数据应用开发与服务(Python)

职业技能等级标准

（2021 年 1.0 版）

北京中软国际信息技术有限公司　制定

2021 年 3 月　发布

目 次

前　言

本标准按照GB/T 1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本标准起草单位：北京中软国际信息技术有限公司、中软国际科技服务有限公司、中软国际(中国)科技有限公司、中国软件与技术服务股份有限公司、中国软件行业协会、上海华腾软件系统有限公司、大连华信计算机技术股份有限公司、北京北控三兴信息技术有限公司、腾讯云计算（北京）有限责任公司、广州番禺职业技术学院、北京信息职业技术学院、陕西国防工业职业技术学院、南京信息职业技术学院、南通职业大学、山西机电职业技术学院、江苏经贸职业技术学院、北京工业职业技术学院、北京交通职业技术学院、北京市大兴区第一职业学校。

本标准主要起草人：周海、余明辉、杨鹏、李修霖、姜楠、成焕、王晓华、宋丹、范路桥、詹增荣、刘玉海、罗春红、吴洪贵、孟繁增、何淼、苏维刚、方圆、李粉霞、陈根琴。

1 范围

本标准规定了数据应用开发与服务(Python)职业技能等级对应的工作领域、工作任务及职业技能要求。

本标准适用于数据应用开发与服务(Python)职业技能培训、考核与评价，相关用人单位的人员聘用、培训与考核可参照使用。

2 规范性引用文件

下列文件对于本标准的应用是必不可少的。 凡是注日期的引用文件,仅注日期的版本适用于本标准。凡是不注日期的引用文件,其最新版本适用于本标准。

GB/T 5271.6-2000 信息技术 词汇 第 6 部分:数据的准备与处理

GB/T 5271.31-2006 信息技术 词汇 第 31 部分：人工智能机器学习

GB/T 35295-2017 信息技术 大数据 术语

GB/T 37721-2019 信息技术 大数据分析系统功能要求

3 术语和定义

GB/T 5271.6-2000、GB/T 5271.31-2006、GB/T 35295-2017、GB/T 37721-2019 等界定的以及下列术语和定义适用于本标准。

3.1 数据采集（数据获取） Data Acquisition

收集并输入数据的过程。

[GB/T 5271.6-2000 定义 06.02.10]

3.2 数据清洗 Data Cleaning

数据清洗是指发现并纠正数据文件中可识别的错误的一道程序，包括检查数据一致性，处理无效值和缺失值等。

[GB/T 37721-2019, 6.2 数据清洗功能要求]

3.3 **数据转换(数据变换)** Data Converting(Data Transforming)

把数据的表示从一种形式改成另一种形式但不改变数据所传达的信息； 按照规定的规则改变数据的形式但基本上不改变数据的含义。

[GB/T 5271.6-2000 定义 06.03.06 定义 06.03.04]

[GB/T 37721-2019 6.3 数据转换功能要求]

3.4 **数据分析** Data Analytics

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总和理解并消化，以求最大化地开发数据的功能，发挥数据的作用。数据分析是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

[GB/T 35295-2017，定义 2.1.48]

3.5 **数据建模** Data Modeling

数据建模指的是对现实世界各类数据的抽象组织，建立一个适合的模型对数据进行处理。常见的算法有分类（有明确类别）、聚类（无明确类别）、关联、回归等。

[GB/T 37721-2019,7.2.2 支持算法的要求]

3.6 **数据可视化** Data visualization

数据可视化是关于数据之视觉表现形式的研究；其中，这种数据的视觉表现形式被定义为一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。简单来说，数据可视化是用图形、图像、图表等方式来表征数据的规律。

[GB/T 37721-2019 7.4 可视化功能要求]

3.7 **机器学习** Machine Learning

功能单元通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程。

[GB/T 5271.31-2006，定义 31.01.02]

4 适用院校专业

中等职业学校：软件与信息服务、计算机网络技术、计算机应用、移动应用技术与服务、物联网技术应用等专业。

高等职业学校：人工智能技术服务、大数据技术与应用、云计算技术与应用、软件技术、软件与信息服务、计算机应用技术、计算机信息管理、计算机网络技术等专业。

应用型本科学校：人工智能、数据科学与大数据技术、软件工程、信息管理与信息系统、计算机科学与技术等专业。

5 面向职业岗位（群）

【数据应用开发与服务(Python)】（初级）：主要面向软件与信息服务企业、数据分析处理服务企业、人工智能服务企业、互联网企业，以及向数字化转型的传统型企事业单位和政府机构的信息化部门，从事 Python 编程、本地小数据量的数据管理、数据文档整理、应用程序部署和维护等工作岗位。

