CDA LEVEL III 考试大纲

CERTIFIED DATA ANALYST LEVEL III EXAMINATION OUTLINE

一、总体目标

CDA（Certified Data Analyst），即“CDA 数据分析师”，是在数字经济大背景和人工智能时代趋势下，面向全行业的专业技能认证，旨在提升全球用户数字技能，助力企业数字化转型，推动行业数字化发展。「CDA 数智化人才考核标准」是面向全行业数据相关岗位的一套科学化、专业化、国际化的人才技能准则，CDA 考试大纲规定并明确了数智化认证考试的具体范围、内容和知识点，考生可按照大纲要求进行相关知识的学习，获取技能，成为数智化专业人才。

二、考试形式与试卷结构

考试方式：一年四届 （3、6、9、12 月的最后一个周六），线下统考，上机答题。

考试题型：客观选择题（单选 45 题，多选 15 题，材料 10 题，每题 1 分），案例实操题（1 题，30 分）

考试时间：90 分钟（客观选择题），120 分钟（案例实操题），共 210 分钟

考试成绩：分为 A、B、C、D 四个层次，A、B、C 为通过考试，D 为不通过

考试要求：客观选择题为闭卷上机答题，请勿携带与考试无关的物品。案例实操题考生须自行携带电脑操作（安装好带有数据挖掘功能的软件，如 PYTHON（推荐）、SQL、SPSS MODELER、R、SAS 等，电脑须具备 USB 拷贝功能及相关解压软件，进行案例操作分析。案例数据将统一提供 CSV 文件）。

