

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, 面向全行业的专业权威国际资格认证, 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 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据分析师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 60 题+多选 30 题+内容相关 10 题)

案例实操题 (1 题)

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用品。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，并能够使用数据挖掘工具(如Python等)进行实操。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%）

- a. 数据挖掘概要（2%）
- b. 数据挖掘方法论（2%）
- c. 基础数据挖掘技术（3%）
- d. 进阶数据挖掘技术（3%）

◆ PART 2 高级数据处理与特征工程（占比 15%）

- a. 高级数据处理（3%）
- b. 特征工程概要（1%）
- c. 特征建构（2%）
- d. 特征选择（3%）
- e. 特征转换（3%）
- f. 特征学习（3%）

◆ PART 3 自然语言处理与文本分析（占比 10%）

- a. 自然语言处理概要（占比 1%）
- b. 分词与词性标注（占比 2%）
- c. 文本挖掘概要（占比 1%）
- d. 关键词提取（占比 2%）
- e. 文本非结构数据转结构（占比 4%）

◆ PART 4 机器学习算法（占比 30%）

- a. 正则化的回归模型（2%）
- b. 决策树（分类树及回归树）（5%）
- c. 支持向量机（1%）
- d. 集成方法（9%）
- e. 聚类分析（4%）
- f. 关联规则（3%）
- g. 序列模式（1%）

h. 模型评估 (5%)

◆ **PART 5 数据挖掘实战 (占比 10%)**

- a. Pipeline (2%)
- b. 类别不平衡问题 (4%)
- c. 模型优化与调参 (4%)

◆ **PART 6 深度学习算法 (占比 13%)**

- a. 感知机 (1%)
- b. BP 神经网络 (3%)
- c. 卷积神经网络 (Convolutional Neural Networks, CNN) (2%)
- d. 循环神经网络 (Recurrent Neural Networks, RNN) (2%)
- e. 优化算法 (2%)
- f. 深度学习中的正则化 (2%)
- g. 自编码器与表示学习 (1%)

◆ **PART 7 大语言模型与人工智能(NLP) (占比 12%)**

- a. 注意力机制与Transformer (3%)
- b. 大语言模型及其应用 (5%)
- c. 微调与知识学习 (3%)
- d. Agent (1%)

五、科目内容

PART 1 数据挖掘概论

◆ **1、数据挖掘概要**

【领会】

数据挖掘在政府部门及互联网、金融、零售、医药等行业的应用

【熟知】

数据挖掘的起源、定义及目标

数据挖掘的发展历程

【应用】

根据给定的数据建立一个数据挖掘的项目

◆ **2、数据挖掘方法论**

【熟知】

数据挖掘步骤（字段选择、数据清洗、字段扩充、数据编码、数据挖掘、结果呈现）

数据挖掘技术的产业标准（CRISP-DM 及 SEMMA）

【应用】

运用数据挖掘进行不同文件格式的数据导入，并进行初步的数据探索，探索的内容包含数值型字段的描述性统计分析、直方图（需与目标字段做连接）、缺失值分析及类别型字段的描述性统计分析、条形图（需与目标字段做连接、缺失值分析。数据探索的结果可进行初步的字段筛选。

◆ 3、基础数据挖掘技术**【领会】**

可视化技术（能使用相关工具根据业务问题做出可视化数据报告）

【熟知】

案例为本的学习(Case-based Learning)：KNN(K-Nearest Neighbor)原理

数据的准备

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

【应用】

运用数据挖掘中的 KNN 算法进行分类预测、数字预测及内容推荐。建模的过程需考虑将数据进行适当的转换以获得更优的分析结果。

◆ 4、进阶数据挖掘技术**【熟知】**

数据挖掘技术的功能分类

描述型数据挖掘/无监督数据挖掘（关联规则、序列模式、聚类分析）

预测型数据挖掘/有监督数据挖掘（分类、预测）

PART 2 高级数据处理与特征工程**◆ 1、高级数据预处理****【领会】**

数据过滤（理解如何通过数据过滤的方式，建立区隔化模型，以提升模型的预测效果）

内/外部数据的扩充方法

【熟知】

缺失值的高级填补技术，包括 KNN 填补、XGBoosting 填补

高级数据转换技术，包括数据泛化(Generalization)、数据趋势离散化(Trend Discretization)

