简答题候选 4 道题，出其中 2 道题。

简述“乱数据”和“脏数据”的概念及两者的区别。

数据问题可分为两种，一种是数据质量有缺陷，这种数据叫做脏数据（dirty data）经过数据清洗（data cleansing）后变为干净数据（clean data）。还有一种是数据模态不符合计算要求，这种数据叫做 vm 乱数据（messy data）经过数据规整化（data tidying）变为规整数据（tidy data）

简述机器学习中“参数”和“超参数”的概念及两者的区别。

参数（Parameter）：可分为算法参数和模型参数。其中，算法参数又称为“超参数”，而模型参数统简称为“参数”。（模型）参数用于描述一个具体的模型。通常，同一个算法所训练出的模型的参数个数和类型是一致的，区别在于参数取值。

超参数（Hyper-parameter）：控制机器学习过程并确定学习算法最终学习的模型参数值的参数。例如：训练集和测试集的分割比例、优化算法中的学习率、聚类算法中的聚类数、损失函数的选择、神经网络学习中的激活函数的选择、隐藏层数及迭代次数等。

简述数据缺失的三种类型，并举例说明。

完全随机缺失（MCAR）：缺失数据与该变量的真实值无关，与其他（观测或未观测）变量也无关。例：老师抱着批改完的卷子，不小心摔倒丢失了几张卷子，导致几位同学没有成绩。这种缺失不是因为成绩这个变量本身高或低而丢失的，也与姓名等无关。

随机缺失（MAR）：缺失数据与其他观察变量有关，但与未观测变量无关。例：要统计某班学生的基本信息，包括名字、性别、身高、体重等。如果某学生的体重这一变量缺失，则这一事件最可能发生在女生，即与已知变量性别相关。

非随机缺失（NMAR）：缺失数据依赖于该变量本身。例：收集数据时，收入一栏很容易缺失，发生这种情况的原因可能是填写人收入过高或过低。

试对 CAP 理论给予比较全面的解释。

CAP 理论的基本思想如下：一个分布式系统不能同时满足一致性（Consistency）、可用性（Availability）和分区容错性（Partition Tolerance）等需求，而最多只能同时满足其中的两个特征。

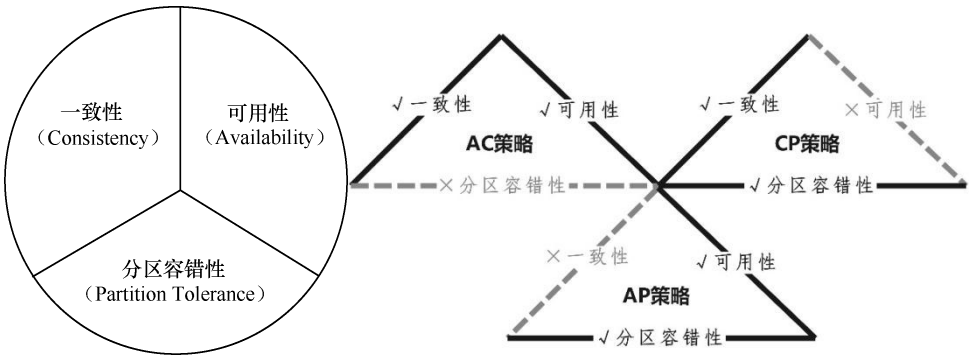
CAP 理论告诉我们，数据管理不一定是理想的——一致性、可用性和分区容错性中的任何两个特征的争取，可能导致另一个特征的放弃，如图 6-11 所示。

（1）一致性。指强一致性，即所有节点都具有最新的数据副本。客户在哪个节点都能得到相同的最新数据。

（2）可用性。系统响应请求的能力，即（非故障）节点收到的每个请求，总能在“给定时间”之内得到“所需要的结果”。

（3）分区容错性。即使发生网络分区，系统也能继续运行的能力。

解释：①没有发生网络分区（P）时，一致性（C）和可用性（A）可以得到保证；②但一旦发生 P，就要在 C 和 A 之间选择之一；③如果选择 C，则在解决 P 之前，系统不可用（A）；如果优先选择 A，则更新的数据就无法到达断开的节点，导致数据不一致（C）。



大家对重要词汇的英文原词（仅限课件上注释了的）要有所掌握。

关联规则、假设检验、KNN 也要会

第一章

1、**DIKW 模型：**

智慧：运用知识，并结合经验创造性的预测、解释、发现等

知识：从多条消息中发现的共性规律、模式、模型、理论、方法等

信息：尤其是多条数据所共同反应的现实世界中的现象

数据：现实世界的记录

2、在数据科学中，各种符号（如字符、数字等）的组合、语音、图形、图像、动画、视频、多媒体和富媒体等统称为**数据（Data）**

3、**DIKUW 模型**分别对应了 **Data → Information → Knowledge → Understanding → Wisdom** 也就是数据，信息，知识，理解，智慧

4、**结构化数据**是直接可以用传统关系数据库存储和管理的数据，先有结构，后有数据；

非结构化数据是无法用关系数据库存储和管理的数据，是没有或难以发现统一结构的数据，比如语音，图像文件等；

半结构化数据是经过一定转换处理后可以用传统关系数据库存储和管理的数据，先有数据，后有结构（或比较容易发现其结构）比如 html、xml 文件等

5、**原始数据（没有经过预处理）**：往往存在缺失值、噪声、错误或虚假数据等质量问题，是零次数据

干净数据（预处理过的数据）：经过清洗、变换、集成等处理，是一次数据

增值数据（分析处理的结果）：经过深度处理或分析，包括脱敏、规约、标注等，是二次数据

洞见数据（直接可以用于决策）：进行洞察分析，包括：统计分析，数据挖掘，机器学习，可视化分析等，是三次数据

6、数据：对客观事物或现象直接记录下来后产生的数据。

例：**一本书的内容**

元数据：数据的数据，可以是数据内容的描述信息。

例：**一本书的作者、出版社、...**

数据对象：对数据内容与其元数据进行封装或关联后得到的更高层次的数据集。

7、**涌现**：再从小数据演变为大数据的过程中，出现了一种名为涌现（Emergence）的现象，，涌现是指系统的整体性能或特性大于其组成元素之和，在不同的层次结构中表现出新的质量

8、大数据的 **4V 特征：Volume Variety Value Velocity**，也即数据量大，类型多，价值密度低，速度快，四个单词的意思为数据量 多样性 价值密度 处理速度

9、数据问题可分为两种，一种是数据质量有缺陷，这种数据叫做脏数据（dirty data）经过数据清洗（data cleansing）后变为干净数据（clean data）。还有一种是数据模态不符合计算要求，这种数据叫做乱数据（messy data）经过数据规整化（data tidying）变为规整数据（tidy data）

10、数据科学处于数学与统计知识、3C 精神与技能、领域实务三大领域的交叉点

11、3C 精神，即创造性地做事，批判性思考，好奇性地提出问题，也即
Creative Woriking、Critical Thinking、Curious Asking

