人工智能数据处理
职业技能等级标准

（2021年1.0版）

科大讯飞股份有限公司　制定
2021 年 4 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：科大讯飞股份有限公司、南京信息工程大学、安徽信息工程学院、深圳职业技术学院、常州信息职业技术学院、重庆电子工程职业学院、深圳信息职业技术学院、淄博职业学院、武汉职业技术学院、襄阳职业技术学院、重庆工商职业学院、广东水利电力职业技术学院、湖北职业技术学院、贵州交通职业技术学院、徐州工业职业技术学院、广西经贸职业技术学院、广西理工职业技术学院、广东科学职业技术学院、安徽林业职业技术学院、广东理工职业学院、苏州信息职业技术学院、云南交通职业技术学院、湖南软件职业学院、苏州市职业大学、重庆财经学院、广西科技师范学院、广东财经大学、广东技术师范大学、广东开放大学、广西大学行健文理学院、广西外国语学院、南宁师范大学、南宁学院、四川大学锦江学院、淮阴工学院。

本标准主要起草人：陈涛、周佳峰、吴华安、李栋学、莫少林、吴有富、蔡铁、武春岭、胡方霞、魏本征、贺敏伟、曾文权、孙宾、张卫东、杨勇、钱银中、凌明胜、刘小华、李粤平、王宝成、胡昌杰、肖政宏、吴砥、刘晓、桂诚、丁德成、胡江院、崔小蕾、马季、雷大正、殷振华、张涛、张进兵、于俊、李雅洁、丁辉、程礼磊、陈小贝

1 范围

本标准规定了人工智能数据处理职业技能等级对应的工作领域、工作任务及职业技能要求。

本标准适用于人工智能数据处理职业技能培训、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。

国家、行业、团体有关标准如下:

GB/T 36625.1-2018 智慧城市 数据融合 第 1 部分：概念模型

GB/T 36625.2-2018 智慧城市 数据融合 第 2 部分：数据编码规范

GB/T 36339-2018 智能客服语义库技术要求

GB/T 37721-2019 信息技术 大数据分析系统功能要求

GB/T35295-2017 信息技术 大数据 术语

GB/T35589-2017 信息技术 大数据 技术参考模型

T/CESA1040—2019《信息技术 人工智能 面向机器学习的数据标注规程》

T/CESA 1039-2019　信息技术 人工智能 机器翻译能力等级评估

T/CESA 1034-2019　信息技术 人工智能 小样本机器学习样本量和算法要求

3 术语和定义

国家、行业标准界定的以及下列术语和定义适用于本文件。

3.1. 数据采集（data collection）

从数据源中选择和收集针对某种特定需要的数据。

3.2. 数据存储（data storage）

使用计算机或其他设备通过记录介质来保存数据。

3.3. 数据清洗（data cleaning）

检测和修正数据集合中错误数据项以及对数据进行平滑处理等操作的数据预处理过程。

3.4. 数据补全（data completion）

在含有遗失值的数据集上通过补全策略进行有效的数据填补过程。

3.5. 数据标注（data annotation）

对文本、图像、语音、视频等待标注数据进行归类、整理、编辑、纠错、标记和批注等操作，为待标注数据增加标签，生成满足机器学习训练要求的机器可读数据编码。

3.6. 数据分析（data analysis）

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；或对已存在的或计划的系统中的数据及其流程系统性的调查研究。

