数据科学与工程导论

教学大纲

课程名称（中文）：数据科学与工程导论

课程名称（英文）：Introduction to Data Science and Engineering

课程性质：专业核心课程

建议学分：3

理论学时：32

实践学时：32

授课对象：数据科学与大数据技术专业学生

先修课程：高等数学、线性代数、程序设计

一、课程简介（中文）

我们正处于一个快速发展的信息化时代，数据极大地影响着我们的生活。数据科学与工程是研究探索网络空间中数据自然界奥秘的理论、方法和技术，研究的对象是数据自然界，研究认识数据的各种类型、状态、属性及变化形式和变化规律，其目的在于揭示自然界和人类行为现象和规律。数据科学与工程导论是为数据科学与大数据技术专业本科生开设的专业基础课，该课程将系统性地讲述与数据科学与工程相关的各方面知识。

本课程定位为数据科学与工程的入门课程，为学生搭建起通向“数据科学与工程知识空间”的桥梁和纽带。课程将系统梳理总结数据科学与工程相关原理、技术和实践案例，帮助学生形成对数据科学与工程知识体系及其应用领域的轮廓性认知，为学生在数据科学与工程领域“深耕细作”奠定基础、指明方向，最终形成数据思维。

数据科学与工程导论的主要内容是以数据为中心，通过计算思维与数据思维的方法，来理解我们所处的世界（科学），以及对现实问题的求解（工程）。其最重要的思维方式是数据思维，简单来说就是以数据为中心的问题求解。

这些内容通过五条线贯穿起来：

第一部分关注数据思维：第一单元绪论介绍了信息文明与数据简史、数据科学与工程的基本内涵、第四范式、以及数据科学的应用；第二单元数据思维与问题求解介绍了数据思维与数据科学过程、计算思维与分析思维、像数据科学家一样思考、数据驱动的问题求解实例。

第二部分关注数据、计算与基础设施：第三单元介绍了比特与数据、数据的二进制表示、数据的模型、数据的结构；第四单元介绍了数据算法、算法分析与局限性、数据结构与算法的关系、计算机编程语言；第 5 单元计算基础设施介绍了通用机器的思想、程序是如何执行的、计算机系统结构、云计算与数据中心；第 6 单元数据的全生命周期管理介绍了数据采集、数据存储、数据管理、数据计算、数据分析、数据展示；第 7 单元数据库系统介绍了数据库的起源与发展、关系数据库、数据仓库与 OLAP、数据管理技术新格局、结构化查询语言 SQL；第 8 单元大数据系统介绍了大数据的基本概念、Hadoop 和 Spark 生态、SQL 与 Hadoop 的组合、大数据系统实例。

第三部分关注分析方法：第 9 单元数据科学过程介绍了数据科学过程基础、数据科学工作流；第 10 单元统计分析原理介绍了数据科学的数学基础、概率统计基础、统计建模；第 11 单元机器学习方法介绍了机器学习的发展历史、机器学习的方法、机器学习的最新发展；第 12 单元深度学习介绍了深度学习的基本概念、深度学习的拓展、深度学习的应用、深度学习的工具；第 13 单元数据挖掘基础介绍了数据挖掘的概念、数据挖掘标准流程、数据挖掘的技术、大数据挖掘；第 14 单元非结构化数据挖掘介绍了自然语言处理、语音信号分析、图像处理与理解。

第四部分关注数据应用与社会问题：第 15 单元数据应用综合应用介绍了数据科学与工程在智慧城市、智能运维、机器视觉中的综合应用实践；第 16 章数据道德与职业行为准则介绍开放的世界、职业规划、数据隐私与社会问题。

最后是开源实践：每个单元中，我们均选区了主流的开源编程语言与软件工具，指导大家充分的在数据上进行实践，主要包括：Python 语言、SQL 语言、KNIME 工具等。

本课程有两项难度系数。普通难度课程为 2 学分，课堂学时 32 个，着重介绍数据科学与工程的主体领域。进阶难度课程为 3 学分，课堂学时 48 个，主要包括一些高级与前沿的课题。

二、课程目标

该课程旨在帮助学生了解如何在大数据时代背景下运用各门与数据相关的技术和理论来服务社会，着重培养学生成为数据工程师所需要的技能与思维。通过对该课程的学习，使学生掌握如何利用计算工具，包括简单的统计、机器学习和数据可视化工具，用以模型化数据和理解数据，在从模糊的问题陈述到解决问题的计算表述的过程中，培养研究能力。了解一系列有用的算法和技巧。培养学生运用统计分析、机器学习、分布式处理等技术，从大量数据中提取对科学研究和生产实践有意义的信息，以可视化等技术通过通俗易懂的形式传达出来的能力。