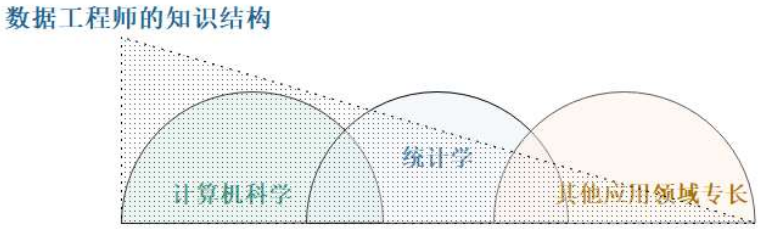
12、描述性分析：一种将数据转换为信息的分析过程 Descriptive analysis
预测性分析：一种将信息转换为知识的分析过程 Predictive analysis
规范性分析：一种将知识转换为智慧的分析过程 Prescriptive analysis
数据 信息 知识 智慧

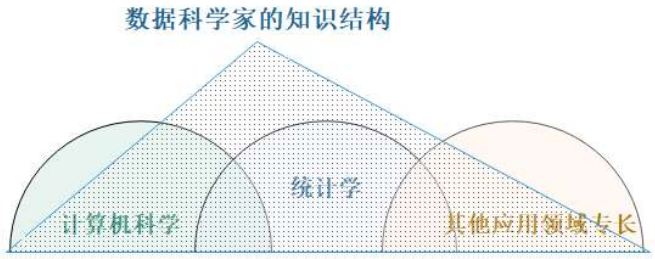
13、第一范式：经验科学，如伽利略斜塔落体实验，斜面实验

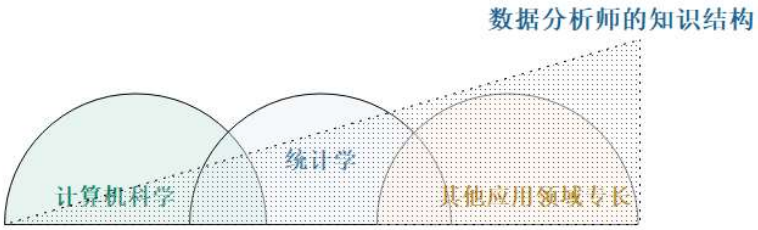
第二范式：理论科学，如热力学第一定律，开普勒从第谷的天文观侧资料中总结出行星运动三大定律，并被进一步的理论和实践所证实

第三范式：计算科学，如计算机模拟仿真，密度泛函理论，分子动力学
第四范式：数据密集型科学发现，如预测分析，聚类，关系挖掘，异常检测

14、







15、**探索型数据分析**（Exploratory Data Analysis，**EDA**）

探索型数据分析是指对已有的数据（特别是调查或观察得来的原始数据）在尽量少的先验假定下进行探索，并通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

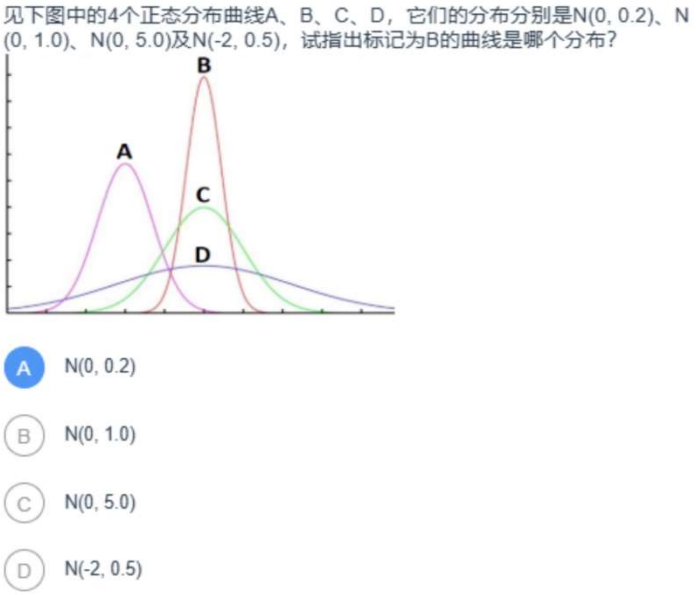
当数据科学家对数据及其相关业务没有足够的经验，且不确定应该采用何种传统统计方法进行分析时，经常通过探索型数据分析方法达到数据理解的目的。

第二章

1、从思维方式看，传统统计学方法可以分为两大类—描述统计和推断统计
描述统计：集中趋势分析，离中趋势分析，相关分析
推断统计：常用两种 参数估计，假设检验（还有抽样分布不是很常见）

2、总平方和（SST）=回归平方和（SSR）+ 残差平方和（SSE）

3、方差越小，正态分布的图像越高



4、假设检验经常存在两类错误，即“弃真错误”和“取伪错误”，在统计学中分别称为 α 错误和 β 错误

5、

- **通常，样本集x和y的相关系数r的计算公式如下：**

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\cdot\sum(y-\bar{y})^2}}$$

式中，$\bar{x}$和$\bar{y}$分别为样本集x和y的均值；r 为样本相关系数。

6、时间序列的四个要素有：趋势，季节变动，循环波动，不规则波动；只含有随机波动的时间序列也称为平稳序列

7、**假设检验（A/B 测试）**
定义：假设检验是一种统计方法，用于判断样本数据是否支持某个假设（如两组数据是否存在显著差异）。其核心思想是通过计算 P 值与显著性水平（如 α =0.05）比较，决定是否拒绝原假设（H0）。
关联规则分析（如 Apriori 算法）
定义：用于发现数据中频繁出现的项集及其关联规则（如“购买 X 的同时常购买 Y”）。常用指标包括支持度（项集出现频率）和置信度（规则成立的概率）

示例：超市购物数据中，发现“啤酒→尿布”的规则，支持度=10%（10%交易同时包含两者），置信度=70%（买啤酒的用户 70%会买尿布），据此调整货架摆放

决策树分类（如 ID3/C4.5 算法）

定义：通过树形结构对数据进行分类，每个节点基于特征属性（如年龄、收入）进行划分，叶节点表示类别。常用算法包括 ID3（信息增益）、C4.5（信息增益率）和 CART（基尼指数）

示例：银行用决策树判断是否批准贷款，根据“收入>5 万”“信用评分>700”等条件逐层分支，最终输出“批准”或“拒绝”

