CDA LEVEL Ⅱ 考试大纲

CERTIFIED DATA ANALYST LEVEL Ⅱ EXAMINATION OUTLINE

CDA 考试大纲是 CDA 命题组基于 CDA 数据分析师等级认证标准而设定的一套科学、详细、系统的考试纲要。考纲规定并明确了 CDA 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 CDA 考试大纲进行相关知识的复习。

CDA 建模分析师考试大纲

基础理论 (占比 20%)

a. 数据挖掘简介(2%)
b. 数据挖掘方法论(7%)
c. 基础数据挖掘技术(5%)
d. 进阶数据挖掘技术简介(6%)

数据前处理 (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(4%)
e. 数据分区(4%)
f. 关键变量挖掘技术(5%)

预测型数据挖掘模型 (占比 40%)

a. 贝氏网络(5%)
b. 线性回归(3%)
c. 决策树(分类树及回归树)(10%)
d. 神经网络(6%)
e. 罗吉斯回归(2%)
f. 支持向量机(4%)
g. 集成方法(5%)
h. 模型评估(5%)

描述型数据挖掘模型 (占比 15%)

a. 聚类分析(6%)
b. 关联规则(6%)
c. 序列模式(3%)

CDA 建模分析师大纲解析

根据 CDA 数据分析师认证考试大纲，经管之家 CDA 数据分析研究院给出了详细解析，以“领会”，“熟知”，“应用”三个不同的级别将每一个知识点进行分解，建议考生应该按照不同的知识掌握程度有目的性的进行复习。

1. 领会：要求应考者能够记忆规定的有关知识点的主要内容，并能够了解规定的有关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，并能根据考核的不同要求，做出正确的解释、说明和阐述。
2. 熟知：要求应考者必须熟悉的理论知识，并能够正确理解和记忆相关的理论方法，根据考核的不同要求，做出逻辑严密的解释、说明和阐述。
3. 应用：要求应考者必须掌握知识点的主要内容，并能够结合工具进行商业应用，根据考核的具体要求，做出问题的具体实施流程和策略。

PART 1

基础理论部分

- 数据挖掘简介
  1. 领会：数据挖掘在政府部门及各行业的应用。
  2. 熟知：数据挖掘的起源、定义及目标，数据挖掘的发展历程。
  3. 应用：根据给定的数据建立一个数据挖掘的 Project。

- 数据挖掘方法论

1. 熟知：数据库中的知识发掘步骤(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 (IBM SPSS)及 SEMMA (SAS))。

2. 应用：运用数据挖掘软件进行不同文件格式的数据汇入，并进行初步的数据探索。探索的内容包含数值型字段的叙述性统计分析，直方图(需与目标字段做链接)，遗缺值分析及类别型字段的叙述性统计分析，分布图(需与目标字段做链接)，遗缺值分析。数据探索的结果可进行初步的字段筛选。

- 基础数据挖掘技术

1．领会：叙述性统计，可视化技术, KNN(K Nearest Neighborhood)原理，KNN 电影推荐案例。

2．熟知：案例为本的学习 (Case-based Learning)，数据的准备，距离的计算(Manhattan Distance / City-Block Distance, Euclidean Distance)。

3．应用：运用数据挖掘软件中的 KNN 进行分类预测。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。

- 进阶数据挖掘技术简介

1．领会：数据挖掘技术的功能分类，数据挖掘网站(KDnuggets & Kaggle)。

2．熟知：描述性数据挖掘/非指导性数据挖掘(关联规则、序列模式、聚类分析)，预测性数据挖掘/指导性数据挖掘(分类、预测)，数据挖掘技术的绩效增益(混乱矩阵(正确率、响应率、捕捉率、F-指标)、Gain Chart、Lift Chart、Profit Chart)。

PART 2

数据前处理

- 字段选择

1．领会：数据整合，数据过滤。

2．应用：运用数据挖掘软件进行数据过滤，以建立区隔化模型。

- 数据清洗

1．熟知：错误值、离群值、遗失值的侦测及处理。

2．应用：运用数据挖掘软件进行错误值、离群值、遗失值的侦测及处理。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（盖帽法或 Winsorize 法）的离群值处理方式。遗失值的处理：常值填充、均值或众数填充、利用建模方式填充。

- 字段扩充

1．领会：内/外部数据的扩充方法。

2．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。

- 数据编码

1．熟知：数据转换(数据正规化、数据一般化、连续性指派、数据离散化)，数据精简(记录精简、域值精简、字段精简、WOE)，连续变量分箱技术。

2．应用：评估不同的数据转换方法对模型效能的影响。

- 数据分区

1．熟知：模型泛化与过渡拟合，数据集的切割(随机取样切割法、分层抽样切割法)，三种数据集（训练、验证及测试数据集）在数据挖掘中的作用。

2．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)。

- 关键变量挖掘技术

1．领会：无效变量，不相关变量，多余变量。

2．熟知：统计方式的变量选择(卡方检定、ANOVA 检定及 T 检定、连续变量相关检验（斯皮尔曼（SPEARMAN）秩相关系数）、信息价值（IV）)，模型方式的变量选择(决策树、罗吉斯回归)。