3.7. 数据建模（data modeling）

用于定义和分析数据的要求和其需要的相应支持的信息系统的过程，将分散多样的数据规则化、标准化，持续提升数据质量。

3.8. 数据治理（data governance）

组织中涉及数据使用的一整套管理行为，对数据进行处置、格式化和规范化的过程。

3.9. 特征工程（feature engineering）

利用领域知识和现有数据，确定哪些特征可能在机器学习训练模型中使用，然后将日志文件及其他来源的原始数据转换为所需特征的过程。

3.10. 数据降维（data dimensionality reduction）

在某些限定条件下，降低随机变量个数，得到一组“不相关”主变量的过程。

3.11. 特征学习（feature learning）

利用技术自动提取数据特征的过程，允许计算机学习使用特征的同时，也学习如何提取特征。

3.12. 数据生成（data generation）

根据实际业务数据分布规律来生成满足实际业务需要的数据，通过生成数据来模拟真实数据集在业务中的作用。

4 适用院校专业

中等职业学校：计算机应用、软件与信息服务、数字媒体技术应用、电子与信息技术、林业信息技术与管理、数字媒体技术应用、统计事务、电子商务等专业。

高等职业学校： 人工智能技术服务、计算机应用技术、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、计算机系统与维护、智能产品开发、智能控制技术、卫生信息管理、信息统计与分析、数字媒体应用技术、大数据技术与应用、商务数据分析与应用等专业。

应用型本科学校：人工智能、计算机应用工程、软件工程、计算机科学与技术、电子与计算机工程、计算机应用工程、智能科学与技术、智能控制技术、信息与计算科学、信息工程、信息管理与信息系统、数字媒体技术、数据计算及应用、数据科学与大数据技术、大数据管理与应用、大数据技术与应用等专业。

5 面向职业岗位（群）

主要面向人工智能、大数据、互联网、软件开发等IT类公司，以及政府机关、企事业单位的信息管理与服务部门，从事人工智能数据收集、处理、维护，人工智能数据建模、分析，人工智能数据治理、生成，人工智能算法应用等工作任务。面向的主要岗位包括数据标注员、人工智能数据分析师、人工智能数据训练师、数据建模工程师、人工智能算法工程师等。

6 职业技能要求

6.1 职业技能等级划分

人工智能应用服务开发职业技能等级分为三个等级：初级、中级、高级，依次递进，高级别涵盖低级别技能要求。

【人工智能数据处理】（初级）：主要面向岗位为在人工智能、大数据、互联网、软件开发等IT类相关公司中，以及政府机关、企事业单位的信息管理与服务部门，从事人工智能数据收集、人工智能数据处理、人工智能数据标注岗位。

可以对一些基本的结构化和半结构化数据进行基础数据库操作，完成结构化、机械化的数据采集、数据存储、数据清洗、数据补全、数据标注工作，可以使用excel等通用软件完成一些简单的数据分析和数据可视化等工作。

【人工智能数据处理】（中级）：主要面向岗位为在人工智能、大数据、互联网、软件开发等IT类相关公司中，以及政府机关、企事业单位的信息管理与服务部门，从事人工智能数据维护、人工智能数据建模、人工智能数据分析岗位。可对非结构化数据进行数据采集、数据清洗等工作，能够完成一定的数据智能分析及可视化、数据仓库、人工智能数据建模与数据治理以及特征工程等工作。

【人工智能数据处理】（高级）：主要在人工智能、大数据、互联网、软件开发等IT类相关公司中，以及政府机关、企事业单位的信息管理与服务部门，面向人工智能算法工程师、大数据工程师、系统架构师等岗位。可以辅助进行人工智能算法的应用，熟悉各类业务数据，可完成高阶完整的数据建模以及数据治理、完整的特征工程，为各类算法的成功落地完成特征工程、数据降维、数据生成等工作。

6.2 职业技能等级标准描述

表 1　人工智能数据处理职业技能等级要求（初级）

工作领域	工作任务	职业技能
1.数据获取与储存	1.1 基础软件功能使用与基本数据结构操作	1.1.1 能够在 Windows、Linux 系统上安装 Python 以及人工智能常用集成软件（Spyder，Pycharm，anaconda 等），并且能够使用 conda 或 pip 包管理工具，在命令行配置和管理需要的 Python 库。 1.1.2 能够完成相应软件编译环境的配置，能够在脚本和控制台两种模式下编译程序。 1.1.3 能够对基本的数据类型，例如字符串、整型、浮点型、布尔类型等，完成数据生成和类型转换等操作。 1.1.4 能够对基本数据结构，例如列表、