聚类（如 K-means 算法）

定义：将相似数据分到同一组（簇），最大化组内相似性、最小化组间相似性。无需预先标注类别，属于无监督学习

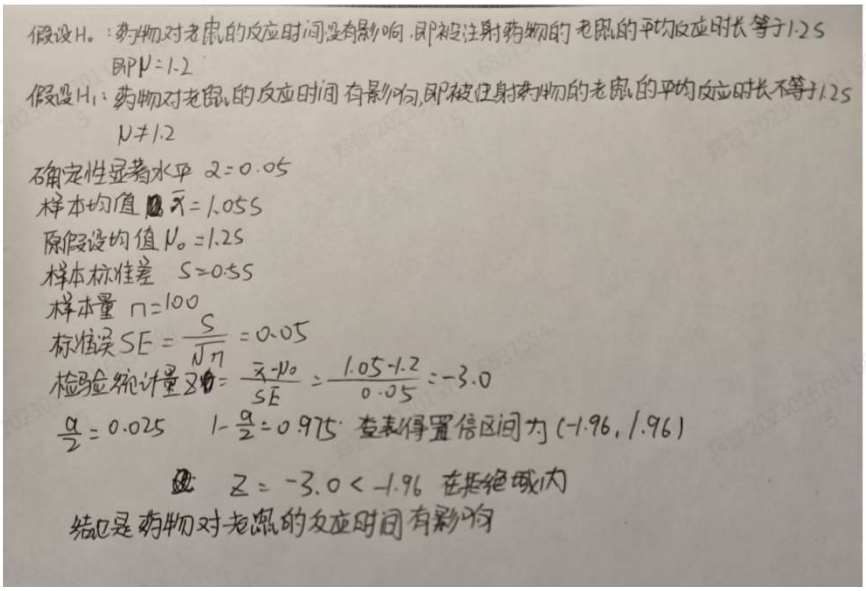
示例：电商对用户聚类，根据“购买频率”“消费金额”分为“高价值客户”“潜在流失客户”等，针对不同群体制定营销策略

8

10.假设检验 (5分)

一位神经学家正在通过给100只老鼠注射一种药物来测试这种药物对老鼠反应时间的影响。他知道没有被注射药物的老鼠的平均反应时间为1.2s，这100只被注射老鼠的平均反应时间为1.05s，标准差为0.5s。请通过假设检验来推断这种药物对老鼠的反应时间是否有影响。

我的答案



9、**贝叶斯网络**是基于**概率推理**的数学模型

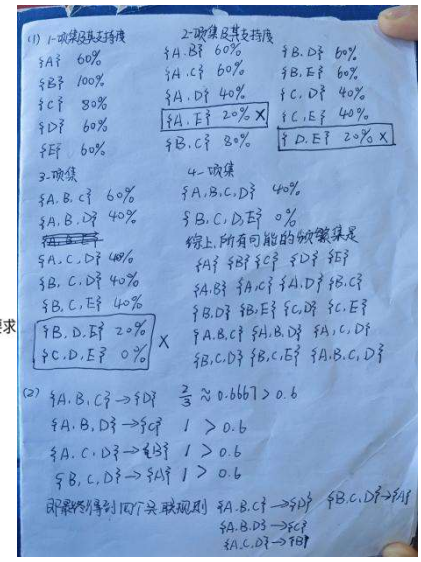
10、

见下面的交易记录，假设“最小支持度阈值”为0.4，“最小置信度阈值”是0.6。要求

（1）按照Apriori算法生成所有合法的频繁集。

（2）只从4-项频繁集中生成右侧只包含一个商品的所有关联规则。

序号	商品
1	A，B，C，D
2	B，C，E
3	A，B，C，E
4	B，D，E
5	A，B，C，D



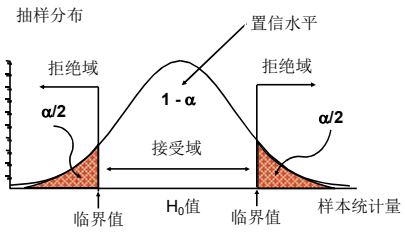
* 关于标准正态分布的查表问题：

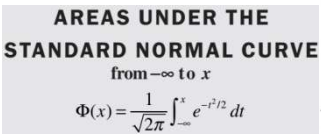
见下面的图和表。钟形曲线下的总面积为 1，α 称为显著性水平（指达到这个水平就可以拒绝），所以 1-α 就称为置信水平。由于图形对称，所以两端阴影面积各占 α/2。表的作用是根据 Φ(x) 的值来查 x 的值，Φ(x) 就是曲线下面，横坐标从−∞到 x 进行积分所得到的面积（见右边那个图）。

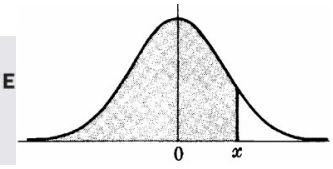
例如：要通过查表求得在显著性水平 α=0.03 下的置信区间

首先，因 α/2=0.015，故 1-α/2=0.985（即 Φ(x)=0.985），现在就需要通过查表来求 x。

其次，在表中查与 0.985 最接近的数值，发现这个值在第 2.1 行、第 7 列上，将行、列值串起来得 2.17，此即是 Φ(x)=0.985 的 x 值，即需要求得的临界值，根据图形的对称性，中心点是 0，故置信区间是（-2.17, 2.17）。







x	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5754
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7258	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7996	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

*假设检验问题：用一个例子来说明。

某食品公司生产一种罐头，按标准每罐净重为 227 克，根据以往生产经验罐头重量的标准差为 3 克。现随机抽查该公司产品 100 罐（即 n），测得平均净重为 228 克，判断这批罐头是否符合标准？（设显著性水平 α =0.05 ）

解：

根据题意：μ0=227， σ =3，n=100，$\bar{x}$=228， α =0.05

现在要检验的是参数 μ 是否与 μ0 一致（即在显著性水平 α =0.05 下可接受）。

（1）提出：原假设 H0：μ=μ0，备择假设 H1：μ≠μ0

（2）选取和计算检验统计量。可知$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$=3.33，z 服从标准正态分布 N(0, 1)。

检验统计量的基本公式为： $$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

1. **单个正态变量标准化**

   若原始变量 $X \sim N(\mu, \sigma^2)$，直接标准化为：

   $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

   （分母仅需总体标准差 σ）

2. **样本均值标准化**

   由于样本均值 $\overline{X}$ 的分布为 $N(\mu, \frac{\sigma^2}{n})$，其标准差为 $\frac{\sigma}{\sqrt{n}}$，因此标准化形式为：

   $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

   （分母需调整样本量影响）

（3）根据显著性水平确定置信区间。 由 α =0.05，得 α /2=0.025，1- α /2=0.975，查上面的正态分布表得置信区间（即接受域）为（-1.96，1.96）。

（4）进行比较并做出决策。 上面计算出的 z=3.33>1.96，在拒绝域内（即在接受域外），故拒绝 H0，结论是这批罐头不符合标准。

* 关联规则的 Apriori 算法：用一个例子来说明。

首先，见下面的交易记录，假设“最小支持度阈值”为 0.5，“最小置信度阈值”是 0.8。要求：

（1）按照 Apriori 算法生成所有合法的频繁集。

（2）只从 3-项频繁集中生成右侧只包含一个商品的所有关联规则。

序号	商品
1	A，C，D
2	B，C，E
3	A，B，C，E
4	B，E

解答：

（1）第一步，生成 1-项集及其支持度，如图所示。其中：支持度= 包含某商品的交易数/总交易数。其中的{D}因支持度小于“最小支持度阈值”，不属于频繁集，也不参与后续创建更大的频繁集。

候选项集 C1	支持度
{A}	50%
{B}	75%
{C}	75%
{D}	25%
{E}	75%

第二步，生成 2-项集及其支持度。

【生成方法：若有两个 k-1 项集，每个项集按字母顺序进行排序。如果两个 k-1 项集的前 k-2 个项相同，而最后一个项不同，则说明它们是可连接的，即可连接生成 k 项集。