3．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选择方法对模型效能的影响。

- 变量压缩

3．熟知：连续变量压缩技术(主成分、变量聚类)、分类变量压缩技术（水平聚类、WOE 打分）。

4．应用：评估不同的数据转换方法对模型效能的影响。

PART 3

预测型数据挖掘模型

- 贝氏网络（贝叶斯网络）

1．熟知：简单贝氏网络(独立性假设、概率的正规化、概率为 0 的问题、空值的问题)、贝氏网络。

2．应用：运用数据挖掘软件建立贝氏网络模型，解读模型结果，并评估模型效能。

- 线性回归

1．领会：参数估计（最小二乘法、矩估计、极大似然估计），目标函数设置（普通最小二乘法、加二阶惩罚项（脊回归）和加一阶惩罚项（Lasso 算法(LARS)））

2．熟知：简单线性回归，复回归，相关系数，回归模型的效能评估(MAE, MSE, RMSE, R^2, Adjusted R^2, AIC & BIC)。

3．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。

- 决策树(分类树及回归树)

1．领会：PRISM 决策规则算法、CHAID 决策树算法(CHAID 的字段选择方式)

2．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 算法的问题)，C5.0 决策树算法(C5.0 的字段选择方式、C5.0 的数值型字段处理方式、C5.0 的空值处理方式、C5.0 的砍树方法)，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 分类树的砍树方法)，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)。

3．应用：运用数据挖掘软件建立分类树及交互式分类树模型，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。

- 神经网络

1．领会：神经网络概述、倒传递神经网络概念与算法、理解径向基函数

2．熟知：多层感知器；倒传递神经网络，倒传递神经网络的架构方式，神经元的组成，神经网络如何传递讯息，神经网络如何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)、倒传递神经网络与罗吉斯回归、线性回归及非线性回归间的关系；径向基神经网络

3．应用：运用数据挖掘软件建立神经网络模型，解读模型结果，并评估模型效能。

- 罗吉斯回归（逻辑回归或 logistic 回归）

1．领会：逻辑回归的极大似然估计，加惩罚项的逻辑回归的极大似然估计

2．熟知：罗吉斯回归与倒传递神经网络的关系，罗吉斯回归的字段选择方式(前向递增法、后向递减法、逐步回归法)。

3．应用：运运用数据挖掘软件建立罗吉斯回归模型，解读模型结果，并评估模型效能。

- 支持向量机

1．领会：支持向量机概述，线性可分，最佳的线性分割超平面，决策分界线，线性不可分，软间隔最大化、SMO。

2．熟知：支持向量，线性支持向量机，非线性转换，核心函数(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。

3．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。

- 集成方法

1．领会：集成方法概述。

2．熟知：训练数据上的集成方法(袋装法、提升法)，输入变量上的集成方法(随机森林)。

3．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。

- 模型评估

1．熟知：混乱矩阵(正确率、响应率、捕捉率、F-指标)，KS Chart，ROC Chart & GINI Chart，Response Chart, Gain Chart, Lift Chart，Profit Chart、Average Squared Error。

2．应用：运用数据挖掘软件比较不同模型间的优略。

- 阈值设置

1．熟知：先验概率调整、混淆矩阵设定阀值。

2．应用：可以根据利润最大化或成本最小化设定阈值；可以结合业务需求进行阈值调整。

- 预测的监测

1．熟知：数据稳定性检验、评分稳定性指标、特征分布指标、模型正确性指标、变量有效性指标。

2．应用：在模型运用之前运用前端监控技术进行模型的稳定性。在模型运用之后运用后端监控模型的正确性。

PART 4

描述型数据挖掘模型

- 聚类分析

1．领会：聚类的概念。

2．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)，距离的计算(Manhattan Distance / City-Block Distance、Euclidean Distance)，聚类算法(阶层式聚类法（也称作层次聚类或系统聚类）、分割式聚类法)、阶层式聚类算法(单一链结法、完全链结法、平均链结法、中心法、Ward's 法)，分割式聚类算法(K-Means 法、K-Medoids 法、两步法)，群数的判断(R-Squared (R^2)、Semi-Partial R-Squared、Root-Mean-Square Standard Deviation (RMSSTD))。

3．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。

- 关联规则

1．领会：关联规则的概念。

2．熟知：关联规则的评估指针(支持度、信赖度（置信度）)，Apriori 算法(暴力法的问题、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)，支持度与信赖度的问题(提升度指标)，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)。

3．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。

- 序列模式

1．领会：序列模式的概念。

2．熟知：序列模式的评估指针(支持度、信赖度)，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)。

3．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。

参考书目

- 数据挖掘：概念与技术（第 3 版），作者：（加）韩家炜，堪博 著，范明，孟小峰 译，机械工业出版社。
- 数据挖掘导论，[美]Pang-Ning Tan，Michael Steinbach，Vipin Kumar 著，译者：范明 范宏建，人民邮电出版社。
- 经济计量分析，[美]威廉 H. 格林著，译者：王明霞等
- Data Mining: A Tutorial Based Primer，作者：Roiger, Richard, Geatz, Michael，Addison-Wesley。
- Data Mining：Concepts and Techniques(2nd.Ed)，作者：Jia Han, Micheline Kamber

CDA 考试报名链接: exam.cda.cn

经管之家 CDA 数据分析研究院

2016.3