

# 《大数据分析技术》实验教学大纲

## (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、目的

作为大数据“采集-存储-分析-应用”中最重要环节的“分析”，本课程是它的技术实现。所以本课程是以实验为主的学科。

实验教学目的有两个：一是培养学生熟练掌握 excel 操作能力和 Python 编程能力；二是让学生熟练掌握运用 excel 和 Python 实现数据预处理、因果推断、聚类分析、模式识别、特征重要性挖掘、深度学习等数据分析技术，通过自己设计和动手操作，从实验中学习、探索和发现数据规律，体验数据价值发现的挫折和欢乐，从而激发学生学习数据规律、挖掘数据价值的兴趣。

### 二、要求

#### 1. 知识要求

（1）理解数据预处理的概念，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欢乐和挫折。

(5) 写作能力，独立撰写实验报告的能力。

## 二、实验项目内容及学时分配

共设计了 15 个实验，其中 13 个必做，2 个选做。实验课题会在专业建设、人才培养的探索中与时俱进、增量更新。

### 实验一、数据预处理 1-大型百货商场会员消费分析（2 学时）

#### 1. 实验目的要求

(1) 熟悉 Python 操作环境、工作原理与命令形式，熟悉 numpy、pandas 等工具箱的应用，熟练编写 Python 程序。

(2) 掌握 Python 中数据的输入与输出操作。

(3) 掌握 Python 中两个数据集的运算：交、并、差，等。

(4) 掌握在 excel 中进行删除脏信息等基本数据清洗方法。

(5) 掌握应用 Python 进行缺失数据填充等基本数据清洗方法的操作，了解数据集成的方法。

#### 2. 实验主要内容

(1) 应用 excel 和 python 从大型百货商场消费明细进行会员消费情况分析。

3. 实验类别：基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

### 实验二、数据预处理 2-大型百货商场会员消费效益分析（2 学时）

#### 1. 实验目的要求

(1) 熟练编写 Python 程序，熟悉 numpy 和 pandas 工具箱在相关领域的应用。

(2) 掌握 Python 中数据的输入与输出操作。

(3) 掌握 Python 的数值计算方法。

(4) 掌握应用 Python 进行数据变换等数据预处理方法。

#### 2. 实验主要内容

(1) 应用 excel 和 python 从大型百货商场消费明细进行会员消费效益分析。

(2) 应用 python 进行数值计算。

3. 实验类别：基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

### 实验三、商务数据分析（2 学时）

#### 1. 实验目的要求

- (1) 在实验一和实验二的基础上进行大型百货商场会员管理策略分析。
- (2) 熟练应用 Python 进行描述性统计分析。
- (3) 了解大型百货商场会员制的商业模式。

#### 2. 实验主要内容

- (1) 对会员和非会员消费特征进行描述性统计分析，对商场收益进行对比；
- (2) 验证会员管理策略基本结论：实行会员制比不实行会员制有更高的收益。

#### 3. 实验类别：基础

#### 4. 实验类型：验证、综合

#### 5. 实验要求：必做

#### 6. 主要仪器：计算机

### 实验四、水箱水流量问题（2 学时）

#### 1. 实验目的要求

- (1) 了解插值的含义及基于 Python 的实现。
- (2) 掌握应用插值技术进行缺失数据的填充。

#### 2. 实验主要内容

- (1) 根据测得的“时刻-水位”数据，估计水箱中的水在任意时刻的流出速度；
- (2) 插值的 Python 实现。

#### 3. 实验类别：专业基础

#### 4. 实验类型：综合

#### 5. 实验要求：选做

#### 6. 主要仪器：计算机

### 实验五、工件轮廓线问题（2 学时）

#### 1. 实验目的要求

- (1) 了解拟合的含义及基于 Python 的实现，了解插值与拟合的异同。
- (2) 掌握应用拟合技术进行缺失数据的填充和因果推断。
- (3) 掌握应用 python 进行奇异点识别的方法和步骤。

#### 2. 实验主要内容

- (1) 根据测得的工件轮廓线的坐标数据，推断两个坐标值之间的因果关系；
- (2) 拟合的 Python 实现。

#### 3. 实验类别：专业基础

#### 4. 实验类型：综合

#### 5. 实验要求：必做

6. 主要仪器：计算机

## 实验六、税收与 GDP 关系探究（2 学时）

### 1. 实验目的要求

（1）理解一元线性回归模型的概念，掌握应用 Python 实现一元线性回归模型的参数估计及模型评估，能应用 Python 做各种统计检验。

（2）掌握应用 Python 工具箱 stats, scipy, sklearn 解线性回归模型的方法和步骤。

（3）应用 python 进行奇异点识别。

### 2. 实验主要内容

探究税收与 GDP 之间的关系。

3. 实验类别：专业基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验七、产品销量与广告媒体投入之间关系分析（2 学时）

### 1. 实验目的要求

（1）理解多元线性回归模型的概念，掌握应用 Python 实现多元线性回归模型的参数估计及模型评估，能应用 Python 做各种统计检验。

（2）掌握应用 Python 工具箱 stats, scipy, sklearn 求解线性回归模型的方法和步骤。

（3）了解广告投入与产品销量之间的关系，收获一些商业模式相关的知识。

### 2. 实验主要内容

分析产品销量与广告媒体投入之间的关系。

3. 实验类别：专业基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验八、中国新冠疫情趋势分析（2 学时）