例如有两个 3 项集： $\{A,B,C\}$ 和 $\{A,B,D\}$，这两个 3 项集就是可连接的，它们可以连接生成 4 项集 $\{A,B,C,D\}$ 。

又如两个 3 项集 $\{A,B,C\}$ 和 $\{A,D,E\}$ ，这两个 3 项集是不能连接生成 4 项集的。】

得到下面的 2-项集。注意，{A，B}的支持度=同时购买了 A 和 B 的交易数/交易总数。

因{A，B}和{A，E}的支持度小于“最小支持度阈值”，不属于频繁集，也不参与后续创建更大的频繁集。

候选项集 C2	支持度
{A，B}	25%
{A，C}	50%
{A，E}	25%
{B，C}	50%
{B，E}	75%
{C，E}	50%

第三步，生成 3-项集及其支持度

得到下面的 3-项集。

只有一行，不能生成更大的频繁集了。

候选项集 C3	支持度
{B，C，E}	50%

第四步，生成关联规则：

综上，所有可能的频繁集是：{A}，{B}，{C}，{E}，{A，C}，{B，C}，{B，E}，{C，E}，{B，C，E}

（2）根据题意，只从 3-项频繁集{B，C，E}生成右侧只包含一个商品的关联规则，结果如下所示：

{B，C}→{E}　置信度（{B，C}→{E}）= 2/2=1 > “最小置信度阈值”，故该关联规则成立。

{B，E}→{C}　置信度（{B，E}→{C}）= 2/3=0.67 < “最小置信度阈值”，故抛弃该关联规则。

{C，E}→{B}　置信度（{C，E}→{B}）= 2/2=1 > “最小置信度阈值”，故该关联规则成立。

置信度的计算方法为：同时包含左、右侧商品的交易数/只包含左侧商品的交易数。

最终得到两个关联规则：{B，C}→{E}和{C，E}→{B}。

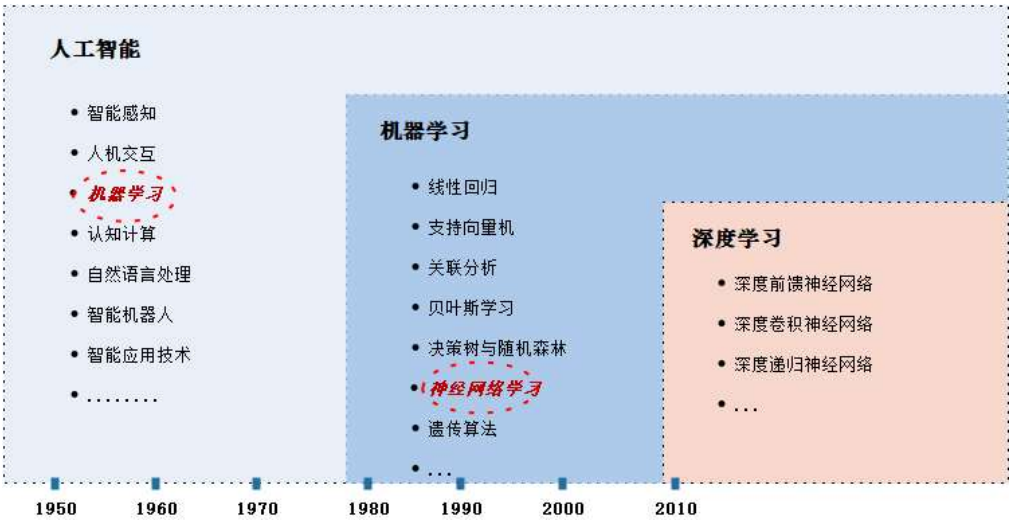
第三章

1、判断一个系统是否具有智能的标准是，看其是否能通过图灵测试

2、炭智能一般是指有碳基生命体，尤其是人类所展现的只能。

硅智能是指基于硅的计算设备，例如计算机和相关系统，所实现的智能，这主要是通过人工智能算法和技术实现的

3、



4、机器学习，是指从有限的观测数据中学习出具有一般性的规律，并利用这些规律对未知数据进行预测的方法

5、有监督学习，使用已知模式预测数据，其使用前提是训练集为带标签数据（Labeled data）常见的有监督学习算法有最近邻，朴素贝叶斯，决策树，随机森林，线性回归，支持向量机（Vector Machines，SVM）和神经网络等

无监督学习：从数据中发现未知的模式信息，当训练集中是不带标签的信息时，通常从用无监督学习法。常见的无监督学习算法有 k-means 聚类，主成分分析，关联规则分析等

监督学习：当训练集中的部分样本缺少标签信息时，通常采用半监督学习。常见的半监督学习算法有：半监督分类方法、半监督回归方法、半监督聚类方法和半监督降维方法。

6、机器学习的**目标函数**由两部分组成，即**误差函数**和**正则化项，**加入正则化项的目的是**防止过拟合**

（5）模型的精度（Precision）

- 在所有判别为正例的结果中，模型正确预测的样例所占的比例，即：

$$Precision=\frac{TP}{(TP+FP)}$$

（6）模型的召回率（Recall）

- 在所有正例中，模型正确预测的样本所占的比例，即。

$$\text{Recall}=\frac{TP}{(TP+FN)}$$

7、机器学习的数据集一般分为：训练集，测试集，验证集

8、（1）聚类（Clustering）

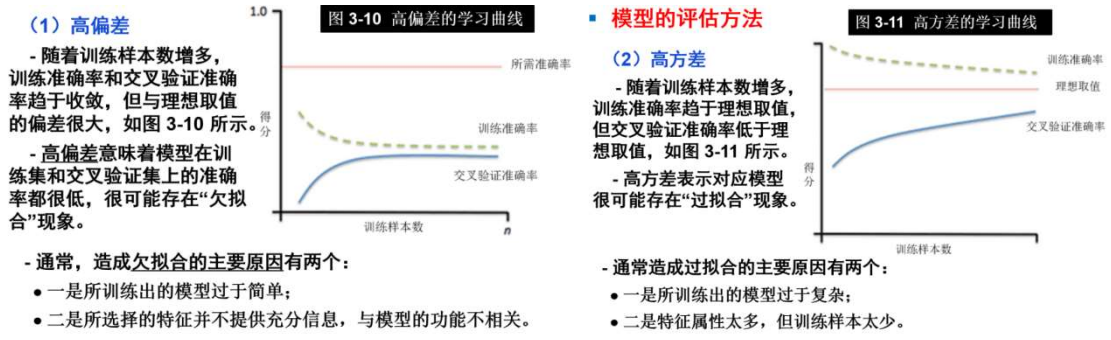
属于一种无监督学习算法，所涉及的属性为连续型属性（Continuous Attribute）。

常见的聚类算法有：k-means 聚类、高斯混合聚类（Gaussian Mixture Model，GMM)、学习向量量化（Learning Vector Quantization，LVQ）和聚集嵌套（Agglomerative Nesting，AGNES）算法等。

