《数据科学与大数据通识导论》题库及答案

1. 数据科学的三大支柱与五大要素是什么？

答：数据科学的三大主要支柱为：

Datalogy (数据学)：对应数据管理 (Data management)

Analytics (分析学)：对应统计方法 (Statistical method)

Algorithmics (算法学)：对应算法方法 (Algorithmic method)

数据科学的五大要素：

A-SATA模型

分析思维 (Analytical Thinking)

统计模型 (Statistical Model)

算法计算 (Algorithmic Computing)

数据技术 (Data Technology)

综合应用 (Application)

1. 如何辨证看待“大数据”中的“大”和“数据”的关系？

字面理解

Large、vast和big都可以用于形容大小

Big更强调的是相对大小的大，是抽象意义上的大

大数据是抽象的大，是思维方式上的转变

量变带来质变，思维方式，方法论都应该和以往不同

计算机并不能很好解决人工智能中的诸多问题，利用大数据突破性解决了，其核心问题变成了数据问题。

1. 怎么理解科学的范式？今天如何利用这些科学范式？

科学的范式指的是常规科学所赖以运作的理论基础和实践规范，是从事某一科学的科学家群体所共同遵从的世界观和行为方式。

第一范式：经验科学

第二范式：理论科学

第三范式：计算科学

第四范式：数据密集型科学

今天，是数据科学，统一于理论、实验和模拟

1. 从人类整个文明的尺度上看，IT和DT对人类的发展有些什么样的影响和冲击？

以控制为出发点的IT时代正在走向激活生产力为目的的DT（Data Technology）数据时代。

大数据驱动的DT时代

由数据驱动的世界观

大数据重新定义商业新模式

大数据重新定义研发新路径

大数据重新定义企业新思维

1. 大数据时代的思维方式有哪些？

“大数据时代”和“智能时代”告诉我们：

数据思维：讲故事 🡪 数据说话

总体思维：样本数据 🡪 全局数据

容错思维：精确性 🡪 混杂性、不确定性

相关思维：因果关系 🡪 相关关系

智能思维：人 🡪 人机协同（人 + 人工智能）

1. 请列举出六大典型思维方式；

直线思维、逆向思维、跳跃思维、归纳思维、并行思维、科学思维

1. 大数据时代的思维方式有哪些？

同5

1. 二进制系统是如何实现的？

计算机用0和1来表示和存储所有的数据，它的基数为2，进位规则是“逢二进一”，用1表示开，0表示关

1. 解释比特、字节和十六进制表示。

比特：一位 0、1 ；字节：8比特，从000到111，十六进制：0000到1111十六个数分别用0,1,2,3,4,5,6,7,8,9，A，B，C,D,E,F 来表示

1. 请辨析现象、数据、信息和知识这几个概念。

信息是关于世界、人和事的描述，它比数据来得抽象；而数据则是信息的载体。

知识比信息更高一个层次，也更加抽象，它具有系统性的特征。

比如通过测量星球的位置和对应的时间，就得到数据；通过这些数据得到星球运转的轨迹，这就是信息；通过信息总结出开普勒三定律，就是知识。

从现象、数据到信息、知识，抽象层次是越来越高的。

人类的进步就是靠使用知识不断地改变我们的生活和周围的世界，而数据是知识的基础。

1. 简述冯·诺依曼计算机工作原理

存储程序控制

程序和数据都用二进制数表示

机器以CPU为中心

1. 简述GB2312编码和Unicode编码原理;

GB2312编码适用于汉字处理、汉字通信等系统之间的信息交换，通行于中国大陆;新加坡等地也采用此编码。中国大陆几乎所有的中文系统和国际化的软件都支持GB 2312。

基本集共收入汉字6763个和非汉字图形字符682个。整个字符集分成94个区，每区有94个位。每个区位上只有一个字符，因此可用所在的区和位来对汉字进行编码，称为区位码。

Unicode(统一码、万国码、单一码)是计算机科学领域里的一项业界标准，包括字符集、编码方案等。Unicode 是为了解决传统的字符编码方案的局限而产生的，它为每种语言中的每个字符设定了统一并且唯一的二进制编码，以满足跨语言、跨平台进行文本转换、处理的要求。