【应用】

运用高级数据预处理技术进行数据过滤，以建立区隔化模型

运用高级数据预处理技术进行缺失值的侦测及填补

运用高级数据预处理技术进行数据泛化的处理

运用高级数据预处理技术进行数据趋势离散化的处理

评估上述不同的数据处理方法对模型效能的影响

◆ 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

特征建构的方法

特征转换的方式

特征的自动学习

以 AI 促进 AI

◆ 3、特征建构

【领会】

特征建构前的准备

特征的空值处理

特征的标准化的

【熟知】

类别型特征的编码

顺序型特征的编码

数值型特征的分箱

建构多项式特征

建构交互特征

特征的归一化

【应用】

运用数据挖掘对特征进行适当的建构，以作为下阶段特征选择的输入

◆ 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（卡方检验、ANOVA 检验及 T 检验）

模型为基础的变量选择（决策树、逻辑回归、随机森林）

高度相关特征的选择

递归式的特征选择

【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（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。

◆ 6、特征学习

【熟知】

关联规则为基础的特征学习

神经网络为基础的特征学习

深度学习为基础的特征学习

词嵌入为基础的文本特征学习

【应用】

运用数据挖掘进行自动的特征学习。同时，评估不同的特征学习方法对模型效能的影响。

PART 3 自然语言处理与文本分析**◆ 1、自然语言处理概要****【领会】**

中文语意平台

自然语言处理的研究范畴

分词

词根还原

词性标注

同义词标订

概念标订

角色标订

◆ 2、分词与词性标注**【领会】**

词性的种类及意义

【熟知】

N-Gram 分词

分词及词性标注的难点

法则式分词法

统计式分词法

词性标注

【运用】

运用中文分词及词性标注技术对多篇文章进行分词及词性标注

◆ 3、文本挖掘概要**【领会】**

信息检索技术之全文扫描

信息检索技术之签名文件

信息检索技术之逐项反转

控制字汇

关键词索引

文本可视化

文本挖掘的应用

【熟知】

信息检索技术之向量空间模型

文本挖掘的处理流程

【应用】

将多篇文件及查询转为向量格式，并计算查询与文件间的相似度。

◆ 4、关键词提取

【熟知】

TF、DF 及 IDF

词性

关键词的提取方法

【应用】

对多篇文件及查询中的词，计算 TF、DF、IDF 及词性并提取重要的关键词。

◆ 5、文本非结构数据转结构

【熟知】

词袋模型

PCA

矩阵分解

词嵌入模型 Glove

词嵌入模型 Word2Vec (Skip-Gram & CBOW)

【应用】

对多篇文件进行词嵌入模型的训练及使用。

将结构化后的文件进行文本分类、情绪分析、文本聚类及文本摘要的应用。

PART 4 机器学习算法

◆ 1、正则化的回归模型

【熟知】

回归模型（线性回归、逻辑回归、模型假设）
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、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。

◆ 3、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

决策边界与支持向量

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

支持向量机与神经网络间的关系

◆ 4、集成方法

【领会】

集成方法概述

【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）

提升法（Adaboost、GBDT、xgboost、LightGBM）

【应用】

运用数据挖掘软件建立组合方法模型，解读模型结果，并评估模型效能。

◆ 5、聚类分析

【领会】

聚类的概念

【熟知】

相似性的衡量（二元变量的相似性衡量、混合类别型变量与数值型变量的相似性衡量）

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

聚类算法（Exclusive vs. Non-Exclusive (Overlapping)的聚类算法、分层聚类法、划分聚类法）

分层聚类算法（单一链结法、完全链结法、平均链结法、中心法、Ward's 法）

划分聚类算法（K-Means 法、EM 法、K-Medoids 法、神经网络 SOM 法、两步法）

密度聚类算法（DBSCAN）

群数的判断 (R-Squared (R²)、Semi-Partial R-Squared、Root-Mean-Square Standard Deviation (RMSSTD)、轮廓系数(Silhouette Coefficient))