（2）分类（Classification）

属于一种有监督学习算法，所涉及的属性为分类型属性（Categorical Attribute）。

常见的分类算法有 ：K-最近邻（K-Nearest Neighbor，KNN）、逻辑回归、朴素贝叶斯、支持向量机（Support Vector Machine，SVM）、决策树与随机森林等算法。



9、 （1）TP（True Positive)

模型“正确地（真/True）”预测了样本的类别，为“正例”。

（2）FN（False Negative）

模型“错误地（假/False）”预测了样本的类别为“负例”，即模型犯了类似于统计学上的第一类错误（Type I Error）。

（3）FP（False Positive）

模型“错误地（假/False）”预测了样本的类别为“正例”，即模型犯了类似于统计学上的第二类错误（Type II Error）。

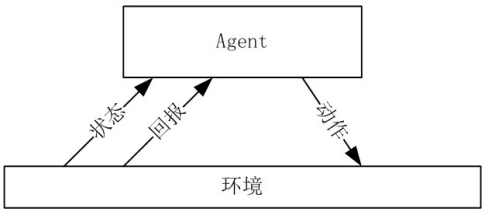
（4）TN（True Negative）

模型“正确地（真/True）”预测了样本的类别为“负例”。

10、贝叶斯分类也是通过概率推断的方法

11、Roc 曲线越靠近左上角效果越好

12、强化学习的原理



13、KNN 分类

下表显示出若干部电影，以及根据打斗和接吻次数给出的分类。现有一部新电影，其中的打斗和接吻次数分别是101和20。设k=3，请严格按照KNN算法的步骤计算并判断该部电影属于哪种类型？

为便于计算，请可采用曼哈顿距离（而不是欧式距离）来衡量两个点之间的距离，两个点（x1, y1）和（x2, y2）之间的曼哈顿距离为：|x1-x2|+|y1-y2|。

电影名称	打斗镜头	接吻镜头	电影类型
乱世家人	1	101	爱情片
魂断蓝桥	5	89	爱情片
英雄本色	108	5	动作片
拳霸	115	8	动作片

新电影坐标（X=101，Y=20）

电影名称	打斗 Xi	接吻 Yi	距离	类型
乱世佳人	1	101	131	爱情
魂断蓝桥	5	89	165	爱情
英雄本色	108	5	22	动作
拳霸	115	8	26	动作

从小到大排序选 K=3 个

动作

动作

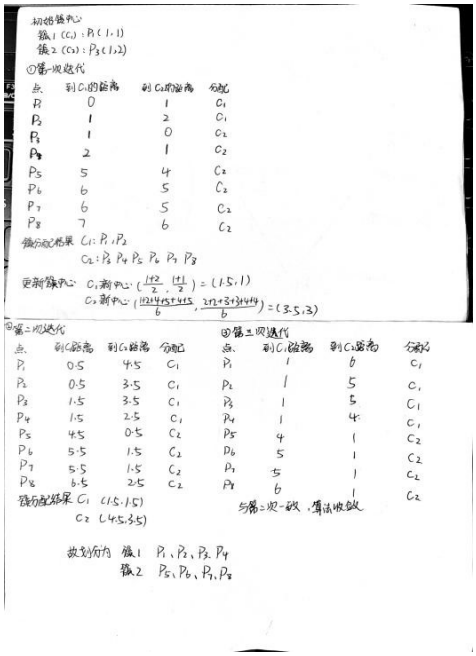
爱情

故是动作片

14、K-Means 聚类

试采用K-Means算法将下表中的8个坐标点分为2个簇。假设在初始时，选择P1和P3分别作为两个簇的中心。为便于计算，请采用曼哈顿距离（而不是欧式距离）来衡量两个点之间的距离，两个点（x1, y1）和（x2, y2）之间的曼哈顿距离为：|x1-x2|+|y1-y2|。

点	P1	P2	P3	P4	P5	P6	P7	P8
坐标	(1, 1)	(2, 1)	(1, 2)	(2, 2)	(4, 3)	(5, 3)	(4, 4)	(5, 4)



第四章

1、数据可视化工作应遵循的基本原则为终于原始数据，尊重目标用户，突出重点，强调用户体验

2、视觉编码涉及两个维度，图形元素和视觉通道

3、从可视化处理视角来看，可以将数据分为四个类型，定类，定序，定距，以及定比

4、通道表现力的评价指标有：精确性，可辨认性，可分离性，视觉突出性

5、 精度对比型

"下列视觉通道中，对数值差异感知最不敏感的是？"

（答案：颜色饱和度）

"在需要精确对比定量数据时，应避免使用哪种通道？"

（答案：面积）

任务适配型

"设计热力图时最核心的视觉通道是？"

（答案：颜色）

当需要展示部分占整体比例时，最优选择是？

（答案：面积/角度）

感知特性型

"人眼对以下哪种通道的微小变化最迟钝？"

（答案：颜色明度）

"在散点图中起核心作用的通道是？"

（答案：位置）

多通道组合

"树状图同时利用了哪两种通道？"

（答案：面积+位置）

"堆叠条形图的核心通道组合是？"

（答案：长度+颜色）

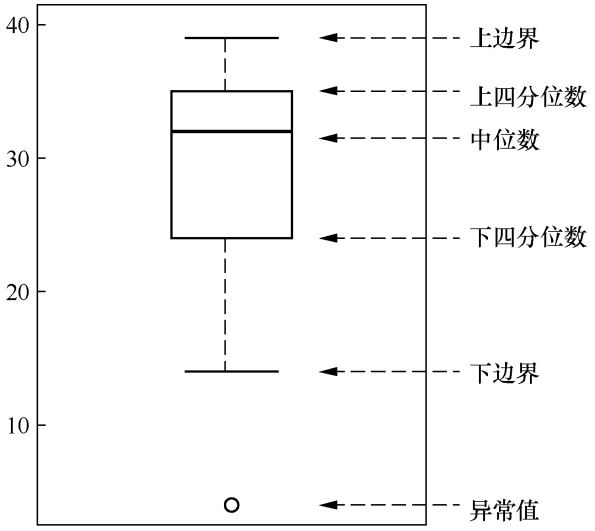
特殊场景 "色盲用户可视化应慎用哪种通道？"

（答案：颜色）
"在小尺寸图表中最易失效的通道是？"
（答案：面积）
6、“两种视觉通道之间不能互相影响对方的表现力”，这表明的是视觉通道表现力评价中的可分离性
7、完图法则人类的视觉感知活动往往倾向于将被感知对象当作一个整体去认知，并理解为与自己经验相关的、简单的、相连的、对称的或有序的以及基于直觉的完整结构。因此，视觉感知结果往往不等同于感知对象的各部分的独立感知结果之和。