三、知识要求

针对不同知识，掌握程度的要求分为【领会】、【熟知】、【应用】三个级别，考生应按照不同知识要求进行学习。

1. 领会：考生能够了解规定的知识点，并能够了解规定知识点的内涵与外延，了解其内容要点之间的区别与联系，并能做出正确的阐述、解释和说明。

2. 熟知：考生须掌握知识的要点，并能够正确理解和记忆相关理论方法，能够根据不同要求，做出逻辑严密的解释、说明和阐述。此部分为考试的重点部分。

3. 应用：考生须学会将知识点落地实践，并能够结合相关工具进行商业应用。能够根据具体要求，给出问题的具体实施流程和策略。

四、考试科目

◆ **PART 1 数据挖掘概论（占比 10%）**

1. 数据挖掘的思维和方法论
2. 算法建模的要素
3. 算法模型的分类
4. 数据挖掘模型的落地
5. 数据挖掘模型的评估（含过拟合、欠拟合、偏差与方差等）

◆ **PART 2 数据处理与特征工程（占比 15%）**

1. 数据探索与数据处理
2. 特征工程概要
3. 特征构造
4. 特征选择
5. 特征转换

◆ **PART 3 机器学习算法（I）（占比 15%）**

1. 正则化的回归
2. 最近邻法（KNN）
3. 朴素贝叶斯
4. 聚类算法
5. 关联规则
6. 序列模式

◆ **PART 4 机器学习算法（II）（占比 30%）**

1. 决策树（分类树及回归树）
2. 支持向量机
3. 集成方法
4. 异常识别算法
5. 神经网络
6. 时间序列算法

◆ **PART 5 优化分析算法（占比 5%）**

1. 运筹优化
2. 流程挖掘

◆ **PART 6 数据挖掘实战方法（占比 20%）**

1. 多分类与多输出
2. 工作流（Pipeline）
3. 类别不平衡问题
4. 标签缺失问题
5. 模型优化与调参
6. 模型解释

◆ **PART 7 数据挖掘模型管理（占比 5%）**

1. MLOps（机器学习研发运营一体化）
2. 模型生命周期管理

◆ **PART 8 数据挖掘实操（案例实操）**

根据题目和给定数据，基于有限的算力和时间，使用任何可能的数据挖掘技术与方法（包括且不限于本考纲展示的技术方法），训练模型并评估结果。

五、科目内容

PART 1 数据挖掘概论

◆ **1、数据挖掘的思维和方法论**

【领会】

企业中数据分析的层级

【熟知】

数据挖掘的标准方法论（CRISP-DM 及 SEMMA）

数据挖掘的知识发现

【应用】

根据给定的需求建立一个数据挖掘的项目

◆ **2、算法建模的要素**

【熟知】

目标函数及其表示

模型的学习策略

模型的寻优方法

模型的评估方法

◆ **3、算法模型的分类**

【熟知】

有监督学习

无监督学习

【应用】

根据需求场景的不同，选择适合的模型

◆ **4、数据挖掘模型的落地**

【熟知】

模型结果的解读

模型的投入产出

模型的生命周期

◆ **5、数据挖掘模型的评估（含过拟合、欠拟合、偏差与方差等）**

【应用】

回归问题的模型评估方法

分类问题的模型评估方法

无监督学习的评估方法

可视化的模型评估（含学习曲线等）

模型预测的偏差与方差（含两者间的关联）

过拟合与欠拟合，及其常用解决方法（含正则，剪枝，早停等）

PART 2 数据处理与特征工程

◆ **1、数据探索与数据处理**

【领会】

数据处理的重要意义

【熟知】

数据探索的常用方法（统计指标探索，可视化探索等）

错误值的识别与处理

离群值的识别与处理

缺失值的识别与处理（含机器学习模型对缺失值的处理方法）

分类型特征的处理（含线性编码、独热编码等方法）

数据的标准化、归一化

【应用】

根据数据与建模的情况，决定是否对数据执行处理，选择合适的数据处理方法

◆ 2、特征工程概要

【领会】

特征工程的重要性特征理解

特征改进（数据清洗对特征的影响）

【熟知】

特征工程的涵盖范围

特征选择的目的

特征构造的目的

特征转换的目的

特征的自动学习

◆ 3、特征构造

【领会】

特征构造前的准备

【熟知】

运用外部数据的特征建构方法

运用数据探索的特征建构方法

运用专家经验的特征建构方法

运用数据分析的特征建构方法

建构多项式特征及交互特征

◆ 4、特征选择

【熟知】

无效变量（不相关变量、多余变量）

统计为基础的特征选择（过滤式）

模型为基础的变量选择（嵌入式）

递归式的特征选择（包裹式）

高度相关的特征选择

【应用】

运用数据挖掘进行关键特征的选择。同时，评估不同的关键特征选择方法对模型效能的影响。

◆ **5、特征转换**

【领会】

类间可分性最大化的特征转换-线性判别分析（LDA）

矩阵分解法的特征转换-非负矩阵分解法（NMF）

对稀疏矩阵进行特征转换-（截断）奇异值分解法（SVD、TSVD）

【熟知】

线性特征转换-主成分分析（PCA）

非线性的特征转换-核主成分分析（Kernel PCA）

【应用】

运用数据挖掘进行特征的转换。同时，评估不同的特征转换方法对模型效能的影响。

PART 3 机器学习算法（I）

◆ **1、正则化的回归**

【熟知】

回归算法（线性回归、逻辑回归、其他非线性回归）

算法假设

正则的分析学意义、代数学意义、几何意义

正则化的回归模型

【应用】

运用数据挖掘软件建立回归模型，解读模型结果，并评估模型效能。

◆ **2、最近邻法（KNN）**

【领会】

惰性算法的优缺点及应用场景

【熟知】

最近邻法（KNN）的原理

样本点间距离的计算(Manhattan Distance、 City-Block Distance、 Euclidean Distance etc.)