【数据应用开发与服务(Python)】（中级）：主要面向软件与信息服务企业、数据分析处理服务企业、人工智能服务企业、互联网企业，以及向数字化转型的传统型企事业单位和政府机构的信息化部门，从事中等数据量的数据采集、特征处理、机器学习建模、可视化以及与数据相关的 web 应用开发等工作岗位。

【数据应用开发与服务(Python)】（高级）：主要面向软件与信息服务企业、

数据分析处理服务企业、人工智能服务企业、互联网企业，以及向数字化转型的传统型企事业单位和政府机构的信息化部门，从事海量数据采集、特征选取、模型优化以及与数据服务应用程序的设计与开发等工作岗位。

6 职业技能要求

6.1 职业技能等级划分

数据应用开发与服务(Python)职业技能等级分为三个等级：初级、中级、高级，三个级别依次递进，高级别涵盖低级别职业技能要求。

【数据应用开发与服务(Python)】(初级)：主要面向本地结构化数据的统计、清洗和转换，以及客户端应用程序的开发。能够根据给定的文档和规范，独立完成基于小规模文件和关系型数据的数据采集、数据统计、数据清洗和可视化图表制作等工作，并能够在开发团队中承担终端应用程序开发、接口测试、文档编制、系统部署等任务。

【数据应用开发与服务(Python)】(中级)：主要面向结构化和非结构化数据的数据预处理和机器学习建模，以及服务端应用程序的开发。能够从多种格式和多种数据来源，采集和存储中等规模的数据；能够在了解一定的数学和算法原理的基础上，对数据进行归一化、文本数值化、离散化等预处理操作，构建基本的回归、分类和聚类等机器学习模型；并能够在开发团队中承担 Web 服务应用的开发和部署等工作任务。

【数据应用开发与服务(Python)】(高级)：主要面向大数据平台的采集处理，模型的优化，以及模型的高级应用。能够设计和实现针对大数据系统的采集；能够针对给定的业务目标，采用主流算法构建和优化数据模型；并能够在开发团队中承担富功能可视化应用和数据服务应用设计、开发工作任务。

6.2 职业技能等级要求描述

表 1 数据应用开发与服务(Python)职业技能等级要求（初级）

工作领域	工作任务	职业技能要求
1. 数据应用程序模块开发	1.1 开发环境搭建	1.1.1 能够在 Windows 上正确安装和配置 Python 3.x 运行环境; 1.1.2 能够分别使用 Pip 和 Conda 完成 Python 包的安装、卸载、升级、查询操作; 1.1.3 能够安装、配置和使用开发工具(VSCode)进行 Python 代码的编写、运行和调试; 1.1.4 能够正确安装和配置 Jupyter Lab，使之在单机上正常运行 Python 代码。
	1.2 终端应用程序开发	1.2.1 能够根据简单的业务流程和业务规则，运用 Python 语言的流程控制语句、函数、模块等功能编制应用程序; 1.2.2 能够以简洁高效的方式处理应用程序中的常见数据，包括：用户输入输出、字符串、日期和时间; 1.2.3 能够合理选择数据结构存取数据，包括：tuple、list、set、dict; 1.2.4 能够使用类、对象、继承等面向对象方法封装、扩展业务功能; 1.2.5 能够通过合理的异常处理增强程序的容错能力和健壮性。
	1.3 软件开发过程管理	1.3.1 熟悉应用程序开发的过程、阶段及每个阶段的任务; 1.3.2 能够在理解开发任务要求的基础上，制定本人工作计划; 1.3.3 能够在开发过程中正确使用源代码管理工具进行版本管理。
2. 数据采集	2.1 文件和目录操作	2.1.1 能够使用 file 对象读写文本文件; 2.1.2 能够分别使用 file 对象和 numpy 读写二进制文件; 2.1.3 能够使用 os 模块对目录和文件进行操作,包括:创建目录、遍历文件和子目录、复制、更名。
	2.2 格式化文件读写	2.2.1 能够使用 csv 模块、numpy 模块和 pandas 模块读写 csv 格式的文件; 2.2.2 能够使用 xml 模块、json 模块从 xml 和 json 格式文件中读取数据; 2.2.3 能够使用 xlrd、xlwt、pandas 模块读写 excel 文件数据。
	2.3 数据库数据获取	2.3.1 能够使用 pymysql 模块连接到 MySQL 数据库服务并进行数据操作;