【应用】

运用数据挖掘软件建立聚类模型，解读模型结果，并提供营销建议。

◆ 6、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，并提供营销建议。

◆ 7、序列模式

【领会】

序列模式的概念

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，并提供营销建议。

◆ 8、模型评估

【熟知】

混淆矩阵（正确率(Accuracy)、查准率(Precision)、查全率(Recall)、F-指标(F-Measure)）

KS 图（KS Chart）

ROC 图（ROC Chart）

GINI 图（GINI Chart）

回应图（Response Chart）

增益图（Gain Chart）

提升图（Lift Chart）

收益图（Profit Chart）

平均平方误差（Average Squared Error）

【应用】

运用数据挖掘软件比较不同模型间的优劣

PART 5 数据挖掘实战**◆ 1、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。

◆ 2、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）

模型惩罚技术

【应用】

能运用类别不平衡的处理技术，提升模型的效能

◆ 3、模型优化与调参**【领会】**

模型参数优化的目的与方法

建模阈值优化的目的与方法

【熟知】

网格搜索
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

PART 6 深度学习算法

◆ 1、感知机**【领会】**

感知机 (Perceptron) 的由来
感知机 (Perceptron) 及感知机的极限
多层感知机 (Multi-Layer Perceptron)

◆ 2、BP神经网络**【领会】**

BP 神经网络概述 (理解神经网络的由来及发展历程)

【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 (分类模型的数据准备、预测模型的数据准备)
BP 神经网络与逻辑回归、线性回归及非线性回归间的关系

【应用】

运用数据挖掘软件建立 BP 神经网络模型, 解读模型结果, 并评估模型效能。

◆ 3、卷积神经网络**【领会】**

卷积神经网络 CNN 的由来及发展历程

【熟知】

卷积的重要思想 (稀疏交互, 参数共享, 等变表示)

卷积运算

池化（不变性）

【应用】

运用数据挖掘软件建立卷积神经网络模型，并评估模型效能。

◆ 4、循环神经网络

【领会】

循环神经网络 RNN 的由来及发展历程

长短期记忆网络

【熟知】

循环神经网络的架构方式

双向循环神经网络

【应用】

运用数据挖掘软件建立循环神经网络模型，并评估模型效能。

◆ 5、优化算法

【领会】

神经网络优化中的挑战（局部问题，梯度消失和爆炸，长期依赖）

基于梯度的优化方法（梯度下降）

随机梯度下降

二阶近似方法（牛顿法等）

自适应学习率算法（Adam等）

◆ 6、深度学习中的正则化

【领会】

参数范数惩罚（L1, L2 正则）

【熟知】

提前终止（early stop）

Dropout

◆ 7、自编码器与表示学习

【领会】

欠完备与正则自编码器

表示能力、层的大小和深度

表示学习的基本概念

无监督预训练

迁移学习和领域自适应

PART 7 大语言模型与人工智能(NLP)**◆ 1、注意力机制与Transformer****【领会】**

注意力机制的由来

残差连接

【熟知】

注意力与多头注意力

自注意力和位置编码

Transformer架构

◆ 2、大语言模型及其应用**【领会】**

大语言模型的发展历程与未来展望

常用大语言模型的架构

Base Model 与 Pretraining

【熟知】

Tokenization

Prompt技术

多模态

【应用】

运用 AI 工具辅助进行数据挖掘。

◆ 3、微调与知识学习**【领会】**

Supervised Fine-tuning

Reinforcement Learning

【熟知】

大语言模型的知识学习

Fine-tuning

Prompt tuning

【应用】

根据业务需求，使用相关框架微调大语言模型

◆ **4、Agent**

【领会】

向量数据库

大语言模型的外挂知识库

上下文记忆

代理与链

六、推荐学习书目

说明：推荐学习书目中，部分书籍结合软件，考试中客观选择题部分不考查软件操作使用，案例实操部分需要考生运用相关软件进行建模分析，考生可根据自身需求选择性学习。参考书目不需全部学完，根据考纲知识点进行针对性学习即可。

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

[2] 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）

[3] 常国珍, 赵仁乾, 张秋剑. Python数据科学, 技术详解与商业实践. 机械工业出版社, 2018.（必读）

[4] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 动手学深度学习（第二版）, 2023.（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）

- [7] 开源模型网站: Hugging Face (<https://huggingface.co/>) (拓展学习)
- [8] 数据挖掘网站: Kaggle (<https://www.kaggle.com/>) (拓展学习)

CDA 数据分析认证考试委员会

CDA Institute