权重的添加与计算

计算复杂度

【运用】

运用数据挖掘软件实现KNN，并能结合其他机器学习方法优化KNN

◆ 3、朴素贝叶斯

【领会】

条件概率与贝叶斯公式

【熟知】

朴素贝叶斯的原理

朴素贝叶斯的模型假设

对数转换与拉普拉斯变换

◆ 4、聚类算法

【领会】

聚类的概念

【熟知】

相似性的衡量与样本点间距离的计算

聚类算法的使用场景

K-Means 聚类算法（含 MiniBatch，K-means++ 等优化方法）

密度聚类算法（DBSCAN）

其他聚类算法（谱聚类，高斯混合聚类等）

聚类结果的评估与群数选择（轮廓系数、决策树评估等）

【应用】

运用数据挖掘软件建立聚类模型，评估与解读模型结果。

◆ 5、关联规则

【领会】

关联规则的概念

【熟知】

关联规则的评估指标（支持度、置信度、提升度）

Apriori 算法（暴力法的弊端、Apriori 算法的理论基础、候选项目组合的产生、候选项目组合的删除）

支持度与置信度的问题（提升度指标）

关联规则的生成

关联规则的延伸（虚拟商品的加入、负向关联规则、相依性网络）

【应用】

运用数据挖掘软件建立关联规则模型，解读模型结果，并提供业务建议。

◆ 6、序列模式

【领会】

序列模式的概念

【熟知】

序列模式的评估指标（支持度、置信度）

AprioriAll 算法（暴力法的问题、AprioriAll 算法的理论基础、候选项目组合的产生、候选项目组合的删除）

序列模式的延伸（状态移转网络）

【应用】

运用数据挖掘软件建立序列模式模型，解读模型结果，并提供业务建议。

PART 4 机器学习算法（II）

◆ 1、决策树（分类树及回归树）

【领会】

PRISM 决策规则算法

CHAID 决策树算法（CHAID 的字段选择方式）

【熟知】

ID3 决策树算法（ID3 的字段选择方式、如何使用决策树来进行分类预测、决策树与决策规则间的关系、ID3 算法的弊端）

C4.5 决策树算法，包括 C4.5 的字段选择方式、C4.5 的数值型字段处理方式、C4.5 的空值处理方式、C4.5 的剪枝方法（预剪枝法、悲观剪枝法）

CART 分类树算法（分类树与回归树、CART 分类树的字段选择方式、CART 分类树的剪枝方法）

CART 回归树算法（CART 回归树的字段选择方式、如何利用模型树来提升 CART 回归树的效能）

【应用】

运用数据挖掘软件建立分类树模型，解读模型结果，并评估模型效能。

运用数据挖掘软件建立回归树模型，解读模型结果，并评估模型效能。

◆ **2、支持向量机**

【领会】

支持向量机概述

线性可分

最佳的线性分割超平面

决策边界

【熟知】

支持向量

线性支持向量机非线性转换

核函数（Polynomial Kernel、Gaussian Radial Basis Function、Sigmoid Kernel）非线性支持向量机

支持向量机与其他算法间的关系

【应用】

运用数据挖掘软件建立分类树模型，解读模型结果，并评估模型效能。

◆ **3、集成方法**

【领会】

集成方法的基本概念

【熟知】

抽样技术

训练数据上的抽样方法

输入变量上的抽样方法

袋装法、随机森林（含随机性与过拟合的关系，袋外样本，热启动等）

提升法（Adaboost、GBDT）、XBGoost（含模型约束）、LightGBM

【应用】

运用数据挖掘软件建立集成方法模型，解读模型结果，并评估模型效能。

◆ **4、异常识别算法**

【领会】

异常识别与离群值的关系

【熟知】

孤立森林

局部异常因子

【应用】

运用数据挖掘软件建立异常识别模型，并提供业务建议。

◆ **5、神经网络**

【领会】

感知机（Perceptron）及感知机的极限

多层感知机（Multi-Layer Perceptron）

【熟知】

BP 神经网络的架构方式

神经元的组成：组合函数（Combination Function）与激活函数（Activation Function）

BP 神经网络如何传递信息

修正权重值及常数项

训练模型前的数据准备

BP 神经网络与逻辑回归、线性回归及非线性回归间的关系

◆ **6、时间序列算法**

【领会】

时间序列分析的基本概念

时间序列的统计分析方法和机器学习方法的联系与区别

循环神经网络 RNN 及长短期记忆网络 LSTM 的发展历程

【熟知】

趋势分解法、ARMA方法、ARIMA 方法的差异和适用场景

ARIMA 方法的建模流程

循环神经网络的架构方式

双向循环神经网络

长短期记忆网络

【应用】

运用数据挖掘软件建立时间序列模型，并提供业务建议。