工作领域	工作任务	职业技能要求
		2.3.2 能够编写 SQL 语句从多个表中查询给定条件的数据; 2.3.3 能够使用 SQL 聚合函数对数据进行求和、求平均值、求极值、计数等统计操作; 2.3.4 能够通过 SQL 语句执行插删改操作; 2.3.5 能够编写 SQL 语句向数据库中批量写入数据。
3. 数据检验处理	3.1 科学计算程序编制	3.1.1 能够选择合理的方式创建 numpy.array 和 pandas.DataFrame 来以存放结构化数据; 3.1.2 熟练使用 math 模块和 numpy 模块中的科学计算函数; 3.1.3 能够根据数据分布要求使用 random 和 numpy 模块生成随机数; 3.1.4 能够通过数据切片和条件筛选，获取 array 和 DataFrame 中的部分数据; 3.1.5 能够对 array 和 DataFrame 进行合并、拆分操作以产生需要的新数据集。
	3.2 数据统计与取样	3.2.1 能够分别以自定义函数和调用 numpy 和 pandas 库函数的方式获取数据的描述性统计，包括：计数、求和、均值、极值、百分位数、方差、标准差; 3.2.2 能够使用 matplotlib 模块绘制散点图、折线图、柱状图和饼图; 3.2.3 能够设置所绘制图形的尺寸、标签、图例、刻度和颜色等属性; 3.2.4 能够通过子图的方式在单张大图中整合多幅图像; 3.2.5 理解训练集、验证集和测试集的意义，并能够分别使用自定义函数和 sklearn 库函数，从原始数据集中随机划分子集。
	3.3 数据清洗与转换	3.3.1 能够分别使用 numpy 和 pandas 库函数识别数据中的缺失值; 3.3.2 能够分别使用删除法、平均值填补、临近值填补、众数填补法填充缺失值; 3.3.3 能够使用 pandas 库函数识别和处理数据中的重复值; 3.3.4 能够调用库函数实现类型转换，包括：字符串、数字和日期之间进行类型转换；array 和 DataFrame 中元素类型的转换。

表 2 数据应用开发与服务(Python)职业技能等级要求（中级）

工作领域	工作任务	职业技能要求
1. 数据应用程序系统开发	1.1 开发和运行环境搭建	1.1.1 能够在 Linux 上安装 Python 3.X 运行环境、VSCode 和 Jupyter Lab 开发环境; 1.1.2 能够熟练使用 Jupyter Lab 进行代码运行、文档编写、文件管理等工作; 1.1.3 能够根据要求部署基于 Django 的 Web 应用程序和数据库，确保其正常运行; 1.1.4 能够编制 Web 系统的安装、部署手册。
	1.2 Web 服务程序开发	1.2.1 能够使用 Django 和 djangorestframework 快速搭建 Web 服务应用程序框架; 1.2.2 能够在 Django 应用程序中针对数据库进行常规的 CRUD 数据操作; 1.2.3 能够分别以按需和 Singleton 模式创建模型对象实例，并将其推理功能以 RESTful 接口发布; 1.2.4 能够从 Python 客户端调用 RESTful 接口，传递正确的参数，接收并解析返回值。
	1.3 团队协作与任务管理	1.3.1 能够基于多方收集的需求信息，在团队内部讨论和确定系统需求并形成文档; 1.3.2 了解并能在实际工作中遵守国内外主要的信息安全和数据隐私保护法律法规，保证数据应用的合规性; 1.3.3 能够基于模板编制模块技术文档，包括：概要设计说明、详细设计说明、接口规范。
2. 多源数据采集	2.1 基于网络协议数据获取	2.1.1 能够使用多线程实现多任务并发，并采用恰当的方法对公共数据进行线程同步保护; 2.1.2 能够使用 re 模块和正则表达式查询和匹配文本; 2.1.3 能够使用 urllib 模块和 requests 对象通过 http 协议获取网页数据; 2.1.4 能够使用 urllib 模块通过 ftp 协议获取文件服务器的目录列表、下载文件数据。
	2.2 爬虫框架使用	2.2.1 能够正确安装、创建和配置基于 Scrapy 框架的爬虫应用程序; 2.2.2 能够正确解析 HTML 网页标签和内容; 2.2.3 能够实现数据翻页、自动链接跳转功能以爬取多个网页的数据; 2.2.4 能够将爬虫框架采集的数据保存到文本文件和 MySQL 数据库中。
	2.3 非结构化数据采集与存储	2.3.1 能够在单机环境下安装和配置 MongoDB、Redis 服务器; 2.3.2 能够使用 pymongo 模块连接到 MongoDB 服务