		元组、字典、字符串等，完成增删改查、数据存储等操作。 1.1.5 能够使用 Python 条件控制和循环控制实现逻辑处理功能。 1.1.6 能够使用 Python 模块调用和自定义的方法实现模块化设计。
	1.2 互联网数据常规工具获取	1.2.1 能够按需求在互联网中搜索并下载公开数据集。 1.2.2 能够熟悉 HTML 原理和 HTML 结构。 1.2.3 能够遵守网络爬虫相应的法律规制，使用爬虫技术实现文本数据、图片、音频和视频的爬取。 1.2.4 能够使用正则表达式、XPath、beautifulSoup 完成 HTML 文本解析。 1.2.5 能够通过模拟登陆的方式爬取需要登陆才能访问的页面数据，能够爬取 Ajax 技术传输的网站数据。 1.2.6 能够将数据持久化到 MongoDB、Redis 和 MySQL 等数据库中。
	1.3 数据存储常规工具使用	1.3.1 能够完成常用的数据库以及数据管理工具的安装配置。 1.3.2 能够使用基本的数据库语言完成数据的删除和存储。 1.3.3 能够将 Python 等编译工具与数据库连接，完成数据存储。 1.3.4 能够将获取的外部数据在数据库中存储。 1.3.5 能够使用数据存储工具，实现结构化数据、半结构化数据、非结构化数据的存储。
2.数据预处理与数据标注	2.1 数据预处理常规工具使用	2.1.1 能够完成数据预处理相关常用工具的安装以及调试使用；

		2.1.2 能够使用 sql 和 excel 完成数据基本的清洗、补全等操作。 2.1.3 能够使用常规工具读取多种存储类型的数据（csv，json，xlsx 等），并进行基础的数据预处理工作。 2.1.4 能够使用常见平台对数据进行基础的预处理工作。
	2.2 常规数据预处理编程	2.2.1 能够使用 Python 读取数据，能够使用 os、Numpy、Pandas 等模块实现文件存储路径的读取以及文件的读写。 2.2.2 能够对 Pandas 基本数据结构，例如 Series、Data Frame，以及 NumPy 的数组、矩阵等结构实现基本的操作。 2.2.3 能够使用 NumPy、Pandas 等模块进行数据的简单处理，包括数据的清洗补全和转换，以及分组和聚合。 2.2.4 能够使用 os、Pandas 完成文件操作（创建、复制、删除、读写、更改文件名等）。 2.2.5 能够使用 cv2,PIL 等模块完成图像数据的预处理操作（读取、保存、去重、裁剪、修改尺寸等）。 2.2.6 能够使用 wave 等模块完成语音数据的预处理操作（读取、播放、录音、清洗、加窗等）。
	2.3 数据标注	2.3.1 能够按标注规范和要求，使用人工智能标注平台对文本、图片数据进行标注。 2.3.2 能够根据特定的需求场景，通过人工智能标注平台制定标注模板以及标注任务。 2.3.3 能够对使用人工智能标注平台过

		程中产生的数据进行收集和分析。 2.3.4 能够整理、反馈数据标注质量并输出相应的报告。 2.3.5 能够使用 Python 等工具编写脚本实现不同类型的数据批量化标注。
3.基础数据分析与可视化	3.1 数学计算	3.1.1 能够使用常用工具计算数据的基本数字特征（总和、均值、众数、中位数等）。 3.1.2 能够使用常用工具计算数据的进阶数字特征（方差、标准差等）。 3.1.3 能够使用常用工具完成数据与数据之前的数学计算（差值等）
	3.2 基础数据分析	3.2.1 能够使用 excel 中的常用函数（SUM,AVERAGE 等）对使用 excel 存储的数据进行简单分析。 3.2.2 能够使用 NumPy、Pandas 等模块的常用函数（sum,max，mean，median 等)对使用 python 读取的数据进行简单分析。 3.2.3 能够使用 sql 中的常用聚合函数（MAX(),SUM()等)对使用数据库存储的数据进行简单分析。 3.2.4 能够对数值型数据的统计结果进行分析解释，说明统计结果反映数据的情况。
	3.3 数据可视化工具（平台）使用	3.3.1 能够将结构化的数据与图表进行连接，使用 excel、word、ppt 等常用工具或平台输出图表。 3.3.2 能够根据数据输出可视化图像，如柱状图、饼图、折线图、散点图、雷达图等。 3.3.3 能够对数据分析的结果选择合适