PART 5 优化分析算法

◆ **1、运筹优化**

【领会】

目标函数的设定原则

线性规划的组成部分、标准形式

整数规划与去尾法线性规划的差异性

非线性规划的组成部分、标准形式

【熟知】

线性规划的计算方法与步骤（含图形法，单纯形法）

整数规划的计算方法与步骤

非线性规划的计算方法与步骤（只考察梯度下降及其衍生方法）

非线性规划在机器学习的模型学习过程中的应用

【应用】

根据题目要求给出目标函数和约束条件，并应用数据挖掘软件求解

◆ 2、流程挖掘

【领会】

业务流程挖掘的概念

【熟知】

业务流程与业务流程管理

流程发现

流程监控

流程挖掘的分析方法和工具

【应用】

运用数据挖掘软件进行流程挖掘，并提供业务建议。

PART 6 数据挖掘实战

◆ 1、多分类与多输出

【领会】

多分类、多标签、多输出的联系与区别

【熟知】

多分类、多标签、多输出的 y 的形式

One Vs Rest, One Vs One

多分类问题的模型评估（只考察混淆矩阵及其衍生指标）

多类多输出

多输出回归

分类器链与回归器链的使用

【应用】

运用数据挖掘软件进行多分类多输出问题的建模，并评估模型结果。

◆ 2、工作流（Pipeline）

【领会】

Pipeline 的基本概念

支持 Pipeline 的常见库

【熟知】

Pipeline 自动数据预处理的方法

Pipeline 自动机器学习的模型建置方法

Pipeline 的调参方法

【应用】

运用 Pipeline 技术，快速应用模型。

◆ 3、类别不平衡问题

【领会】

不平衡数据定义

不平衡数据场景

传统学习方法在不平衡数据中的局限性

类别不平衡所造成的问题

【熟知】

类别不平衡问题的检测方法

过采样技术（Over-sampling）

欠采样技术（Under-sampling）

基于模型的类别不平衡处理技术

【应用】

能运用类别不平衡的处理技术，提升模型的效能

◆ 4、标签缺失问题

【领会】

标签缺失问题的定义

基于业务的打标签方法

【熟知】

自我训练方法（Self Training）

标签传播及其衍生方法（Label Propagation）

◆ **5、优化算法与调参**

【领会】

参数优化的目的

【熟知】

交叉验证

网格搜索（及其衍生方法）

随机参数搜索

贝叶斯搜索

【应用】

运用模型参数优化建立更精准的数据挖掘模型

◆ **6、模型解释**

【领会】

白盒模型与黑盒模型的区别与联系

【熟知】

特征重要性的分析方法（基于模型的方法，基于假设检验的方法，基于后处理的方法）

特征影响的量化分析（Partial Dependence）

决策路径分析与规则提取

代理模型

其他模型解释方法

◆ **7、其他数据挖掘高级方法**

【领会】

内生性问题及其解决思路（如工具变量，双重机器学习等）

实验设计（如AB测试）及其前提假设

自定义损失函数，代价敏感学习

生存分析（含截尾数据的处理与建模）

PART 7 数据挖掘模型管理

◆ **1、MLOps（机器学习研发运营一体化）**

【领会】

MLOps 的背景

MLOps 倡导的机器学习建模与应用流程

MLOps 的设计框架

MLOps 中的数据/概念漂移

2、模型生命周期管理

【领会】

模型生命周期的概念

模型生命周期的主要阶段

模型管理（含模型的监控与迭代等）

模型服务（只考察实时/准实时服务与批量服务）

工作流管理

权限管理

PART 8 数据挖掘实操（案例实操）

根据题目和给定数据，基于有限的算力和时间，使用任何可能的数据挖掘技术与方法（包括且不限于本考纲展示的技术方法），训练模型并评估结果

六、推荐学习书目

1.CDA数据科学研究院.CDA 三级认证教材:敏捷数据挖掘[M].电子工业出版社,2025.(必读)

2.周志华.机器学习（第二版）.清华大学出版社,2016.（必读）（西瓜书）

3.Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 深度学习 DEEP LEARNING, 人民邮电出版社, 2017.（必读）（花书）

4.魏建国、曾珂、翟锟、常国珍.金融数据分析和数据挖掘案例实战[M].电子工业出版社,2025.（选读）

5.爱丽丝·郑,阿曼达·卡萨丽.精通特征工程.人民邮电出版社,2019.（选读）

6.Chris Albon. Python 机器学习手册:从数据预处理到深度学习.电子工业出版社,2019.（选读）

CDA 数据分析认证考试委员会

CDA Institute

注：考试题库中约有 5% 的超纲题