工作领域	工作任务	职业技能要求
		器并执行数据查询和修改操作; 2.3.3 能够使用 redis 模块连接到 MongoDB 服务器并执行数据查询和更新操作。
3. 数据预处理	3.1 数据统计与可视化展现	3.1.1 能够针对给定数据集进行方差分析：计算协方差、相关性; 3.1.2 能够通过 matplotlib 和 seaborn 库绘制箱图、误差图、直方图、热力图，以查看数据的统计信息和关联信息; 3.1.3 能够通过等值线图、3D 曲面图等绘制高维数据图像。
	3.2 数据清洗与转换	3.2.1 能够调用 sklearn 库函数对数据进行标准化和归一化处理; 3.2.2 能够调用 sklearn 和 pandas 库函数将文本标签转换成数值形式; 3.2.3 能够调用 sklearn 库函数将数值转换成 Onehot 编码; 3.2.4 能够分别调用 numpy、pandas 和 sklearn 库函数对数据进行分箱处理; 3.2.5 能够调用 sklearn 库函数生成 K 折交叉验证数据集。
4. 数据分析与建模	4.1 回归模型构建	4.1.1 能够使用 sklearn 模块的算法包构建和训练线性回归模型; 4.1.2 理解线性回归的评价方法，并能够采用合适的尺度评价模型的性能; 4.1.3 理解最小二乘法原理，并能够编写梯度下降算法程序求解线性模型的最优参数解; 4.1.4 能够以可视化的方式展现模型的拟合效果。
	4.2 分类模型构建	4.2.1 能够调用 sklearn 库函数构建和训练逻辑回归模型实现分类，并能够绘制分类决策边界线; 4.2.2 能够调用 sklearn 库函数构建和训练朴素贝叶斯模型实现文本分类; 4.2.3 能够调用 sklearn 库函数构建和训练 KNN 模型实现分类; 4.2.4 能够调用 sklearn 库函数构建和训练决策树模型实现分类; 并使用 graphviz 可视化查看决策树的决策过程; 4.2.5 理解分类模型的评价指标，包括：Accuracy、Precision、Recall、F1 Score，并能够使用 sklearn 库函数计算上述指标。
	4.3 聚类模型构建	4.3.1 理解并能够使用编程方式计算向量之间的距离，包括：曼哈顿距离、欧式距离、余弦夹角、汉明距离;

工作领域	工作任务	职业技能要求
		4.3.2 理解非监督学习与监督学习的区别，能够调用 sklearn 库函数构建和训练 K-Means 模型对数据进行聚类; 4.3.3 能够以可视化的方式展现数据的聚类效果。
	4.4 数据建模全流程应用	4.4.1 理解数据建模的通用流程和关键环节，并能按照流程执行建模任务; 4.4.2 能够根据建模目标，评估和选取恰当的回归、分类或聚类方法; 4.4.3 能够按照选定模型的建模要求，准备好必要的训练和测试样本数据; 4.4.4 完成模型训练，并能将模型通过 Web 服务对外发布; 4.4.5 能够编写客户端程序测试模型 Web 服务的正确性和可用性。

表 3 数据应用开发与服务(Python)职业技能等级要求（高级）

工作领域	工作任务	职业技能要求
1．数据应用系统开发与部署	1.1 系统部署与使用	1.1.1 能够在单机上搭建基础的大数据平台及常用组件，包括：HDFS、HBase、Hive、Flume、Kafka; 1.1.2 熟悉 HDFS、HBase、Hive 的基本命令操作，并能使用这些命令管理数据; 1.1.3 能够部署和维护基于 Flask 的 Web 服务应用程序。
	1.2 服务性应用程序开发	1.2.1 能够使用 Flask 创建 Web 应用程序，并访问 MySQL、MongoDB、Redis 中的数据; 1.2.2 能够定义 RESTful 函数，解析调用参数并以 json 格式返回结果; 1.2.3 能够在 Flask 应用程序中通过异步调用处理长时间任务。
	1.3 团队协作与项目管理	1.3.1 能够与用户及团队外部人员进行有效沟通和协作，充分获取需求及资源支持; 1.3.2 能够基于项目需求和限制，在团队内部合理分工，制定项目开发计划; 1.3.3 能够使用恰当的工具追踪项目开发进度并合理调整。
2．大数据采集	2.1 大数据采集	2.1.1 能够使用 hdfs 模块连接到 HDFS 服务，并进行文件和目录操作; 2.1.2 能够使用 happybase 模块连接到 HBase 服务，并进行表的创建、数据添加、数据查询等操作;

