第1章

一、简答题

1. 大数据是指数量非常庞大、种类非常复杂，无法使用常规软件工具和技术手段进行采集、管理和处理的数据集合。

2. 非结构性、不完备性、时效性、完全性以及可靠性。

3. 大数据有五个基本特征：数据规模大(volume)、数据种类多(varity)、数据要求处理速度快(velocity)、数据价值密度低(value)、数据的真实性(veracity)，即所谓的5V特性。

4. 结构化数据是指具有数据结构描述信息的数据，非结构化数据是指不方便用固定结构来表现的数据。前者先有结构，再有数据，且主要以种类表格形式呈现；后者只有数据，没有结构，主要以图形、图像、音频、视频信息形式呈现。

5. 数据处理的过程可以概括为五个步骤，分别是数据采集与记录，数据抽取、清洗与标记，数据集成、转换与归约、数据分析与建模，数据解释与应用。

6. 数据挖掘(data mining)是采用数学的、统计的、人工智能和神经网络等领域的科学方法，如记忆推理、聚类分析、关联分析、决策树、神经网络、基因算法等，从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。

当前的主要功能如下：

(1)分类：按照分析对象的属性、特征，建立不同的组类来描述事物。

(2)聚类：识别出分析对象的内在规则，按照这些规则把对象分成若干类。

(3)关联规则：关联是某种事物发生时其他事物会发生的一种联系。

(4)预测：把握分析对象发展的规律，对未来的趋势做出预见。

(5)偏差的检测：对分析对象的少数的、极端的特例的描述，揭示内在的原因。

7. (1)数据采集：系统日志采集方法、网络数据采集方法、其他数据采集方法。

(2)数据清洗：遗漏数据处理、噪声数据的识别、对不一致数据检测、数据过滤与修正等。

(3)数据存储：根据对一致性要求的强弱不同，数据存储策略分为ACID和BASE两种。

(4)数据集成：将相互关联的分布式异构数据源集成到一起，使用户能够以透明的方式访问这些数据源。

(5)数据转换：基于规则或元数据的转换、基于模型与学习的转换等。

(6)数据规约：维归约、数据归约、数据抽样等技术。

(7)数据分析与挖掘：已有数据的分布式统计分析技术和未知数据的分布式挖掘、深度学习技术。分析方法主要包含假设检验、显著性检验、显著性分析、相关分析、T检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析(简单回归分析、多元回归分析)、逐步回归、回归预测与残差分析、曲线分析、因子分析、聚类分析、主成分分析、判别分析、对应分析、多元对应分析等；数据挖掘方法包括 k-means 聚类算法、SVM统计学习算

法和 NaiveBayes 分类算法等。

(8)数据可视化：利用云计算、标签云、关系图等呈现。

8. 大数据的数据类型丰富，包括结构化数据和非结构化数据，其中，前者占 10%左右，主要是指存储在关系数据库中的数据，后者占 90%左右，种类繁多，主要包括邮件、音频、视频、微信、微博、位置信息、链接信息、手机呼叫信息、网络日志等。

9. 电商大数据、医疗大数据、教育大数据、金融大数据、农业大数据、旅游大数据、气象大数据。大数据还在城市建设与规划、环境保护、电力行业、零售、交通、公共设施、舆情监控、城市治理等方面也得到了广泛应用，涉及现代社会管理、生活、服务等各个方面。

10. 大数据是一把双刃剑。一方面，海量数据的迅速增长为社会发展提供了更多宝贵的数据资源。网络和数据库中所记载的各种巨量数据，是现实生产劳动的真实反映。人们可以利用这些数据分析问题、解决问题，并且促成新的理论和技术。伴随着数据处理能力的提升、运算与存储成本的井喷，以及越来越多的设备中嵌入各种传感技术，便利数据的收集、存储与分析正处于一个近乎无限上升的趋势。另一方面，大数据前所未有的运算能力也给人们带来了挑战，不可控的持续爆炸式增长的大数据正向人们的数据中心基础设施和数据处理及分析的各个环节发起严峻的挑战，也给人们的法律、伦理及社会规范发起挑战，考验人们能否在大数据的世界中保护隐私和其他价值观。同时企业对大数据人才需求激增，复合型人才需求更甚。

二、选择题

1. A　2. D　3. A　4. ABC　5. ABCD　6. AC　7. ABC