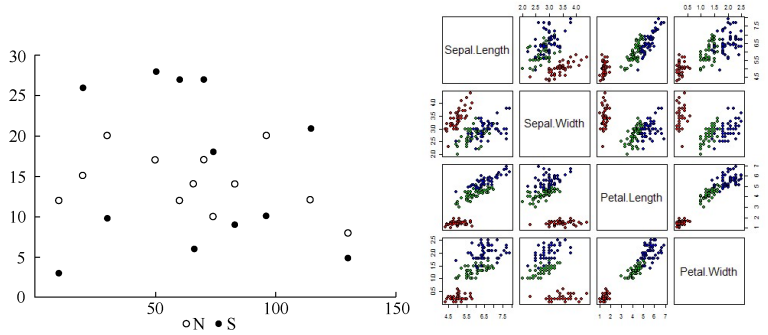
8、视觉假象
是指给目标用户产生的错误或不准确的视觉感知，而这种感知与数据可视化者的意图或数据本身的真实情况不一致。原因：
（1）**可视化视图所处的上下文（周边环境）可能导致视觉假象。**
（2）人们对亮度和颜色的相对判断容易造成视觉假象。
（3）目标用户的经历与经验可能导致视觉假象。
9、可视分析学的特点有：强调数据到知识的转化过程，强调可视化分析与自动化建模之间的相互作用，强调数据映射和数据挖掘的重要性，强调人机交互的重要性，强调数据加工活动的必要性

10、饼图主要用于表示整体与部分之间的关系

箱线图是由约翰·图基（John W. Tukey）发明的一种用于可视化数据分布的制图方法，如图 4-13 所示（1）箱（长方形盒子）。表示数据的大致范围，一般为数据取值范围的 25%～75%。需要注意的是，数据的实际取值范围用盒子上方和下方的两根横线表示。（2）线（盒子中的横线）。表示中位数的位置。



散点图主要用于显示数据点在笛卡儿坐标系中的分布情况，每个点所对应的横、纵坐标代表的是该数据在对应维度上的属性值，如图 4-14 所示。在实际应用中，我们经常采用散点图矩阵的方式表示多维



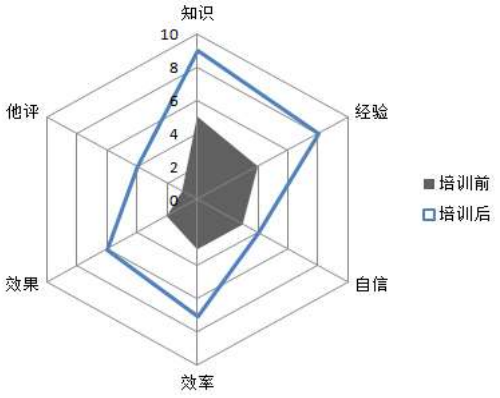
维恩图一种表示数据集合运算的可视化方法—用平面上的封闭图形元素之间的重叠关系表示数据集合的并与交等集合运

热地图一般以地图为基础，采用不同色彩（如颜色、亮度、透明度等）表示数据值的大小。常用于显示数据的分布和变化趋势。近年来，基于位置的服务（Location Based Services，LBS）系统的兴起推动了包括热点图在内的、基于地图的数据可视化方法的广泛应用。

等值线主要用于显示等值数据的分布情况， 其画法为将多维空间中的具有相同值的数据点相互连接后投影到二维平面上，一般为三无（无相交、无分支、无中断）封闭线路。

等值线在地理（如等高线等）、气象（如等温线、等压线、等降水量线等）、物理（如等磁线、等势线等）等领域具有较为广泛的应用。

雷达图主要用于可视化多个属性的数据



11、视觉隐喻可以在实际图像上进行，也可以在抽象化后的图像上进行

12、若要对因果关系进行隐喻，最适合采用鱼刺图

13、目前的地铁的线路图图最早起源于伦敦地铁线路图

第五章

1、数据质量的基本属性有正确性，完整性，和一致性，扩展属性有形式化程度，时效性，精确性，子描述性

正确性（Correctness）：数据是否实事求是地记录了客观现象。

完整性（Integrity）：数据是否未被未授权篡改或损坏。

一致性（Consistency）：数据内容之间是否存在自相矛盾的现象。

形式化程度（Formalization）：数据的形式化表示程度。程度越高，越易于被计算机自动处

理。
时效性（Timeliness）：数据是否被及时记录下来，确保数据与客观世界之间的同步性。
精确性（Accuracy）：数据的精度是否满足后续处理的要求。
自描述性（Self-Description）：数据是否带有自描述信息，如数据模式信息，有效性验证方法。

2、数字签名消息鉴别保护双方的数据交换不被第三方侵犯，但不保证双方互相欺骗或抵赖，故需要数字签名来实现实体鉴别的功能。

3、第一数字定律（First-Digit Law，也称本福特定律）描述的是自然数“1”到“9”的使用频率，设 $d \in \{1,2,3,4,5,6,7,8,9\}$，其中,数字“1”的使用最多接近三分之一，“2”为17.6%，“3”为12.5%，依次递减，“9”的频率是 4.6% 该定律可用来检查各种数据是否有造假的可能。

小概率原理基本思想：一个事件如果发生的概率很小的话，那么它在一次试验中是几乎不可能发生的，但在多次重复试验中几乎是必然发生的，数学上称之小概率原理。

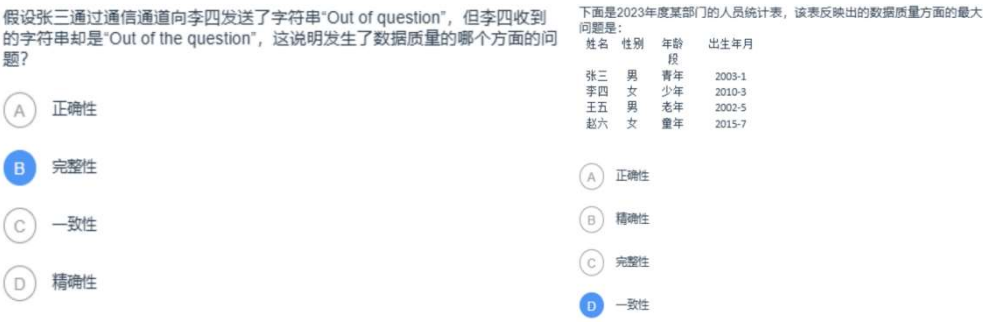
语言学规律：各个字母的使用次数不一样，有的偏高，有的偏低，这种现象称为偏用现象。连接特征/重复特征

4、对于具有耐抗性的分析结果，当数据的一小部分被新的数据代替时，即使它们与原来的数值差别很大，分析结果也只会有轻微的改变

5、表示数据分布峰态的术语是 Kurtosis,而表示数据分布对称性的术语是 Skewness / skewness

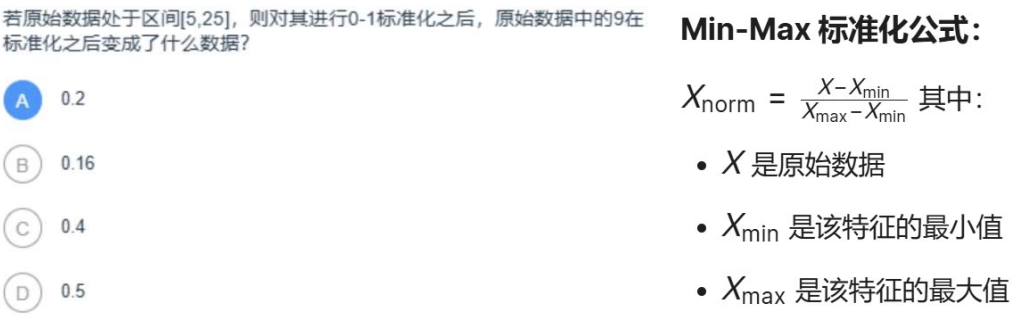
6、脏数据体现在缺失数据，异常数据，冗余数据，错误数据

7、



8、非随机缺失的解决方法较为复杂，可以采用模型选择法和模式混合法等

9、.