1. 简述摩尔定律。

摩尔定律是由英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）提出来的。其内容为：当价格不变时，集成电路上可容纳的元器件的数目，约每隔18-24个月便会增加一倍，性能也将提升一倍。换言之，每一美元所能买到的电脑性能，将每隔18-24个月翻一倍以上。这一定律揭示了信息技术进步的速度。

1. 为什么计算机系统要往并行与异构的方向发展？

单核CPU已经走到尽头，不能承载更多的晶体，所以采用多核和GPU及FPGA等并行与异构，多台计算机一起工作，进一步提升计算性能

1. 什么是云计算？

云计算的定义：

云计算是一个模型，这个模型可以方便地按需通过网络访问一个可配置的计算资源（例如，网络、服务器、存储设备、应用程序以及服务）的公共及。这额资源可以迅速提供并发布，同时最小化管理成本或服务提供商的干涉。

1. 为什么说数据上云是一种趋势？

大数据上云，数据上云，分析上云，人工智能上云，数据上云是一种趋势

1. 从技术体系和资源结构两方面谈一下云计算的逻辑平台组成。

云平台的逻辑组成

技术体系：IaaS, PaaS, SaaS（Technically）

资源结构：计算、存储、网络

1. 数据获取过程可分为哪几个步骤？

数据获取阶段的任务是以数字形式将信息聚合, 以待存储和分析处理，数据获取过程可分为三个步骤：

数据采集

数据传输

数据预处理

1. 数据预处理需要做哪些工作？

数据预处理

由于数据源的多样性，数据集由于干扰、冗余和一致性因素的影响具有不同的质量。

一些数据分析工具和应用对数据质量有着严格的要求。因此在大数据系统中需要数据预处理技术提高数据的质量。

数据集成（Data integration）

数据清洗（Data cleansing）

冗余消除（Redundancy elimination

1. 数据的存储方式有哪些？

数据的存储方式：纸带

磁带

数字存储

随机存取存储器（Random access memory，RAM）

磁盘（HDD）和磁盘阵列

存储级存储器：闪存、SSD

光盘

1. 什么是大数据的计算模式？

所谓大数据计算模式，即根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象（abstraction）或模型（model）。

1. 简述大数据的四种主要分析技术。

大数据的四种主要分析技术

（1）统计分析

（2）机器学习

是一门研究机器获取新知识和新技能，  
并识别现有知识的学问。

（3）数据挖掘

从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

（4）可视化分析

统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术。

1. 简述数据库的事务ACID性质

数据库的事务处理（ACID）性质：

原子性（Atomicity）：不可分割

一致性（Consistency）：前后一致

隔离性（Isolation）：并发

持久性（Durability）：永久

1. 关系代数的五种基本运算分别是什么？

关系代数的五种基本运算分别是并、差、投影、选择、笛卡尔积

1. 简述数据仓库的四种类型。

数据仓库的四种类型：

传统数据仓库

实时处理数据仓库

关联发现数据仓库

数据集市

1. 什么是GFS?请简述GFS的特点？

GFS是一个可扩展的分布式文件系统，用于大型的、分布式的、对大量数据进行访问的应用。它运行于廉价的普通硬件上，将服务器故障视为正常现象，通过软件的方式自动容错，在保证系统可靠性和可用性的同时，大大减少了系统的成本。

系统分为三类角色：

Master（主服务器）

Chunk Server（数据块服务器）。

Client（客户端）

1. 请简述HDFS的特点。

Hadoop是一个分布式系统基础架构，由Apache基金会开发。用户可以在不了解分布式底层细节的情况下，开发分布式程序，充分利用集群的威力高速运算和存储。Hadoop实现了一个分布式文件系统（Hadoop Distributed File System），简称HDFS。

HDFS有着高容错性的特点，并且设计用来部署在低廉的硬件上。

HDFS 集群包含：

一个主节点（ NameNode ）

多个从属节点（DataNodes ）

多个客户端访问

1. 大数据系统为什么要采用分布式的架构？

大数据是数据库的自然延伸：分布式。

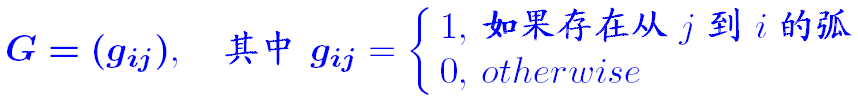
由于大数据要处理大量、非结构化的数据，所以在各处理环节中都可以采用并行处理。

目前，Hadoop、MapReduce和Spark等分布式处理方式已经成为大数据处理各环节的通用处理方法。

1. 什么是有向图、邻接矩阵、超链接矩阵。

一个有向图D是指一个有序三元组(V(D)，A(D)，ψD)，其中ψD)为关联函数，它使A(D)中的每一个元素(称为有向边或弧)对应于V(D)中的一个有序元素(称为顶点或点)

