商业分析总复习指南

测验题

说明：请用2-3句话简要回答以下每个问题。

1. 什么是情感分析？请列举至少两个商业应用。

2. 请解释中心极限定理（CLT）及其在统计推断中的重要性。

3. 什么是跨行业数据挖掘标准流程（CRISP-DM）？请列出其六个阶段。

4. 请解释相关性与因果关系的区别，并根据课程资料给出一个例子。

5. 决策树是如何对一个新样本进行分类的？

6. 在模型构建中，什么是“过度拟合”？请描述一种避免过度拟合的方法。

7. 简要比较线性回归和逻辑回归在目标变量和输出上的主要区别。

8. K-最近邻（KNN）分类算法的基本思想是什么？

9. 在关联规则学习中，支持度（Support）、置信度（Confidence）和提升度（Lift）这三个指标分别衡量什么？

10. 根据课程内容，商业分析有哪三个维度？请分别说明它们回答了什么问题。

答案

1. 情感分析是一项自然语言处理任务，旨在通过分析文档来获取作者表达的正面或负面感受、观点或情绪。其商业应用包括品牌分析、新产品认知、产品基准比较，以及通过分析公众情绪来预测选举结果或市场趋势。

2. 中心极限定理（CLT）指出，对于从总体中抽取的大样本，其均值的抽样分布会近似服从正态分布。这一定理非常重要，因为它是Z检验和t检验等假设检验的基础，允许我们仅基于一个样本的统计量来对总体进行推断。

3. CRISP-DM是数据挖掘项目的标准流程，它提供了一个结构化的方法论。其六个阶段分别是：业务理解（Business Understanding）、数据理解（Data Understanding）、数据准备（Data Preparation）、建模（Modelling）、评估（Evaluation）和部署（Deployment）。

4. 相关性衡量两个变量之间线性关联的强度和方向，但并不意味着一个变量的变化导致了另一个变量的变化，即“相关关系不等于因果关系”。例如，一个地区的银行分支机构数量和总存款额可能呈正相关，但这并不意味着开设更多的分支机构就会直接导致更多的存款。

5. 决策树通过一系列基于属性的测试来对新样本进行分类。样本从根节点开始，根据其在每个节点上的属性值选择相应的分支向下移动，直到到达一个叶节点。该叶节点所代表的类别（或类别概率分布）即为该新样本的预测结果。

6. 过度拟合是指模型在训练数据上表现非常好，但在未见过的新数据上表现不佳的现象，因为它学习了训练集中的细节和噪声而非普遍规律。避免过度拟合的一种方法是剪枝（Pruning），通过简化决策树（如限制树的深度或在生成完整树后移除节点）来防止其变得过于复杂。

7. 线性回归和逻辑回归的主要区别在于目标变量的类型和模型的输出。线性回归用于预测连续的数值型变量（如房价、销售额），其输出是一个连续值。而逻辑回归是一种分类模型，用于预测类别型变量（如“是/否”、“通过/未通过”），其输出是属于某一特定类别的概率。

8. K-最近邻(KNN)算法的基本思想是“物以类聚”。对于一个未知类别的新样本，算法会计算它与训练集中所有样本的距离，找出距离最近的K个邻居，然后根据这K个邻居的类别，通过多数投票等方式来决定新样本的类别。

9. 支持度衡量某个项集(如{啤酒，尿布})在所有交易中出现的频率，反映其受欢迎程度。置信度衡量规则的可靠性，例如在购买了A的交易中，有多大比例也购买了B。提升度衡量规则的有趣性，即购买A对购买B的概率的提升程度，如果大于1则说明规则有意义。

10. 商业分析的三个维度是:描述性分析、预测性分析和指导性分析。描述性分析回答“发生了什么？”的问题;预测性分析回答“将要发生什么以及为什么？”的问题;指导性分析则回答“我应该做什么以及为什么？”的问题。

论述题

说明：以下问题旨在考察综合理解与应用能力。

1. 结合Robinhood应对GameStop事件的商业案例，详细阐述如何设计一个利用推特数据的情感分析项目。请描述该项目在CRISP-DM流程的“业务理解”和“数据理解”阶段需要完成的关键任务，并解释自然语言处理（NLP）在其中的作用。

2. 假设您是一家电信公司的业务分析师，目标是预测哪些客户可能会流失。请比较并论述使用决策树和逻辑回归这两种模型来解决此问题的优劣。您的论述应包括模型的可解释性、对数据关系的处理能力以及各自可能遇到的挑战（如过拟合）。

3. 解释为什么在评估分类模型时，仅使用“准确率”指标可能具有误导性，尤其是在处理类别分布不均衡的数据集时。请详细描述精准率（Precision）、召回率（Recall）以及ROC曲线如何为模型性能提供更全面、更稳健的评估。

4. 比较有监督学习和无监督学习的根本区别。请分别以线性回归和K-均值聚类为例，说明它们的目标、应用场景、数据要求以及评估方法有何不同。

5. “购物篮分析”是关联规则挖掘的经典应用。请阐述如何利用支持度、置信度和提升度这三个核心指标来发现“有趣”且有商业价值的规则。并讨论该技术除了零售业的交叉销售外，还能在哪些其他商业场景中得到应用。