目标 1：了解数据专业全貌，建立数据思维的意识；

目标 2：掌握数据科学与工程的基本内涵和应用模式；

目标 3：培养以数据为中心的问题求解能力，系统性的学习数据科学与工程的核心原理与关键技术；

目标 4：培养开源开放的精神，建立基于开源工具的数据分析与处理意识，并做到初步的数据编程训练；

目标 5：让大家感受到数据与计算的美，数据与计算的愉悦；

目标 6：点燃大家对数据专业的热情与兴趣。

三、课堂教学内容和学时分配

大类知识点	小类知识点	建议学时分配
1、数据科学与工程的基本概念	信息文明与数据简史 数据科学与工程的基本内涵 第四范式：数据密集型科学 数据科学与工程的应用案例	2
2、数据思维与问题求解	问题求解的概念 思维方式的概念 计算思维概念与方法 数据思维概念与方法 数据驱动的问题求解实例	2
3、数据的模型与结构	比特与数据 数据的二进制表示 数据的模型的概念 数据的结构的概念	2
4、数据的计算与程序表达	数据的计算：算法 算法分析与局限性 数据结构与算法的关系 计算机编程语言	2
5、计算基础设施	通用机器的思想 程序是如何执行的 计算机系统结构 云计算与数据中心	2
6、数据的全生命周期管理（系统角度）	数据采集 数据存储 数据管理 数据计算 数据分析 数据展示	2
7、数据库系统	数据库的起源与发展 关系数据库 数据仓库与 OLAP SQL 与 NoSQL 语言 数据管理新技术	2

8、大数据系统	大数据的基本概念 Hadoop 和 Spark 生态 大数据系统实例 SQL on Hadoop	2
9、数据科学过程 （工程角度）	数据获取过程 数据预处理过程 探索性分析过程 数据建模过程 数据可视化过程	2
10、统计分析原理	数据科学的数学基础 概率与统计基础 数据分析的工具 统计建模：线性回归模型	2
11、机器学习方法	机器学习简史 机器学习的原理与方法 典型的机器学习算法介绍 机器学习的最新发展	2
12、深度学习	深度学习的原理 误差逆传播 卷积神经网络 深度学习的工具	2
13、数据挖掘基础	数据挖掘的通用模式 多维数据挖掘 多媒体异构数据挖掘 大数据挖掘	2
14、非结构化数据挖掘	自然语言分析 Web 信息提取 图像数据处理	2
15、数据综合应用	智慧城市实践 智能运维实践 机器视觉实践	2
16、数据道德与职业	开放的世界 数据隐私与社会问题 数据伦理 数据专业的职业规划	2
	合计	32

四、实践教学内容和要求

实践内容	所需知识点	实践要求	学时分配
1、Git 与 Python 基础	1	能用用 Git 进行代码与数据管理，能正确编写并运行 Python 程序	2
2、Python 问题求解	2	能用 Python 解决简单的编程问题，熟悉并掌握 Python 的基本结构	2
3、Python 算法与数据结构基础	3	能用典型的 Python 数据结构进行编程，能初步调试 Python 代码并修改 bug	2
4、Python 程序性能评测	4	能对 Python 的代码性能进行分析*	2
5、Python 数据采集与存储*	5	能用 Python 进行基本的数据采集，例如爬虫 能用 Python 进行简单的数据分析与可视化	2
6、SQL 数据处理与分析	6	能用 SQL 进行基本的数据操作	2
7、NoSQL 数据处理与分析*	6	能用 NoSQL 进行基本的数据操作	2
8、基于 Python 的 MapReduce 数据处理*	7	能用 Python 和 MapReduce 进行基本的数据处理	2
9、Python 数据科学过程	8	能用 Python 进行完整的数据科学过程分析	2

10、Python 统计分析	9	能用 Python 完成典型的统计分析任务	2
11、Python 机器学习*	10	会用 Python 进行典型的机器学习方法的调用与数据分析	2
12、Python 深度学习*	11	能用 Python 处理简单的多维数据、文本、语音与图像数据	2
13、Python 结构化数据挖掘	12	能用 Python 处理简单的多维数据、文本、语音与图像数据	2
14、Python 非结构化数据挖掘*	13	能用 Python 处理简单的多维数据、文本、语音与图像数据	2
15、数据科学与工程综合项目实践	14	能综合运用本课程中学到的方法和工具完成一个实际数据作品	2
		完成一个相对复杂的实际数据作品	2
合计			**32**

五、教材、参考书目和学习材料建议

1. 王伟，《数据科学与工程导论》，华东师范大学出版社，2020。