		的可视化形式。 3.3.4 能够在各类可视化图表中合理添加数据的数字特征。

表 2　人工智能数据处理职业技能等级要求（中级）

工作领域	工作任务	职业技能
1. 进阶数据分析可视化与数据仓库搭建使用	1.1 人工智能数据分析工具使用	1.1.1 能够针对多个业务的源数据进行价值信息的提取和进一步的分析挖掘。 1.1.2 能够根据业务场景和数据类型选择数据分析工具（R,sys,Python 等），并完成工具的安装以及调试使用。 1.1.3 能够使用人工智能统计类数据分析工具进行数据分析工作，包括数据的离散程度，异常值检测等。
	1.2 人工智能可视化工具使用	1.2.1 能够使用较为专业的可视化工具（R 语言的 ggplot2 或 Python 的 matplotlib 等）完成数据可视化工作。 1.2.2 能够使用可视化程序画出多种类型（直方图、3D 图像、等高线、各种条形图、动画等）展示数据集。 1.2.3 能够根据一些高阶的数据关系选择合适的图表类型展示，例如展示多项式函数，使用散点图展示数据集中的两个变量之间的关系。 1.2.4 能够根据语音数据绘制音频信号。 1.2.5 能够使用 Power BI 的 Power Query 实现数据清洗，能够使用 Power BI 的 Power Pivot 实现数据模型搭建，能够使用 Power BI 的 Power View 实现数据的可视化。

	1.3 数据仓库搭建使用	1.3.1 能够使用数据库进行数据的基本操作，包括增删改查等，能够说明数据库的存储和操作机理。 1.3.2 能够完成数据抽取、转换、加载等阶段等数据仓库搭建工作。 1.3.3 能够通过 SQL 编程对数据仓库进行访问和相关的操作。 1.3.4 能够使用通用工具（OLAP/Bi 等）对数据仓库进行取数并分析。 1.3.5 能够根据业务需求，对数据仓库的数据进行整理、分析等操作，为业务需求提供支撑。
2.基础数据建模与数据治理	2.1 人工智能数据建模工具使用	2.1.1 能够使用各类数据建模工具实现数据建模方法。 2.1.2 能够选择合适的变量或者重构变量来建立模型。 2.1.3 能够使用通用 sklearn 模块建立分类、回归、聚类等模型。 2.1.4 能够对模型效果进行评估。
	2.2 进行基础的数据治理工作	2.2.1 能够识别并构建数据资源目录，包括识别元数据、构建数据字典等。 2.2.2 能够基于数据资源目录对数据进行数据规整管理。 2.2.3 能够对数据进行数据质量管理等处理。
3.基础特征工程	3.1 基础特征处理	3.1.1 能够对数据标准化。 3.1.2 能够对数据归一化。 3.1.3 能够对定量特征进行离散化。 3.1.4 能够对定性特征哑编码。 3.1.5 能够对定性特征独热编码。 3.1.6 能对数据进行对数、指数变化。
	3.2 基础特征选择	3.2.1 能够使用 sklearn.feature_selection 中

		VarianceThreshold类移除低方差的特征。 3.2.2 能够使用sklearn.feature_selection中卡方（chi2）类检验定性自变量对定性因变量的相关性。 3.2.3 能够使用minepy包中Mine模块计算互信息和最大信息系数(MIC)来选择特征。 3.2.4 能够根据业务场景和数据类型使用特征选择工具包实现特征选择。