为研究需要，我们定义邻接矩阵



进一步，如果将邻接矩阵中的元素除以对应节点的出度，可以得到该图的超链接矩阵

1. 如何计算PageRank的值?

PageRank算法

第一步：将互联网作为一个有向图，并用邻接矩阵进行表示；

第二步：将该邻接矩阵转换为超链接矩阵；

第三步：求解该超链接矩阵的最大特征向量（如幂迭代法）；

第四步：求得的特征向量中的值即为对应网页的PageRank值。

1. 典型的数据挖掘技术有哪些？

典型的数据挖掘技术：关联分析、序列模式、分类（预言）、聚集、异常检测

1. 数据挖掘的标准流程分为那几个步骤？请分别简述它们。

商业理解：找问题－确定商业目标、 对现有资源的评估， 确定问题是否能够通过数据挖掘来解决， 确定数据挖掘的目标， 制定数据挖掘计划

数据理解：

数据准备：确定数据挖掘所需要的数据， 对数据进行描述， 数据的初步探索， 检查数据的质量

建立模型：对各个模型进行评价， 选择数据挖掘模型， 建立模型

模型评估：评估数据挖掘的结果， 对整个数据挖掘过程的前面步骤进行评估，确定下一步怎么办？是发布模型？还是对数据挖掘过程进行进一步的调整，产生新的模型

模型发布：把数据挖掘模型的结果送到相应的管理人员手中， 对模型进行日常的监测和维护， 定期更新数据挖掘模型

1. 典型的分类方法有哪些？

典型分类方法：判定树归纳分类，贝叶斯分类，后向传播分类，k-最临近分类，基于案例的推理，遗传算法，粗糙集方法，模糊集方法

1. 典型的预测方法有哪些？

典型预测方法：回归方法，线性回归，多元回归，非线性回

1. 请简述分类过程。

数据分类：两步过程

第一步，建立一个模型，描述预定数据类集和概念集

假定每个元组属于一个预定义的类，由一个类标号属性确定

基本概念

训练数据集：由为建立模型而被分析的数据元组形成

训练样本：训练数据集中的单个样本（元组）

学习模型可以用分类规则、判定树或数学公式的形式提供

第二步，使用模型，对将来的或未知的对象进行分类

首先评估模型的预测准确率

对每个测试样本，将已知的类标号和该样本的学习模型类预测比较

模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比

测试集要独立于训练样本集，否则会出现“过分适应数据”的情况

1. 简述预测和分类的相同点和不同点。

预测是构造和使用模型评估无样本类，或评估给定样本可能具有的属性或值空间。

预测和分类的相同点：

两者都需要构建模型

都用模型来估计未知值

预测当中主要的估计方法是回归分析

线性回归和多元回归

非线性回归

预测和分类的不同点：

分类法主要是用来预测类标号（分类属性值）

预测法主要是用来估计连续值（量化属性值）

1. 什么是离群点分析？离群点分析可以应用在哪些领域？

离群点分析

离群点:一些与数据的一般行为或模型不一致的孤立数据

通常孤立点被作为“噪音”或异常被丢弃，但在欺骗检测中却可以通过对罕见事件进行孤立点分析而得到结论。

应用：信用卡欺诈检测，移动电话欺诈检测，客户划分，医疗分析（异常）

1. 什么是异常检测？异常监测可以应用在哪些领域？

异常检测是数据挖掘中一个重要方面，用来发现”小的模式”(相对于聚类)，即数据集中间显著不同于其它数据的对象。

异常探测应用：电信和信用卡欺骗，贷款审批，药物研究，气象预报，金融领域，客户分类，网络入侵检测，故障检测与诊断等

1. 列举四种机器学习方法中的监督学习算法并对它们进行简要解释。

监督学习算法：线性回归，逻辑回归，神经网络，SVM

回归算法有两个重要的子类：即线性回归和逻辑回归

神经网络(也称之为人工神经网络，ANN)，是80年代机器学习界非常流行的方法，其诞生起源于对大脑工作机理的研究。简单来说，就是分解与整合

SVM（支持向量机）

1. 什么是高级编程语言？

高级编程语言：为了克服低级语言的缺点，科学家设计出了更加易用的高级编程语言（high-level programming language）。

