

# 《大数据分析技术》理论教学大纲

## (Big Data Analysis Techniques)

课程代码：0600084

总学时：48 学时（其中：讲课 16 学时、实验 32 学时、讨论 0 学时）

先修课程：《线性代数》、《高等数学》、《概率论》、《统计学》、《数据科学与大数据技术导论》、《C 语言程序设计》、《python 程序设计》，等。

### 一、课程性质

《大数据分析技术》是数据科学与大数据技术专业的必修课程、主干课程、核心课程，是大数据采集、存储、分析与应用中最重要环节“分析”的具体技术化课程，是数学、统计学和计算机技术相结合的计算机类课程。课程包含数据分析理论教学，更主要的是充分应用 excel 和 python 等工具挖掘数据价值的实践性教学。

### 二、课程目的

从宏观和微观两个维度看本课程的目的。

宏观上让学生了解“数据是资源”的含义，了解国家的大数据战略，培养学生“万物皆数据，万物皆产生数据”的宏观数据理念和思维能力。

微观上使学生掌握数据分析经典技术和深度学习的基本思想和方法，能应用数学模型的方法和数据实验的手段来研究数据和挖掘数据的价值，并能应用统计学方法对所建模型进行评估，最终将所挖掘的数据价值应用于生产实践和人工智能等领域使得生产力的发展产生质的飞跃。

### 三、课程任务

《大数据分析技术》是一门实验科学，从数据出发，借助 excel 和 python 等软件，从实验中学习、探索和发现数据规律，体验数据价值发现的欢乐和挫折。

《大数据分析技术》的主要任务有：

（1）介绍数据预处理的方法与步骤，包括数据清洗、数据集成、数据变换和数据规约，定性变量的赋值方法和定量变量的离散化，等；细致说来，主要的具体的任务是数据清洗中缺失数据的填充，数据集成中多源、多类型数据的集成，数据变换中数据的标准化和归一化，数据规约中属性规约和维数规约。

（2）讲述经典数据分析技术及其应用，包括插值、拟合、回归、聚类、模式识别、决策树和随机森林等；

（3）讲述基于 python 语言和 TensorFlow 平台的深度学习方法及其应用，包括经典神经网络、卷积神经网络和循环神经网络。

### 四、教学内容

## 1. 数据预处理

这是一个数据分析中内容相当丰富的模块，包括：

- (1) 数据清洗—数据的脏信息，脏信息的清洗方法与步骤；
- (2) 数据集成—多源数据、多类型数据集成在具有一致性的数据仓库中；
- (3) 数据变换—数据标准化和归一化；
- (4) 数据规约—属性规约、维数规约，等；
- (5) 定性变量赋值方法，定量变量的离散化，等。

## 2. 经典数据分析技术

- (1) 插值与拟合；
- (2) 回归分析：主要包括线性回归分析、非线性回归、logistic 回归分析；
- (3) 聚类分析：FCM、k-means++、KNN、SVM，等；
- (4) 决策树和随机森林：提供最优解决方案和挖掘特征重要性。

## 3. 深度学习

- (1) TensorFlow 平台
- (2) 神经网络模型
- (3) 卷积神经网络
- (4) 循环神经网络

# 五、基本要求

## 1. 知识要求

(1) 理解数据预处理的观念，熟练地通过 excel 和 python 对数据进行预处理，特别是熟练掌握数据填充方法、数据标准化和归一化方法、奇异值识别方法，等。

(2) 了解机器学习的概念，并理解常用的经典的机器学习方法。熟练掌握经典的数据分析技术，特别是能熟练地将数据拟合和回归分析等技术应用于因果推断，将 FCM、k-means++、KNN 和 SVM 等技术应用于聚类分析，将决策树和随机森林应用于问题最优解决方案的搜寻和特征重要性的挖掘。

(3) 理解深度学习的概念，理解神经网络模型的概念，熟练掌握应用基于 python 语言和 TensorFlow 平台的深度学习挖掘大数据价值的方法和步骤。

## 2. 能力要求

- (1) 数学建模能力和编程能力。这是数据分析所必需俱备的基本能力。
- (2) 应用统计能力。应用统计学参数估计、假设检验等方法对模型进行评估。

(3) 注重研究性质。《大数据技术》充分拓展学生的思维，独立提出自己的想法（比如，基于数据建立什么模型，是学生独立行为），然后在计算机上实验，通过细致的观察和深入的分析，得到逼近事实、较为正确的结论。

(4) 独立实验，强调探索过程。自己设计和动手，提出自己的关于数据建模的猜测并对自己的猜测进行评估，找出支持论据，从实验中学学习、探索和发现数据价值及其规律，体验数据价值发现的挫折和欢乐。