10、Z-score 标准化：* Z-score 标准化（Zero-Score Normalization）指使经过处理的数据符合标准正态分布，即均值为 0，标准差为 1。

若两个属性A和B之间的相关系数为rAB，则在下面哪种情况下，可以判断A、B两个属性是冗余的，即其中一个属性可以删除？

A rAB = 1

B rAB = 0

C rAB = 0.5

D rAB = -0.5

11、

12、离群点：模型预测的 y 值与真实的 y 值相差非常大。

高杠杆点：x 值比较异常，通常与因变量值 y 没有关系。

强影响点：对模型有较大影响的点。

13、

* 分箱处理

- 分箱（Binning）处理的基本思路是将数据集放入若干个“箱子”之后，用每个箱子的均值（或边界值）替换该箱内部的每个数据成员，进而达到噪声处理的目的。

- 下面以数据集 Score={60, 65, 67, 72, 76, 77, 84, 87, 90}的噪声处理为例，介绍分箱处理（采用均值平滑技术等的深分箱方法）的基本步骤：

• 第 1 步：将原始数据集 Score={60, 65, 67, 72, 76, 77, 84, 87, 90}放入以下 3 个箱：

箱 1：60, 65, 67

箱 2：72, 76, 77

箱 3：84, 87, 90

▪ 噪声数据及其处理方法

* 分箱处理（续）

• 第 2 步：计算每个箱的均值：

箱 1 的均值：64

箱 2 的均值：75

箱 3 的均值：87

• 第 3 步：用每个箱的均值替换对应箱内的所有数据成员，进而达到数据平滑（去噪声）的目的：

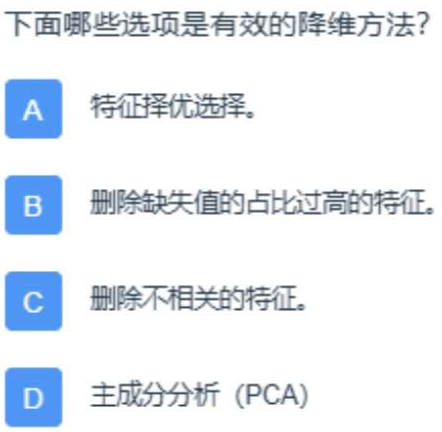
箱 1：64, 64, 64

箱 2：75, 75, 75

箱 3：87, 87, 87

• 第 4 步：合并各箱，得到数据集 Score 经过噪声处理后的新数据集 score*，即 score* = {64, 64, 64, 75, 75, 75, 87, 87, 87}。

14、过滤法、包裹法、嵌入法、主成分分析



14、数据脱敏处理的要求：单向性、无残留、易于实现

15、按标注活动的自动化程度，数据标注可以分为手工标注、自动化标注和半自动化标注。

16、所谓“规整数据”应同时满足以下 3 个基本原则，如图 5-9 所示：

（1）每个观察占且仅占一行。

（2）每个变量占且仅占一列。

（3）每一类观察单元构成一个关系（表）。

17、

例题：假设有一组数据：31, 15, 14, 16, 15, 17, 19, 18，要求通过箱线图的方法识别该数据集中的离群点。

解答：

首先将数据从小到大排序：{14, 15, 15, 16, 17, 18, 19, 31} ；最大值为 31，最小值为 14，数据个数 N=8。

（下四分位数）Q1 的位置 = N * 0.25 = 8 * 0.25 = 2

（中位数）Q2 的位置 = N * 0.5 = 8 * 0.5 = 4

（上四分位数）Q3 的位置 = N * 0.75 = 8 * 0.75 = 6

对应 Q1（第 2 个）位置的数为 15，对应 Q2（第 4 个）位置的数为 16，对应 Q3（第 6 个）位置的数为 18。简言之：Q1 = 15，Q3 = 18。

IRQ = Q3 - Q1 = 18 - 15 = 3

下边缘 = Q1 - 1.5 * IQR = 15 - 1.5 * 3 = 10.5

上边缘 = Q3 + 1.5 * IQR = 18 + 1.5 * 3 = 22.5

上下边缘的区间为：[10.5, 22.5]，在数据{14, 15, 15, 16, 17, 18, 19, 31 }中，只有 31 在该区间之外，可判定为离群点。

还应能绘制相应的箱线图（箱线图画法见第 5 章课件）。

- 1.5 是一个经验值，基于正态分布的性质：
  - 在正态分布中，大约 99.3%的数据落在均值±2.7 个标准差范围内。

- IQR 大约覆盖了正态分布中 50%的数据，IQR ≈ 1.35 个标准差（对于标准正态分布）。
- 1.5 * IQR ≈ 2.7 个标准差，因此上下边缘大约对应均值±2.7 个标准差。
- 这样定义的离群点大约对应正态分布中 0.7%的极端值（即每侧约 0.35%）。

- 1.5 的选择可以捕捉到大多数真实的离群点，同时避免将太多正常数据标记为离群点。
- 如果需要更严格或更宽松的标准，可以调整乘数（例如用 3.0 定义“极端离群点”）。