工作领域	工作任务	职业技能要求
		2.1.3 能够使用 pyhive 模块连接到 Hive 服务，并进行数据的查询操作。
	2.2 数据集成与传输	2.2.1 能够配置 flume 从指定数据源(文件系统、数据库等)获取数据; 2.2.2 能够使用 kafka-python 模块编写数据生产者程序，并与 flume 数据源对接; 2.2.3 能够使用 kafka-python 模块编写数据消费者程序。
3．数据预处理与特征选取	3.1 数据统计与取样	3.1.1 能够使用 Z-Score、IQR 等方法进行异常点/离群点检测; 3.1.2 能够使用 scipy.stats 库对目标数据进行正态性验证; 3.1.3 能够使用 pandas 对数据进行抽样，包括：等距抽样、分层抽样、整群抽样。
	3.2 数据可视化展现	3.2.1 能够使用 PyEcharts 绘制基本的数据图形; 3.2.2 能够使用 matplotlib 和 PyEcharts 库以动画形式展现时间变化数据; 3.2.3 能够使用 PyEcharts 以地图形式展现地理相关数据。
	3.3 特征选取	3.3.1 能够使用 Filter 方法进行特征筛选，包括:方差过滤法、卡方过滤法、F 检验法、互信息法; 3.3.2 能够使用 Wrapper 方法对数据集进行特征选取; 3.3.3 能够使用 Embedded 方法对数据集进行特征选取; 3.3.4 能够使用主成分分析法和线性判别法进行特征降维操作。
4．模型构建与优化	4.1 模型优化	4.1.1 理解线性回归的变体模型(Ridge 回归和 Lasso 回归)，并能够使用 sklearn 构建模型; 4.1.2 理解模型常用的超参数(例如: Learning Rate、L1/L2 等)，并能够通过交叉验证遴选最优超参数组合; 4.1.3 理解 ROC 和 AUC 的概念，能够计算并绘制 ROC 曲线，并通过 AUC 选取最优模型。
	4.2 集成学习模型使用	4.2.1 能够使用 Bagging 方式集成学习模型(决策树)实现分类和回归; 4.2.2 能够使用 Boosting 方式集成学习模型(Adaboost、GBDT、XGBoost)实现分类和回归; 4.2.3 理解 Stacking 方式的工作原理，能够使用该方法获取新的训练数据集并训练集成学习模型。
	4.3 高级模型构建与综合应用	4.3.1 理解关联规则的原理，并能够使用 APRIORI 模型构建模型;

工作领域	工作任务	职业技能要求
		4.3.2 理解协同过滤原理，并能够通过 Python 实现基于 User-Based 和 Item-Based 的个性化推荐模型; 4.3.3 理解 ARIMA 模型原理，并能够通过 Python 实现时间序列模型; 4.3.4 基于上述模型构建数据服务应用程序，对外提供推理和预测服务。

参考文献

[1] GB/T 5271.6-2000 信息技术 词汇 第6部分:数据的准备与处理

[2] GB/T 5271.31-2006 信息技术 词汇 第31部分：人工智能机器学习

[3] GB/T 35295-2017 信息技术 大数据 术语

[4] GB/T 37721-2019 信息技术 大数据分析系统功能要求

[5] 国家职业技能标准编制技术规程（2018年版）

[6] 中华人民共和国职业分类大典

[7] 中等职业学校专业目录

[8] 普通高等学校高等职业教育（专科）专业目录

[9] 普通高等学校本科专业目录

[10] 中等职业学校专业教学标准（试行）

[11] 高等职业学校专业教学标准（2018年）

[12] 本科专业类教学质量国家标准

[13] 教育部《关于职业院校专业人才培养方案制订与实施工作的指导意见》

[14] 职业学校专业（类）定岗实习标准

[15] 职业院校专业实训教学条件建设标准