关键术语词汇表

术语 (Term)	定义 (Definition)
K-Means Clustering (K均值聚类)	一种常用的聚类方法，将数据点分配到预先指定数量（k）的簇中，每个簇由一个质心代表，目标是最小化所有数据点到其所属簇质心的平方距离之和。
K-Nearest Neighbor (KNN) (K最近邻)	一种分类算法，通过查找训练集中与新样本距离最近的K个邻居，并根据这些邻居的类别进行投票，来确定新样本的类别。
Accuracy (准确率)	分类模型评估指标，指预测正确的样本数占总样本数的比例。
Alternative Hypothesis (H1) (备择假设)	假设检验中与原假设相对立的假设，通常是我们希望通过数据证明其成立的论点。
Association Rule Mining (关联规则挖掘)	一种流行的无监督学习技术，用于在大型数据集中发现变量之间有趣的关系或关联模式，常被称为“购物篮分析”。
Business Analytics (商业分析)	分析数据，提取有助于业务决策的有用信息的过程。
Central Limit Theorem (CLT) (中心极限定理)	指出对于从总体中抽取的大样本，其均值的抽样分布遵循近似正态分布。
Classification (分类)	一种监督学习任务，目标是预测一个离散的类别标签。
Clustering (聚类)	一种无监督学习技术，将数据对象分组，使得同一簇内的对象彼此相似，而与其他簇的对象不相似。

Coefficient of Determination (R^2) (判定系数**)**	回归模型评估指标，表示因变量的方差中可由自变量解释的部分有多大，取值范围为0到1，越接近1模型越好。
Confidence (置信度**)**	关联规则的度量指标，衡量在包含条件（前件）的交易中，同时也包含结果（后件）的比例。
Confidence Interval (CI) (置信区间**)**	一个区间估计，以一定的概率（置信水平）包含未知的总体参数。
Correlation (相关性**)**	衡量两个变量之间线性关系的相对强度和方向的指标，取值在-1和+1之间。
CRISP-DM (跨行业数据挖掘标准流程**)**	一个广泛应用的数据挖掘过程模型，包括业务理解、数据理解、数据准备、建模、评估和部署六个阶段。
Cross-Validation (交叉验证**)**	一种模型评估技术，将数据集分成k个子集，轮流使用其中k-1个作为训练集，剩下的1个作为测试集，以获得更稳健的性能估计。
Decision Tree (决策树**)**	一种流行的监督学习模型，它通过一系列基于属性的判断来做出决策或预测，结构易于理解和解释。
Entropy (熵**)**	源于信息论，用于衡量数据集的混乱或不纯度程度。在决策树中，用于计算信息增益。
Hypothesis Testing (假设检验**)**	一种统计方法，用于根据样本数据对关于总体的某个假设（原假设）做出是拒绝还是接受的决策。

Information Gain (信息增益)	在决策树构建中，指由于使用某个属性进行划分而导致的熵的减少量。算法会选择信息增益最大的属性进行划分。
Lift (提升度)	关联规则的度量指标，衡量规则的有趣性，即条件（前件）的出现对结果（后件）出现的概率的提升程度。大于1表示正相关。
Linear Regression (线性回归)	一种回归分析技术，用于建立一个或多个自变量与一个连续的因变量之间的线性关系模型。
Logistic Regression (逻辑回归)	一种分类模型，用于预测一个二元或多元的类别型变量，输出结果为属于某个类别的概率。
Mean (均值)	集中趋势的度量，指一组数据的算术平均值。
Median (中位数)	集中趋势的度量，指将一组数据按大小排序后处于中间位置的数值。
Mode (众数)	集中趋势的度量，指一组数据中出现频率最高的数值。
Natural Language Processing (NLP) (自然语言处理)	计算机科学、人工智能和语言学的交叉领域，关注计算机与人类语言之间的交互，目标是让计算机能够处理、分析和理解自然语言。
Normal Distribution (正态分布)	一种连续概率分布，由均值μ和标准差σ完全表征，其曲线呈钟形。
Null Hypothesis (H0) (原假设)	假设检验中当前持有的信念或被假定为真的陈述，检验的目的就是收集证据来决定是否推翻它。

Overfitting (过度拟合)	模型在训练数据上表现极好，但在新的、未见过的数据上表现不佳的现象。这是因为模型学习了训练数据中的噪声和细节。
P-value (P值)	在原假设为真的前提下，观察到当前检验统计量或更极端结果的概率。P值越小，反对原假设的证据越强。
Precision (精准率)	分类模型评估指标，在所有被模型预测为正类的样本中，实际为正类的比例。
Pruning (剪枝)	简化决策树以防止过度拟合的过程，可以在树生长时进行（预剪枝），也可以在树完全生成后进行（后剪枝）。
Recall (召回率)	分类模型评估指标，在所有实际为正类的样本中，被模型正确预测为正类的比例。
Regression (回归)	一种监督学习任务，目标是预测一个连续的数值。
ROC Curve (ROC曲线)	接收者操作特征曲线，通过绘制不同决策阈值下的真正例率（TPR）与假正例率（FPR）来评估分类模型性能的工具。
Sampling (抽样)	从总体中选择一个子集（样本）的过程，目的是通过分析样本来推断总体的特征。
Sentiment Analysis (情感分析)	一项自然语言处理任务，旨在通过分析文本获取作者表达的感受、观点、情绪等（如正面、负面或中性）。
Significance Level (α) (显著性水平)	在假设检验中预先设定的一个概率阈值，用于判断是否拒绝原假设。通常设为0.05。当p值小于α时，拒绝原假设。

Standard Deviation (标准差**)**	离散程度的度量，衡量数据点与其均值的偏离程度，是方差的平方根。
Support (支持度**)**	关联规则的度量指标，衡量某个项集在所有交易中出现的频率或比例。
Term Frequency (词频**)**	在文本分析中，指某个词在文档中出现的次数，用于衡量一个词在文档中的重要性。
Variance (方差**)**	离散程度的度量，衡量数据与其均值之差的平方的平均值。
Z-Test (Z检验**)**	一种统计检验，当总体方差已知且样本量较大时，用于检验样本均值与总体均值之间是否存在显著差异。