CDA LEVEL Ⅱ建模分析师考试大纲

CERTIFIED DATA ANALYST LEVEL Ⅱ EXAMINATION OUTLINE

一、总则

「CDA 数据分析师人才行业标准」是面向全行业数据分析及大数据相关岗位的一套科学化、专业化、正规化、系统化的人才技能准则。经管之家 CDA 数据分析师认证考试是评判「标准化人才」的唯一考核路径。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

包括客观题和案例操作题两部分：其中客观题（单选+多选）考试时间为 90 分钟，上机答题；案例操作题考试时间为 120 分钟，闭卷，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、SQL、SPSS MODELER、R、SAS、WEKA 等，进行案例操作分析。案例数据将统一提供 CSV 文件）。

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。

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：考生能够领会了解规定的知识点，并能够了解规定知识点的内涵与外延，了解其内容要点和它们之间的区别与联系，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、考试范围

◆ **PART 1 数据挖掘基础理论 （占比 20%）**

a. 数据挖掘概要（2%）

b. 数据挖掘方法和原理（7%）

c. 数据挖掘技术基础（5%）

d. 数据挖掘技术进阶（6%）

◆ **PART 2 数据预处理 （占比 25%）**

a. 字段选择（2%）

b. 数据清洗（8%）

c. 字段扩充（2%）

d. 数据编码（8%）

e. 特征提取技术（5%）

◆ **PART 3 预测型数据挖掘模型 （占比 40%）**

a. 朴素贝叶斯（5%）

b. 线性回归（3%）

c. 决策树（分类树及回归树）（8%）

d. 神经网络与深度学习（6%）

e. 逻辑回归（2%）

f. 支持向量机（4%）

g. 集成方法（5%）

h. 模型评估（7%）

◆ **PART 4 描述型数据挖掘模型 （15%）**

a. 聚类分析（6%）

b. 关联规则（6%）

c. 序列模式（3%）

五、考试内容

PART 1 数据挖掘基础理论

◆ **1、数据挖掘概要**

【领会】

数据挖掘在政府部门及互联网、金融、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 Project

◆ **2、数据挖掘方法和原理**

【熟知】

数据库中的知识发现步骤（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】

运用数据挖掘软件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（需与目标字段做链接）、缺失值分析及类别型字段的描述性统计分析、条形图（需与目标字段做链接）、缺失值分析。数据探索的结果可进行初步的字段筛选。

◆ **3、数据挖掘技术基础**

【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】

描述性统计

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】

运用数据挖掘软件中的 KNN 模块或者算法进行分类预测及 KNN 电影推荐。建模的过程需考虑将数据进行适当的转换以获得较佳的分析结果。

◆ **4、数据挖掘技术进阶**

【熟知】

数据挖掘技术的功能分类

描述性数据挖掘/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）

数据挖掘技术的绩效增益，包括混淆矩阵（正确率、查准率、查全率、F-指标）、Gain Chart、Lift Chart、Profit Chart。

PART 2 数据预处理

◆ 1、字段选择

【领会】

数据整合 (理解不同数据来源的整合问题)

数据过滤 (理解如何透过数据过滤的方式，建置区隔化模型，以提升模型的预测效能)

【应用】

运用数据挖掘软件进行数据过滤，以建立区隔化模型。

◆ 2、数据清洗

【熟知】

错误值、离群值、缺失值的侦测及处理

【应用】

运用数据挖掘软件进行错误值、离群值、缺失值的侦测及处理。离群值的侦测可比较平均值法与四分位数法的差异。同时，需熟悉天花板/地板法（盖帽法）的离群值处理方式。缺失值的处理则需熟悉利用建模的方式来填补缺失值。

◆ 3、字段扩充

【领会】

内/外部数据的扩充方法

【应用】

运用数据挖掘软件进行字段扩充，及评估扩充前后对模型效能的提升程度，并能加以说明原由。

◆ 4、数据编码

【熟知】

数据转换，包括数据正规化(Normalization)、数据泛化(Generalization)、数据离散化(Discretization)

