中国刑事警察学院硕士研究生招生考试

《数据科学基础》考试大纲

（2023 年 12 月）

Ⅰ.考查目标

要求考生具有准确把握数据科学基础知识和基础理论的专业素质，具备分析、判断和解决国家安全警务问题的基本能力。具体包括：

1. 全面掌握数据科学基础知识及其内涵，掌握重要概念。

2. 正确理解数据科学基础理论，把握数据科学基础的科学思想方法与立场。

3. 了解网络爬虫的基本原理及数据获取方法，掌握关键信息的提取方法。

4.了解人工智能基本概念及工作原理，掌握人工智能常用分类与回归、聚类、神经网络、深度学习等方法的原理与使用场景。

Ⅱ.考试形式和试卷结构

一、考试满分及考试时间

试卷满分为 80 分；考试时间为 90 分钟。

二、答题方式

答题方式为闭卷、笔试。

三、试卷题型结构

1. 名词解释，共 20 分。

2. 简答题，共 40 分。

3. 论述题，共 20 分。

Ⅲ.考查内容

第一部分 网络爬虫与信息提取

一、爬虫基本原理及数据获取

1. 爬虫基本流程

2. 数据获取

3. 网络爬虫引发的问题及限制

4. 应用案例分析

二、信息解析

1. BeautifulSoup 库的基本知识

2. BeautifulSoup 的遍历方法

3. 基于 BeautifulSoup 库的 HTML 内容查找方法

4. 应用案例分析

三、关键信息提取

1. 正则表达式的构成

2. 正则表达式的使用

3. 应用案例分析

第二部分 人工智能与机器学习

一、人工智能基础

1. 机器学习的基础

2. scikit-learn 简介

二、分类与回归

1. k 近邻分类

2. 线性回归

3. 支持向量机

4. 决策树

5. 随机森林

三、聚类

1. 聚类性能度量

2. k 均值聚类

四、神经网络与深度学习

1. 感知机

2. 前馈神经网络

3. 深度学习

Ⅳ.参考书目

1. 杨志强 王睿智 肖杨 孙丽君 李湘梅 丛培盛.数据科学基础[M].高等教育出版社, 2022

Ⅴ.参考试题举例（非完整试题，仅为样式与分值说明）

数据科学基础

一、名词解释（共 20 分）

1. 机器学习

二、简答题（共 40 分）

1. scikit-learn 评估器 API 的使用步骤。

三、论述题（共 20 分）

1. 论述网络爬虫的基本流程。

Ⅵ.参 考 答 案

数据科学基础

一、名词解释

1. 机器学习

答：广义上说，机器学习是一种能够赋予机器进行学习的能力，使它不断改善自身。机器学习主要通过计算手段，利用经验来改善系统性能。在计算机系统中，“经验”即为“数据”。机器学习利用“数据”训练出“模型”的算法，然后使用模型进行测。由此可见，数据对于机器学习的意义，可以说数据是原材料，机器学习是加工工具，模型是产品。人们可以利用这个产品做预测，指导未来的行动。

二、简答题

1.scikit-learn 评估器 API 的使用步骤。

答：（1）整理数据，获取特征矩阵和目标数组。

（2）通过从 scikit-learn 中导入适当的评估器类，选择模型类。

（3）设置模型超参数，初始化模型。

（4）调用模型实例的 fit 方法，拟合数据。

（5）应用模型，预测新数据的标签。

三、论述题

1.论述网络爬虫的基本流程。

答：（1）发起请求

向目标站点发起请求，同时包含额外的配置信息，然后等待服务器响应。这个请求的过程就如同在浏览器地址栏中输入网址并单击 Enter 键，将浏览器作为客户端向务器发送了一次请求。

（2）获取响应内容

如果服务器能正常响应，客户端会得到一个响应，即获取到的内容，类型可能是 HTML、Json 字符串，二进制数据(图片、视频等)等。此步骤相当于服务器接收了客户端的请求，并发送网页 HTML 文件给浏览器。

（3）解析内容

如果得到的内容是 HTML，可以使用正则表达式、网页解析库进行内容解析以便提取感兴趣的信息。如果是 Json，可以直接转为 Json 对象解析；如果是二进制数据，可以保存或做进一步处理。这一步相当于浏览器把服务器的文件获取到本地后再进行解释和展现。

（4）保存数据

保存的方式多样，可以把数据存为文本，特定的 JPEG、MP4 等格式文件，也可以把数据保存到数据库中。此步骤相当于用户在浏览网页时下载网页中的图片、视频或其他数据。