### 1. 实验目的要求

（1）理解一元非线性回归模型的概念，掌握应用 Python 实现一元非线性回归模型的参数估计及模型评估，能应用 Python 做各种统计检验。

（2）理解 logistic 回归的概念，应用 logistic 回归模型解决我国新冠疫情发展趋势问题：新冠累计患者数量、自然传播率、拐点等。

（3）掌握应用 Python 工具箱 stats, scipy, sklearn 求解一元非线性回归模型的方法和步骤。

（4）了解我国新冠疫情爆发的前因后果，为战胜疫情贡献绵薄之力。

## 2. 实验主要内容

应用 **logistic** 模型预测我国新冠疫情发展趋势。

## 3. 实验类别：专业基础

## 4. 实验类型：综合

## 5. 实验要求：必做

## 6. 主要仪器：计算机

# 实验九、口罩销量影响因素研究（2 学时）

## 1. 实验目的要求

（1）了解多元非线性回归模型的概念，会应用 **Python** 实现多元非线性回归模型的参数估计及模型评估，能应用 **Python** 做各种统计检验。

（2）了解应用 **Python** 求解非线性回归模型的方法和步骤。

（3）了解空气质量的描述方法，做一个环保的人，为国家“绿水青山就是金山银山”的环保战略贡献一份力量。

## 2. 实验主要内容

口罩销量与空气质量因素“两尘四气”的关系研究。

## 3. 实验类别：专业

## 4. 实验类型：综合

## 5. 实验要求：选做

## 6. 主要仪器：计算机

# 实验十、信贷影响因素研究（2 学时）

## 1. 实验目的要求

（1）理解决策树和随机森林模型的概念，掌握应用决策树求解问题的最优解决方案，掌握应用随机森林挖掘特征的重要性，掌握应用 **Python** 实现决策树和随机森林的方法和步骤。

（2）了解信用贷款的背景和一些操作方法，做一个讲信用、有信誉的人。

## 2. 实验主要内容

根据个人信用数据为银行提供是否发放贷款的最优决策。

## 3. 实验类别：专业

## 4. 实验类型：综合

## 5. 实验要求：必做

## 6. 主要仪器：计算机

# 实验十一、我国城市发展水平分析（2 学时）

## 1. 实验目的要求

（1）理解聚类分析的概念，掌握应用模糊 C 均值聚类、k 均值聚类、k 最邻近聚

类等方法进行聚类分析，掌握应用 Python 实现上述聚类方法的步骤，并能熟练地对各聚类方法进行评估。

(2) 了解我国城市发展水平，对我国国情有更深入的认识。

## 2. 实验主要内容

根据城市发展相关指数，分析我国城市发展水平。

3. 实验类别：专业基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验十二、空气质量评估研究（2 学时）

### 1. 实验目的要求

(1) 了解支持向量机的数学原理，掌握支持向量机在聚类分析中的应用，熟练掌握支持向量机的 Python 实现方法，并能熟练地调节支持向量机中的参数。

(2) 了解空气质量的描述方法，做一个环保的人，为国家“绿水青山就是金山银山”的环保战略贡献一份力量。

### 2. 实验主要内容

根据空气质量中“两尘四气”含量描述空气质量等级，并应用支持向量机对质量等级未知的空气的质量进行识别。

3. 实验类别：专业

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验十三、鸢尾花的识别（2 学时）

### 1. 实验目的要求

(1) 理解掌握经典神经网络模型的概念及建立方法，熟练掌握基于 Python 实现的经典神经网络在聚类分析中的应用。

(2) 了解鸢尾花数据集的来源及相关应用研究。

(3) 掌握应用经典神经网络对鸢尾花进行聚类分析。

### 2. 实验主要内容

应用经典神经网络模型进行鸢尾花的识别。

3. 实验类别：专业基础

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验十四、图像识别（2 学时）

### 1. 实验目的要求

(1) 了解图形识别的概念，掌握应用基于 Python 语言和 TensorFlow 平台的深度学习方法实现图像识别。

(2) 掌握基于 Python 语言和 TensorFlow 平台的深度学习方法。

(3) 理解掌握卷积神经网络在图像识别中的应用。

### 2. 实验主要内容

应用基于 Python 语言和 TensorFlow 平台的深度学习方法“卷积神经网络”实现图像识别。

3. 实验类别：专业

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 实验十五、上证指数预测（2 学时）

### 1. 实验目的要求

(1) 了解时间序列的概念，掌握应用基于 Python 语言和 TensorFlow 平台的深度学习方法实现时间序列的预测。

(2) 理解掌握循环神经网络“LSTM”在时间序列预测中的应用。

(3) 了解金融时间序列的概念。

### 2. 实验主要内容

应用基于 Python 语言和 TensorFlow 平台的深度学习方法“循环神经网络—LSTM”实现上证指数的预测。

3. 实验类别：专业

4. 实验类型：综合

5. 实验要求：必做

6. 主要仪器：计算机

## 三、考核方式

1、实验成绩：操作过程 20%、实验报告 70%、实验记录 10%

2、无期中抽考、有实验课的本课程最终成绩计算方法

最终成绩=平时成绩×0.1+实验成绩×0.6+期末考查成绩×0.3

3、折算等级制：优≥90、良≥80、中≥70、及格≥60、不及格<60

撰写人：蒋剑军

审核人：

日期：2021 年 6 月