(5) 写作能力，独立撰写实验报告的能力。

## 六、教学内容及学时分配

数据科学与大数据技术专业是新兴专业，尚在在摸索中建设。本课程在参考众多有价值的文献的基础上采用自编讲义作为教材。

本课程是一门以实验为主的课程，其主要内容可分为两部分：第一部分是经典数据分析技术，围绕因果推断和聚类分析这两大基本内容，让学生充分应用计算机软件的计算和可视化功能，展示基本概念和结论，体验如何发现、总结和应用数据价值及其规律。第二部分是深度学习，以经典神经网络为基础向高端提升、为中心向边缘科学发散，涉及卷积神经网络、循环神经网络（以 LSTM 为代表），等等。

教学内容			教学要求	重点(☆)	难点(Δ)	学时安排	备注
数据预处理	1	数据清洗	B			9	
	2	数据集成	C				
	3	数据变换	A	☆			
	4	数据归约	C		Δ		
	5	定性变量赋值和定量变量离散化	B	☆	Δ		
插值与拟合	1	插值	B			3	
	2	拟合	B	☆	Δ		
回归分析	1	一元线性回归分析	A	☆		9	
	2	多元线性回归分析	A	☆	Δ		
	3	一元非线性回归分析	A	☆			
	4	多元非线性回归分析	C		Δ		
	5	Logistic 回归	A	☆	Δ		
决策树和随机森林	1	决策树	C		Δ	6	
	2	随机森林	A	☆	Δ		
聚类分析	1	聚类分析导论	C			9	
	2	模糊 C 聚类分析 (FCM)	B				
	3	K 均值聚类分析 (k-mean++)	B				
	4	K 最邻近聚类分析 (KNN)	B	☆			

	5	支持向量机（SVM）	A	☆	Δ		
深度学习	1	TensorFlow 平台简介	A	☆		12	
	2	经典神经网络模型	B	☆			
	3	卷积神经网络（CNN）	A	☆			
	4	循环神经网络（RNN）	A	☆	Δ		
合 计			48				

(教学要求: A—熟练掌握; B—掌握; C—了解)

## 七、建议实验项目及学时分配

序号	实验项目名称	内容提要	学时	实验属性	开出要求
1	实验 1：数据预处理 1-大型百货商场会员消费分析	数据清洗，数据集成	2	综合	必做
2	实验 2：数据预处理 2	数据变换，数据归约	2	综合	必做
3	实验 3：商务数据分析	应用实验 1 和实验 2 预处理好的数据进行商务数据分析	2	验证	必做
4	实验 4：水箱水流量问题	应用插值解决水箱水流量问题	2	验证	选做
5	实验 5：工件轮廓线问题	应用拟合解决工件轮廓线问题	2	综合	必做
6	实验 6：税收与 GDP 关系探究	应用一元线性回归模型解决税收与 GDP 关系问题	2	综合	必做
7	实验 7：产品销量与广告媒体投入之间关系分析	应用多元线性回归模型进行产品销量与广告媒体投入之间关系分析	2	综合	必做
8	实验 8：中国新冠疫情趋势分析	应用一元非线性回归模型研究我国新冠疫情发展趋势	2	综合	必做
9	实验 9：口罩销量影响因素研究	应用多元非线性回归模型进行口罩销量影响因素研究	2	综合	选择
10	实验 10：信贷影响因素研究	应用决策树、随机森林方法研究信贷影响因素	2	综合	必做
11	实验 11：我国城市发展水平分析	应用 FCM、k-mean++或 KNN 方法进行我国城市发展水平分析	2	综合	必做
12	实验 12：空气质量评估研究	应用 SVM 方法进行空气质量评估研究	2	综合	必做
13	实验 13：鸢尾花的识别	应用经典神经网络对鸢尾花进行聚类分析	2	综合	必做
14	实验 14：图像识别	应用卷积神经网络实现图像识别	3	综合	必做
15	实验 15：上证指数预测	应用长短时记忆模型实现上证指数预测	3	综合	必做
合计		32			

## 八、教学方法与教学手段

1. **教学方法:** 课堂讲解、课堂讨论、上机实验、多媒体应用

2. **教学手段:** 多媒体、数据分析软件

## 九、建议教材与参考书目

### 1. 建议教材:

- ①大数据分析挖掘，石胜飞，人民邮电出版社，2018
- ②深度学习—基于 Python 和 TensorFlow 平台，谢琼，人民邮电出版社，2018

### 2. 参考书目:

- ①Python 数据处理与挖掘，吴振宇、李春忠、李建峰，人民邮电出版社，2020
- ②应用多元统计分析，高惠璇，北京大学出版社，2019
- ③数据科学与大数据技术导论，杜小勇，人民邮电出版社，2021

## 十、大纲编写的依据与说明

本课程教学大纲是根据数据科学与大数据技术专业培养目标和基本要求，结合本课程的性质、教学的基本任务和基本要求，及铜陵学院应用型本科院校建设及应用性人才培养方案等来制定的。

撰写人：蒋剑军

审核人：

日期：2021 年 6 月