第5章 数据加工

- **噪声数据及其处理方法**

*** 离群点处理（续）**

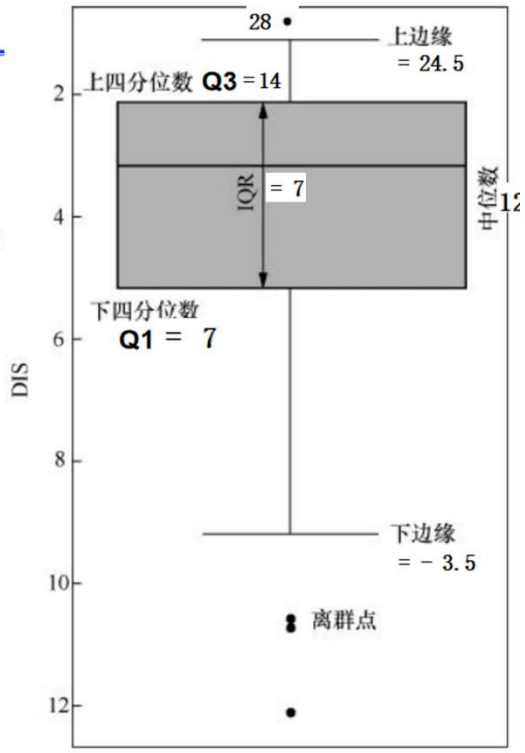
- 常用离群点的识别方法有 4 种：

（1）可视化方法

• 例如图 5-4 所示的绘制散点图的方法；

• 也可以采用箱线图方法，如图 5-5 所示，其中没有包含在箱线中的 3 个独立的点是离群点。

图5-5 箱线图与离群点

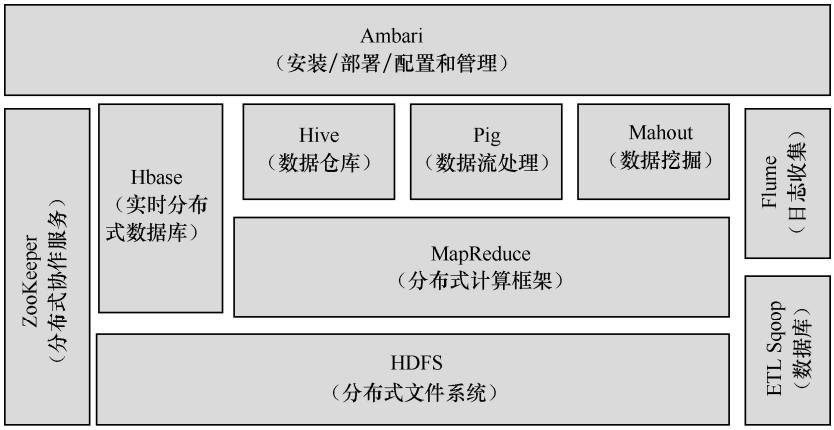


第六七章

1、

类型	含义	举例
HaaS（Hardware as a Service，硬件即服务）	云端将硬件设备以服务形式提供给终端，终端可按需购买或租用云端的硬件设备，安装自己的软件系统，完成各种数据存储或计算任务。	IBM的IDC云计算
IaaS（Infrastructure as a Service，基础设施即服务）	云端将计算资源和存储资源以服务形式提供给终端，终端可按需购买或租用所需的基础设施。	Amazon的EC2
SaaS（Software as a Service软件即服务）	云端将软件系统以服务形式提供给终端，终端可按需购买或租用云端的软件系统，完成各种计算任务。	Salesforce.org
PaaS（Platform as a Service，平台即服务）	云端将软件开发平台以服务形式提供给终端，终端可按需购买或租用云端的开发平台，完成软件系统的研发任务。	Google的App Engine
DaaS（DataBase as a Service，数据库即服务）	云端将数据库及其管理系统以服务形式提供给终端，终端可按需购买或租用云端的数据库或数据库管理系统服务。	Oracle 的云服务

2、Hadoop 生态系统



2、MapReduce 编程模型将问题**抽象为两个阶段—Map 阶段和 Reduce 阶段**；其输入和输出值均为<key, value>型，即“键-值对”

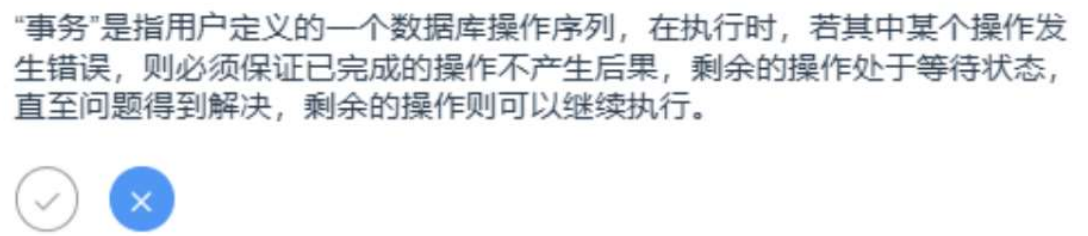
3、早期的 HDFS 是按照 Google 文件系统（Google File System，GFS）的思想设计的，因此**HDFS 通常被认为是 GFS 的开源版本**

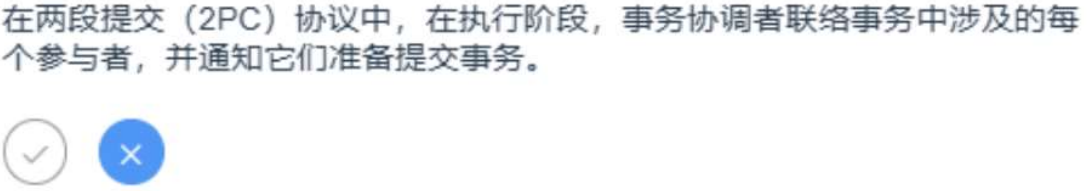
4、与 Hadoop MapReduce 的磁盘计算不同的是，Spark 采用的是**内存**计算模式

5、RDD 是 Spark 的抽象数据模型；

6、Spark 的技术架构可以分为 3 个层：资源管理层、Spark 核心层和服务层。

7、事务是数据库管理系统运行的基本工作单位，也是用户定义的一个数据库操作序列，这些操作要么全部执行，要么全部不执行，是一种不可分割的工作单位





8、NoSQL 指那些非关系型的、分布式的、不保证遵循 ACID 特征的数据存储系统

9、 NoSQL 数据库中采用的主要数据模型有 4 种：key-value、key-document、key-column 和图存储

10、ACID（原子性、一致性、隔离性、持久性）

11、CAP 理论的基本思想如下：一个分布式系统不能同时满足一致性（Consistency）、可用性（Availability）和分区容错性（Partition Tolerance）

12、BASE 原则是基本可用（Basically Available)、柔性状态（Soft State)和最终一致（Eventually Consistent)的缩写。

13、在 NoSQL 中，分片（Sharding）与复制（Replication）是数据分布的两种技术

14、该论文将数据分析的方法、技术和工具—分析工具的应用时代分为 3 个，即商务智能

时代、大数据时代和数据富足供给时代

15、从复杂度及价值高低两个维度，可以将数据分析分为

描述性分析（Descriptive Analytic）、

诊断性分析（Diagnostic Analytics）

预测性分析（Predictive Analytics）

规范性分析（Prescriptive Analytics）

16、

（第七章）下面哪个选项是定义下面表述的最恰当的概念？
“将数据转换为产品的艺术。”

A 数据清洗

B 数据挖掘

C 数据加工

D 数据柔术

19.单选题

现要找出乌鲁木齐每年的最高气温。假设map()函数已经从输入文本中提取出年份和气温信息，得到如下所示的结果。那么，对该结果进行排序和分组、以及经过reduce()函数处理后，得到的结果分别是什么？
(2023, 24)
(2023, 17)
(2023, -12)
(2022, 27)
(2022, 7)

A 排序和分组后：
[(2022, 27) (2022, 7)]
[(2023, 24) (2023, 17) (2023, -12)]
经过reduce()函数处理后：
(2022, 27)
(2023, 24)

B 排序和分组后：
(2022, [27, 7])
(2023, [24, 17, −12])
经过reduce()函数处理后：
(2022, 27)
(2023, 24)

C 排序和分组后：
(2022, [27, 7])
(2023, [24, 17, −12])
经过reduce()函数处理后：
[(2022, 27) (2023, 24)]

D 排序和分组后：
(2022, [27, 7])
(2023, [24, 17, −12])
经过reduce()函数处理后：
(2022, 27, 2023, 24)

17、Mahout 的主要目标是提供可扩展的机器学习算法及其实现，旨在帮助开发人员更加方

便快捷地创建智能应用程序。