表 3　人工智能数据处理职业技能等级要求（高级）

工作领域	工作任务	职业技能
1.进阶数据建模与数据治理	1.1 进阶数据建模	1.1.1 能够完成深度学习开发环境（GPU加速模块 cuda、cudnn）以及深度学习开发框架（tensorflow、pytorch、keras等）的配置。 1.1.2 能够使用人工智能学习平台完成CNN、RNN 模型构建。 1.1.3 能够基于 Transformer 模型完成数据建模。 1.1.4 能够选择两个或以上的模型对比，并通过调整参数迭代优化模型效果。 1.1.5 能够进行模型部署及运行。
	1.2 数据治理体系规划	1.2.1 能够建立数据治理体系。 1.2.2 能够建立数据质量评估框架。 1.2.3 能够使用定性法、统计分析法、层次分析法进行质量评估。 1.2.4 能够使用人工智能工具治理元数据，编写并优化数据资源目录。
2.特征工程	2.1 进阶特征处理	2.1.1 能够实现连续特征离散化，离散特征的连续化。

		2.1.2 能够根据文本构建词袋模型、词频、文档频次、TF-IDF 特征。 2.1.3 能够使用人工智能学习平台构建文本 word2vec 特征。 2.1.4 能够用人工智能学习平台提取图像特征并进行数据增强。
	2.2 进阶特征选择	2.2.1 能够使用 sklearn.feature_selection 中 SelectFromModel 类结合 L1、Tree、stability 进行特征选择。 2.2.2 能够使用人工智能学习平台进行特征选择。 2.2.3 能够使用各类高阶的特征选择法完成特征选择，例如方差选择法、相关系数法、卡方检验、互信息法、基于惩罚项的特征选择法等。 2.2.4 能够根据业务场景和数据类型，对数据进行预处理、特征处理等操作后完成特征选择。
3.数据降维与数据生成	3.1 数据降维	3.1.1 能够使用 Scikit-learn decomposition 模块中主成分分析（PCA）算法实现数据降维。 3.1.2 能够使用 Scikit-learn decomposition 模块中线性判别式分析（LDA）算法实现数据降维。 3.1.3 能够根据业务场景和数据类型使用特征降维工具包实现特征降维。
	3.2 数据生成策略制定	3.2.1 能够设计数据生成策略。 3.2.2 能够根据数据生成策略设计合理的步骤，包括模型训练、特征工程、特征学习等结果。 3.2.3 能够根据业务场景和需求分析，综

		合条件来制定数据生成策略。
	3.3 数据生成工具使用	3.3.1 能够使用 sklearn 中的 dataset.make_regression 类进行回归数据生成。 3.3.2 能够使用 sklearn 中的 dataset.make_classification 类进行分类数据生成。 3.3.3 能够使用 sklearn 中的 datasets.make_blobs 类进行聚类数据生成。 3.3.4 能够使用 random 模块完成符合正态、伯努利、均匀分布等随机数据的生成。

参考文献

[1] 中等职业学校专业目录

[2] 普通高等学校高等职业教育（专科）专业目录

[3] 普通高等学校本科专业目录

[4] GB/T 36625.1-2018 智慧城市 数据融合 第 1 部分：概念模型

[5] GB/T 36625.2-2018 智慧城市 数据融合 第 2 部分：数据编码规范

[6] GB/T 36339-2018 智能客服语义库技术要求

[7] GB/T 37721-2019 信息技术 大数据分析系统功能要求

[8] GB/T35295-2017 信息技术 大数据 术语

[9] GB/T35589-2017 信息技术 大数据 技术参考模型

[10] T/CESA1040—2019《信息技术 人工智能 面向机器学习的数据标注规程》

[11] T/CESA 1039-2019 信息技术 人工智能 机器翻译能力等级评估

[12] T/CESA 1034-2019 信息技术 人工智能 小样本机器学习样本量和算法要求