名称： 大数据与云计算导论**

**学院： 计算机与软件学院**

**专业： 软件工程**

**指导教师： 陆玉武**

**报告人： 郑彦薇 学号：2020151022**

**时间： 2021年4月12日**

**报告提交时间： 2021/4/13**

|  |
| --- |
| **说明：**  （1）本次作业满分为100分。  （2）本次作业截至时间2021年4月23日（周五）23:59。  （3）个人信息：WORD文件名中的“姓名”、“学号”，请改为你的姓名和学号；报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“实验报告提交时间”等信息。  （5）提交方式：截至时间前，请在Blackboard系统中提交，延迟提交无效。  （6）发现抄袭（包括复制&粘贴整句话、整张图），抄袭者和被抄袭者的总成绩记零分，可以参考网络资源。  （7）期末考试阶段补交无效。  （8）提交作业的命名方式为：**大数据与云计算导论-学号-姓名**，请按规定命名并按时提交作业。 |
| **（1）关系数据库的特点是什么？NoSQL数据库的特点是什么？（20分）**  **答：【关系数据库的特点】关系数据库建立在关系数据模型之上，主要用来存储结构化数据并支持数据的插入、查询、更新、删除等操作。关系模型将数据组织成一系列由行和列构成的二维表格，通过关系代数，关系演算等方法来处理表格中的数据，并且可以利用多种约束来保证数据的完整准确。通过标准化的结构化查询语言（SQL），关系数据库中的数据能被灵活地组合、拆分、转换，这使得关系数据库管理系统RDBMS的用户和应用能够非常方便地处理其中的数据。**  **【NoSQL数据库的特点】NoSQL数据库是对于非关系型的一类数据库系统的统称。关系数据库在管理键值对、文档 、图等类型数据上有所不足，NoSQL则是针对各个类型数据的存储和访问特点而专门设计的数据库管理系统。近年来，随着大数据应用的不断扩展，NoSQL数据库系统得到了广泛应用；NoSQL数据库通过采取一些新的设计原则，利用大规模计算机集群实现对大数据的有效管理。**  **（2）大数据分析的方法分为哪些类？统计分析方法中，哪些指标可以描述数据的离散程度？（20分）**  **答：【大数据分析的方法】统计分析方法，基于机器学习的数据分析，图的数据分析，自然语言中的数据分析。**  **统计分析方法中，方差和标准差可以描述数据的离散程度。这两种指标的值越大，数据的分散性程度越高。**  **（3）什么是非监督学习？什么是监督学习？请给出训练集、验证集、测试集的作用。（20分）**  **答：【非监督学习】：非监督学习是建立在所有数据标签，即所属的类别都是未知的情况下使用的分类方法。**  **【监督学习】：监督学习是利用已知的分类信息，得到一些更精准的分类方法。**  **【训练集、验证集、测试集的作用】：定义测试集为需要使用算法来帮助处理分类的数据。除了自动学习的部分以外，机器学习模型一般会有一些需要人工设置的超参数，对于同样的模型和训练集，使用不同的超参数进行学习，最后得出的结果可能相差很大。为了在测试集上取得尽可能好的结果，但又不知道在测试集上数据的真实分类标签，一般需要在知道真实情况的数据集中划分训练集和验证集。在训练模型时，只使用训练集中的数据，而使用验证集中的数据进行模型正确率的检验。这样，通过验证集上的检验，就可以对于不同超参数在当前任务下的好坏有一个大致的了解，从而选择正确的超参数训练模型并在测试集上取得好结果。**  **（4）哪些统计分析方法可以分析数据两个特征之间的关系？它们有什么区别和联系？（20分）**  **答：数据描述性分析与回归分析；数据描述性分析能够概括数据位置特性，分散性、关联性等数字特征，以及能够反映数据整体分布特征；回归分析通过线性回归模型可以使用已知数据去估计未知参数**  **（5）决策树有哪些构建方法？它们有何不同？（20分）**  **答：使用ID3算法进行构建；使用C4.5算法进行构建；**  **不同：相比于ID3算法，C4.5算法继承了ID3算法的优点，并在此基础上进行改进：利用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；在树构造过程中进行剪枝；能够完成对连续属性的离散化处理；能够对不完整数据进行处理。** |

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |