第二章数据基础

人工智能的实现离不开数据，数据是人工智能 3 个要素之一。数据就像 AI 的“学习教材”——想教 AI 认猫狗，就得给它看无数猫狗照片（数据量要大）；照片得清晰正确（数据质量要高），不能把狗标成猫，还要有不同品种、角度（数据要多样）。AI 通过“做题”（用大部分数据训练）和“考试”（用剩余数据测试）来学习，错了就调整方法，就像小朋友改错题。没有数据，再聪明的算法也学不会任何技能，就像没课本的学生没法学习。简单说，AI 的聪明程度=好算法×足够多的好数据，就像学霸=好方法+大量练习。

2.1 数据获取与预处理

想象你需要做一份精致的菜肴（人工智能应用），同时已经拥有了顶级的食谱（算法模型），若没有新鲜优质的食材（数据）和精湛的刀工（预处理技术），最终成品也难以令人满意。因此，在人工智能领域，数据正是这样的核心要素——它的获取与处理将直接决定最终成品的质量，在整个人工智能开发中至关重要。

在人工智能领域，数据涵盖文本、语音、图像等多种类型。而将这些不同类型的数据转化为人工智能算法可识别的形式，是开展所有 AI 开发工作的首要步骤。

2.1.1 文本数据的数值表示

（1）名义尺度数据的含义与转换

名义尺度（Nominal Scale）是统计学中数据测量层次的最低级别，用于对事物进行**分类或命名**，其核心特点是数据仅表示不同的类别或属性，**类别之间无顺序、无数量关系，也无法进行数学运算**。

例 1：人口统计学中的性别信息与民族信息，前者一般只考虑男性与女性，后者取值如汉族、蒙古族等。这些数据类别之间无先后之分，无数量关系，也无法进行数学运算（如“男+女”无意义）。

1-of-N 转换（也称为**独热编码/One-Hot Encoding** 或**二元向量编码**）是机器学习和数据预处理中常用的一种技术，用于将**名义尺度（Nominal Scale）的分类变量**转换为**数值型向量**，以便算法能够处理。其核心思想是为每个类别创建一个独立的二进制（0 或 1）特征，从而消除类别之间的无序性，同时保留分类信息。

示例 1：颜色分类

名义尺度数据：绿色，黄色，蓝色（$N = 3$）

名义尺度	转换后的编码
绿色	001
黄色	010
蓝色	100

示例 2：性别分类

名义尺度数据：男，女（$N = 2$）

名义尺度	转换后的编码
男	10
女	01

示例 3：民族分类

名义尺度数据：汉族，蒙古族，…，珞巴族（$N = 56$）

名义尺度	转换后的编码

汉族	1000…0（56 位，第 1 位为 1）
蒙古族	0100…0（56 位，第 2 位为 1）
…	…
珞巴族	0000…1（56 位，第 56 位为 1）

（2）序数尺度数据的含义与转换

序数尺度（又称顺序尺度）是对事物之间等级或顺序差别的一种测度，用于对数据进行分类和排名，但数字本身不具有实际数量意义。

序数尺度具有以下特点：

（1）序数尺度具有等级性

序数尺度将数据划分为不同的类别，并赋予明确的顺序，如比赛名次（冠军、亚军、季军），教育程度（小学、初中、高中、本科、硕士及以上）

（2）序数尺度无固定间隔

在序数尺度中，相邻等级之间的差距不一定相等。例如“满意”与“非常满意”之间的差距，可能小于“不满意”与“满意”之间的差距，但具体数值无法量化。

（3）序数尺度没有绝对零点

序数尺度没有“无”或“零”的概念。例如教育程度中的“小学”不代表“零教育”

（4）序数尺度有一定的运算限制

序数尺度的数据往往仅支持中位数、众数等统计量，或非参数检验（如秩和检验），不支持均值、方差等需要等距或比率尺度的运算。

序数尺度的数字化是通过将具有顺序关系的定性数据转换为数值形式，以便进行统计分析和可视化展示的过程。其核心原则是保留原始数据的顺序关系，同时避免赋予数值间不合理的数学意义（如等距或比率关系）。几种常见的序数尺度数字化实例如下表所示：

表 1 序数尺度数字化示例

序数尺度数据	数字化示例
满意度调查	非常不满意→1，不满意→2，一般→3，满意→4，非常满意→5
收入等级	低（0~25%）→1，中低（25%-50%）→2，中高（50%-75%）→3，高（75%-100%）→4
疼痛程度	无痛→0，轻度→1，中度→3，重度→5（强调中重度疼痛的差异）

2.1.2 文本数据的预处理

1.文本处理的作用

文本语料在输送给模型前一般需要一系列的预处理工作, 才能符合模型输入的要求, 如: 将文本转化成模型需要的张量, 规范张量的尺寸等, 而且科学的文本预处理环节还将有效指导模型超参数的选择, 提升模型的评估指标。

2. 文本预处理的常用方法

（1）数据清洗

文本数据往往来源于语料库或者爬虫数据，其中往往含有大量无效信息，文本数据清洗是自然语言处理（NLP）中的关键预处理步骤，旨在消除原始文本中的噪声、冗余和错误，提升后续分析或模型训练的效果。因此，文本数据清洗的核心目标是将原始文本转化为结构化、规范化的高质量数据。其主要包含以下几个方面：

提升数据质量：消除文本数据中的拼写错误、乱码、重复内容等。

统一数据格式：涉及到编码、日期、数字等非文本字段的标准化。

降低噪声干扰：过滤与主题无关的符号、广告信息以及 HTML 标签等内容。

增强语义一致性：处理文本中的同义词、缩写以及多语言混合等问题

（2）分词

分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符。分词过程就是找到这样分界符的过程。

分词实例：

原文：“湖南师范大学创建于 1938 年，位于历史文化名城长沙，是国家“211 工程”重点建设的大学”

分词结果：“湖南师范大学/创建/于/1938/年/，/位于/历史/文化名城/长沙/，/是/国家/“211 工程”/重点/建设/的/大学”

（3）文本的向量化表示

文本的向量化表示（Text Vectorization）是将文本数据转换为数值向量的过程，以便机器学习模型（如神经网络、支持向量机等）能够处理。简单来说，向量化表示的目标是让电脑“看懂”文字。目前，主流的文本向量化主要有 One-Hot 编码、词袋模型以及词嵌入技术等。

One-Hot 编码：其核心思想是为每个类别创建一个二进制向量，其中只有一个位置为 1，其余为 0，从而避免类别之间的数值比较带来的误导。

词袋模型（Bag of Words, BoW）：词袋模型将文本视为一个“装满单词的袋子”，忽略单词的顺序和语法结构，仅统计每个单词在文本中出现的频率或存在性。

词嵌入技术：词嵌入是自然语言处理（NLP）中语言模型与表征学习技术的统称。概念上而言，它是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。其核心思想是通过向量空间中的位置关系捕捉单词的语义和语法信息

2.1.3 语音信号的数值表示

语音信号属于声音，而声音是一种波形，如图 1 所示：

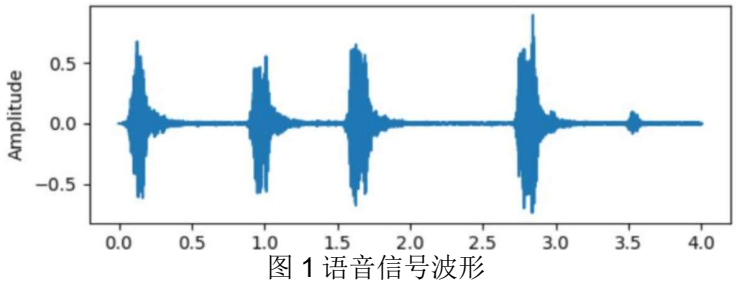

图 1 语音信号波形

在上图中，波形的横坐标是时间，纵坐标是波形的幅度，即声音的大小。

要实现 AI 中的语音识别需要进行以下几个步骤：

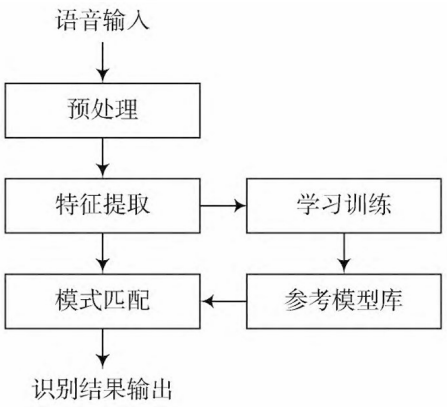

图 2 语音识别流程

当用户进行语音输入后，系统首先会对语音信号进行预处理，通过降噪、归一化、分帧等步骤，将原始信号转化为更适合后续分析的形式，是连接信号采集与高级处理（如识别、合成）的桥梁。

降噪：通过特定的算法消除背景噪声、电路干扰等无关信号，提升语音的清晰度和可懂度，为后续处理打下坚实的基础。

归一化：语音信号的归一化的作用是统一语音信号的幅度范围，消除因录音设备、距离等因素导致的音量差异，提高模型鲁棒性。

分帧：分帧就是把声音切开成一小段一小段，每小段称为一帧。方便后续处理中利用语音的短时平稳性（在短时间内特性稳定）进行分析。

预处理结束后，语音信号就可以送入 AI 模型，进行特征提取、模式匹配等操作，最终，系统将匹配结果作为识别结果输出，完成从语音到文本的转换，为用户提供可读的识别信息。

2.1.4 图像数据的数值表示

图像数据的数值表示是计算机处理和存储图像的基础，它涉及如何将连续的视觉信号（如亮度、颜色）转换为离散的数字形式。这一过程涵盖采样、量化、色彩空间转换等多个步骤，最终生成可被计算机处理的数值矩阵。

1.三基色原理

三基色原理是色彩科学中的核心理论，它揭示了人眼如何感知颜色以及如何通过有限的基本颜色混合出自然界中绝大多数色彩的机制。

在三基色原理中，认为人眼视网膜存在三种感光细胞：分别对红（Red）、绿 Green）、蓝（Blue）三种波长的光最敏感（即 L、M、S 视锥细胞）。任何颜色均可通过调整红、绿、蓝三种基色的强度按比例混合得到。

基于三基色原理，人们提出颜色的数学表达式，即任何颜色C，均可表示为三基色的线性组合：

$$C = R \cdot \boldsymbol{r} + G \cdot \boldsymbol{g} + B \cdot \boldsymbol{b}$$

其中，R, G, B为基色的强度系数，$\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}$为基色的单位向量（通常对应标准红、绿、蓝颜色）。

因此，在显示终端（如显示器，手机）中，可以通过控制红、绿、蓝三色荧光粉或 LED 的亮度，组合出数百万种颜色。

2.图像数据的分类

根据颜色通道和数值类型，图像数据可分为以下类型

（1）二值图像

二值图像是一种图像类型，其中每个像素只有两个可能的值，通常表示为 0（黑色）和 255（白色）。二值图像在机器视觉、机器人视觉和计算机视觉中广泛应用，因为它们便于进行轮廓分析、对象测量、形态学处理和特征提取等操作。
因为二值图每个像素只有两个值，所以每个像素只需要 1bit 即可存储。

（2）灰度图像

由于二值图只有 1bit 存储空间，表示的细节有限，因此人们便引入灰度图。灰度图是在二值图的基础上使用 8bit 来存储每一个像素的值，所以其范围在 0~255 之间。不同于二值图，灰度图的颜色信息更加丰富，不仅仅只有白色和黑色，还有各种灰色，共有 256 个灰度值。

（3）彩色图像

灰度图是单通道图形，即每个像素只有 1 个分量，而基于三基色原理，人们通过 RGB 模型来表示彩色图像。在 RGB 模型中，用红（R）、绿（G）、蓝（B）三原色的强度组合表示颜色。每个通道取值范围为 0-255（8 位），可组合约 1677 万种颜色（24 位真彩色）。

表 2 RGB 模型颜色表示示例

颜色名称	RGB 表示	说明
纯红色	(255,0,0)	红色通道亮度最大(255)，其余通道为 0。
纯绿色	(0,255,0)	绿色通道亮度最大
纯蓝色	(0,0,255)	蓝色通道亮度最大
黄色	(128,0,128)	红绿色通道亮度为中等，混合后为黄色
紫色	(128,0,128)	红蓝色通道亮度为中等，混合后为紫色
白色	(255,255,255)	三个通道亮度均最大，混合后为白色
黑色	(0,0,0)	三个通道亮度均为 0，无光混合，显示为黑色

思考题

1．数据在人工智能中的作用是什么？请结合“AI 的聪明程度=好算法×足够多的好数据”这一观点进行阐述。
2．名义尺度数据和序数尺度数据的主要区别是什么？请分别给出两种数据类型的例子并说明其转换方法。
3.请详细阐述三基色原理，并解释它是如何在 RGB 彩色图像表示中应用的。
4.在图像处理中，二值图像、灰度图像和彩色图像的主要区别是什么？请分别给出它们的应用场景，并解释为什么在这些场景中会选择这种图像类型。

2.4 科学计算

科学计算在人工智能中扮演着核心支撑角色，涉及 AI 的数据构建、算法实现、模型优化、性能评估等各个方面，其本质是将人工智能问题转化为数学问题，通过数值方法寻求近似解的过程，其精度、效率与稳定性直接决定了 AI 系统的性能上限与应用价值。

为提高人工智能领域的开发效率，人们提供了各种各样的科学计算库，这些计算库发挥着十分关键的作用：它们能够为原始数据的预处理提供高效的工具，用于支持数据清洗、格式转换、特征工程等基础操作，确保输入数据的质量与适用性；在模型构建阶段，这些计算库通过封装数学运算、优化算法和机器学习框架，降低了复杂模型（如神经网络）的实现门槛；在训练过程中，它们又能够提供自动微分、并行计算和硬件加速能力，显著提升大规模数据集的训练效率；对于模型评估与优化，库中的统计分析和可视化工具能帮助开发者分析模型性能，并指导超参数调优；此外，科学计算库还支持跨领域任务，如自然语言处理中的文本向量化、计

算机视觉中的图像特征提取，以及强化学习中的环境建模与策略优化等，为 AI 开发人员提供了极大的便利。目前，主流的科学计算库主要包括 Numpy、pandas、Matplotlib 以及 Scikit-learn 等。

2.4.1 Numpy 库

NumPy（Numerical Python）是 Python 的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比 Python 自身的嵌套列表结构要高效得多，支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库，数组数据在内存中连续存储，极大提升了访问和计算速度。

图 3 NumPy 标志

Numpy 计算库的主要功能包含以下几个方面：
1）提供了一个强大的 N 维数组对象 NDArray，支持任意维度的数组（如向量、矩阵、张量），且所有元素类型相同（如整数、浮点数）。
2）具有比较成熟的数学运算工具，支持线性代数以及统计函数的各种运算。
3）支持广播机制，允许不同形状的数组进行算术运算，自动扩展较小数组的维度以匹配较大数组，从而简化代码。

Numpy 底层使用 C 语言进行实现，执行速度远超纯 Python 循环。部分操作可以利用多核 CPU 或 GPU 进行加速，实现并行计算。同时数学运算避免了 Python 列表的冗余开销，适合处理大规模数据。在数据预处理、图像处理、机器学习等领域有着广泛的应用。

2.4.2 pandas 库

pandas 是基于 NumPy 的一种工具，该工具是为解决数据分析任务而创建的，它有着强大的数据分析能力和处理工具。“pandas”的名称源自术语“panel data”，用于描述包含多个时间段观察结果的数据集，强调其专注于处理现实世界数据集的多样化数据结构。pandas 支持数据的增、删、改、查；支持时间序列分析功能；支持灵活处理缺失数据；具有丰富的数据处理函数；具有快速、灵活、富有表现力的数据结构：DataFrame 数据框和 Series 系列。

图 4 pandas 图标

2.4.3 Matpllotlib

Matplotlib 是 Python 中最基础且功能强大的数据可视化库之一，由 John D. Hunter 于 2003 年创建，旨在为科学计算提供类似 MATLAB 的绘图接口。它能够生成静态、交互式或动态的 2D/3D 图表，广泛应用于数据分析、学术研究和工程领域。

图 5 matplotlib 图标

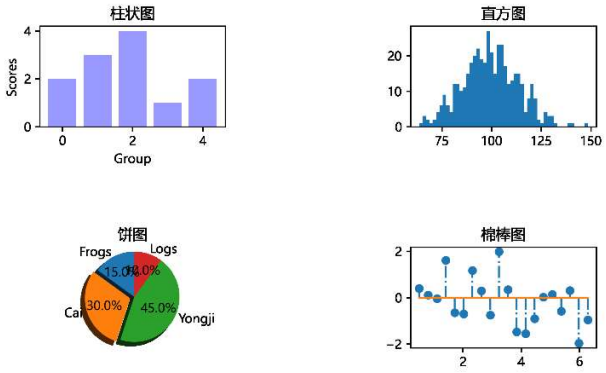

图 6Matlplotlib 绘制的图形示例

2.4.4 Scikit-learn

Scikit-learn（简称 sklearn）是 Python 中最流行、最成熟的机器学习库之一，专注于提供简单高效的工具用于数据挖掘和数据分析。它基于 NumPy、SciPy 和 Matplotlib 构建，支持多种经典机器学习算法，并以其一致的 API 设计和优秀的文档广受开发者欢迎。

图 7 scikit-learn 图标