数据精简（记录精简、域值精简、字段精简）

数据集的切割（随机取样切割法、分层抽样切割法）

【应用】

运用数据挖掘软件进行数据转换及数据集的切割（能将数据切割为训练、验证及测试数据集）。同时，评估不同的数据转换方法对模型效能的影响。

◆ **5、特征提取技术**

【熟知】

无效变量（不相关变量、多余变量）的ㄋ分方式

统计方式的变量选择（卡方检验、ANOVA 检验及 T 检验）

模型方式的变量选择（决策树、逻辑回归、随机森林）

变量提取（PCA、LDA）

【应用】

运用数据挖掘软件进行关键变量的挖掘。同时，评估不同的关键变量ㄙ方法对模型效能的影响。

PART 3 预测型数据挖掘模型

◆ **1、朴素贝叶斯**

【熟知】

朴素贝叶斯（独立性假设、概率的正规化、拉普拉斯转换、空值的问题）

【应用】

运用数据挖掘软件建立朴素贝叶斯模型，解读模型结果，并评估模型效能。

◆ **2、线性回归**

【熟知】

简单线性回归

多元线性回归

相关系数

回归模型的效能评估(MAE、MSE、RMSE、R2、Adjusted R2、AIC & BIC)

【应用】

运用数据挖掘软件建立线性回归模型，解读模型结果，并评估模型效能。

◆ **3、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 决策树算法（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。

◆ **4、神经网络与深度学习**

【领会】

BP 神经网络概述（理解神经网络的由来及发展历程）

卷积神经网络（Convolutional Neural Networks, CNN）（理解卷积神经网络 CNN 的由来及发展历程）

递归神经网络（Recurrent Neural Networks, RNN）（理解递归神经网络 RNN 的由来及发展历程）

【熟知】

感知机(Perceptron)及感知机的极限

多层感知机(Multi-Layer Perceptron)

BP 神经网络的架构方式

神经元的组成：组合函数(Combination Function)与活化函数(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。

◆ 5、逻辑回归

【熟知】

逻辑回归与 BP 神经网络的关系

逻辑回归的字段选择方式（前向递增法、后向递减法、逐步回归法）

【应用】

运用数据挖掘软件建立逻辑回归模型，解读模型结果，并评估模型效能。

◆ 6、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

支持向量机与神经网络间的关系

【应用】

运用数据挖掘软件建立支持向量机模型，解读模型结果，并评估模型效能。

◆ 7、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（袋装法、提升法）

输入变量上的抽样方法（随机森林）

【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。

◆ 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 4 描述型数据挖掘模型

◆ 1、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 (R^2)、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。

◆ **2、关联规则**

【领会】

关联规则的概念

【熟知】

关联规则的评估指针（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。

◆ **3、序列模式**

【领会】

序列模式的概念

【熟知】

序列模式的评估指针（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。

六、推荐学习书目

说明：推荐学习书目中，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] 经管之家. CDA 数据分析师备考手册（电子版）. 2019. (必读)

[2] 经管之家. SPSS Modeler+Weka 数据挖掘从入门到实战，电子工业出版社，2019. (选读)

[3] Jiawei Han, Micheline Kamber, Jian Pei. 数据挖掘：概念与技术（原书第 3 版）[M]. 范明，孟小峰 译, 机械工业出版社，2012. (必读)

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. 数据挖掘导论（原书第 2 版）[M]. 段磊，张天庆译, 机械工业出版社，2019. (必读)

[5] 周志华. 机器学习[M]. 清华大学出版社，2016. (必读)

[6] 赵卫东，董亮. 机器学习[M]. 人民邮电出版社，2018. (选读)

[7] 数据挖掘网站：KDnuggets (https://www.kdnuggets.com/) (拓展学习)

[8] 数据挖掘网站：Kaggle (https://www.kaggle.com/) (拓展学习)

CDA Institute

经管之家 CDA 数据分析研究院