高级语言吸收了人们熟悉的自然语言和数学语言的某些成分，因此非常易学、易用、易读；

高级语言在构造形式和意义方面具有严格定义，从而避免了语言的歧义性；

高 级语言与计算机硬件没有关系，用高级语言写的程序可以移植到各种计算机上执行。

1. 列举出10大数据挖掘的算法。

数据挖掘领域的十大经典算法：

C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART.

1. 简述深度学习、机器学习、人工智能三者的关系。

人工智能包括了机器学习，机器学习包括了深度学习

1. 三大数据编程语言分别是哪三种？它们分别适用于哪些场景？

R语言:最常用数据分析工具之一，兼容性强

Python:编程类数据分析，如文本字符等非机构化数据的处理

SQL： 侧重数据库方面，如数据仓库等，作为Oracle等数据库方面的基础知识不可或缺

1. 数据工程的主要研究内容可分为哪三个方面？请分别简述它们。

数据工程的主要研究内容可概括为三个方面:

数据产品的创新，具体包括非结构化数据的结构化、数据衍生品的创造；

数据产品及其数据衍生品的价值分析，包括数据产品及其数据衍生品内在关系（线性关系与非线性关系）的识别，数据产品及其数据衍生品价值评价模型的构建；

建立在数据有效性基础上的理论探讨，包括理论体系的构建，定律与模型的发现等。

1. 标准的数据科学过程可分为哪六步？

（1）Setting the research goal

（2）Retrieving data

（3）Data preparation

（4）Data exploration

（5）Data modeling

（6）Presentation and automation

1. 为什么要开放公共数据？

把数据当原料应用价值，数据是一种基础设施，开放公共数据本质上是提供一种公共产品，促进合作共创，通过开放数据，利用数据，解决问题创造价值，让社会，供给方，利用方合作共赢。

1. 什么是智慧城市，智慧城市有哪些应用？

智慧城市就是运用信息和通信技术手段感测、分析、整合城市运行核心系统的各项关键信息，从而对包括民生、环保、公共安全、城市服务、工商业活动在内的各种需求做出智能响应。其实质是利用先进的信息技术，实现城市智慧式管理和运行，进而为城市中的人创造更美好的生活，促进城市的和谐、可持续成长。

智慧城市有智慧政府、智慧交通、公共安全等应用

1. 人工智能中的“人工”和“智能”分别怎么理解？

人工智能的定义可以分为两部分，即“人工”和“智能”。“人工”比较好理解，争议性也不大。有时我们会要考虑什么是人力所能及制造的，或者人自身的智能程度有没有高到可以创造人工智能的地步，等等。但总的来说，“人工系统”就是通常意义下的人工系统。

关于什么是“智能”，就问题多多了。这涉及到其它诸如意识（CONSCIOUSNESS）、自我（SELF）、思维（MIND）（包括无意识的思维（UNCONSCIOUS\_MIND））等等问题。人唯一了解的智能是人本身的智能，这是普遍认同的观点。但是我们对我们自身智能的理解都非常有限，对构成人的智能的必要元素也了解有限，所以就很难定义什么是“人工”制造的“智能”了。因此人工智能的研究往往涉及对人的智能本身的研究。其它关于动物或其它人造系统的智能也普遍被认为是人工智能相关的研究课题。

1. AI的五个条件是什么？

AI的五个条件：清晰的领域界限，自动标注数据，超大计算量，海量数据，顶尖AI科学家。

1. 随着人工智能技术的发展，人类和智能机器之间会是一个什么样的关系？

四大趋势：

关键性任务的人工智能 (Mission-critical AI)

个性化人工智能 (Personalized AI)

跨多组织机构的人工智能 (AI across organizations)

后摩尔定律时期的人工智能 (AI demands outpacing the Moore‘s Law)