**《分布式统计方法》课程教学大纲**

一、课程基本信息

课程代码：18100083

课程名称: 分布式统计方法

英文名称：Distributed Statistical Methods

课程类别：专业课

学 时：48

学　　分：3

适用对象：数据科学与大数据技术专业大四学生

考核方式：考试

先修课程：概率论，数理统计，线性代数，机器学习

二、课程简介

中文简介

本课程介绍与分布式统计问题相关的各种算法，包括各种分类、回归、聚类算法，最速下降法、牛顿法、拟牛顿法、BFGS算法等等。

英文简介

This course introduces different types of algorithms related to distributed statistical problems, including classification, regression, clustering, steepest descent method, Newton’s method, quasi Newton method, BFGS methods etc.

三、课程性质与教学目的

本课程是面向大四大数据专业学生开设的专业选修课，也可作为其他专业学生的选修课。课程目的是介绍与分布式统计问题密切相关的各种算法，培养和增强学生解决实际数据分析问题的能力。

课程思政总体思路：分布式统计方法包含很多统计学基本方法和最优化方法，如今广泛应用于人工智能，运筹，物流，国防等领域。世界各国无不将最新的智能算法应用于生活、生产乃至于战争中，以掌握主动权。当前世界局势纷繁复杂，如果我们能够应用好最新的各种与分布式统计相关的算法，这将有助于我国在人工智能，国防等领域抢占先机，争夺话语权。

四、教学内容及要求

1. **人工智能简介、Apache Spark简介**
2. 目的与要求
3. 介绍人工智能的各种应用
4. 介绍Apache Spark编程框架、软件和安装注意事项
5. 教学内容

第一节 人工智能简介

1. 主要内容

人工智能的应用领域

1. 基本概念和知识点

机器学习，数据挖掘，模式识别

1. 问题与应用（能力要求）

了解机器学习，数据挖掘，模式识别的区别与联系

第二节 Apache Spark简介

1. 主要内容

介绍Apache Spark软件的使用技巧，安装过程

1. 基本概念和知识点

Apache Spark的环境配置

1. 问题与应用（能力要求）

会安装Apache Spark软件，并学习Apache Spark的编程语言特点

1. 思考与实践

思考Apache Spark软件与R，python等软件的联系。

1. 教学方法与手段

课堂讲授

1. **Breeze程序包**
2. 目的与要求
3. 介绍Breeze程序包的内容简装技巧
4. 掌握Breeze程序包的基本命令和常用库
5. 教学内容

第一节 创建向量、矩阵

1. 主要内容

利用Breeze包创建向量、矩阵

1. 基本概念和知识点

向量、矩阵

1. 问题与应用（能力要求）

掌握Breeze包创建向量、矩阵的技巧

第二节 整行与整列的运算

1. 主要内容

凸集，凸函数的定义及判别方法，闭包的含义

1. 基本概念和知识点

Breeze包创建整行、整列

1. 问题与应用（能力要求）

掌握Breeze包创建整行、整列的技巧

第三节 常用数学运算

1. 主要内容

常见的spark软件包中的加减乘除运算

1. 基本概念和知识点

数学中的各种运算

1. 问题与应用（能力要求）

会用Breeze包创建各种基本数学运算

第四节 常用分布

1. 主要内容

Breeze包和各种分布的创建

1. 基本概念和知识点

正态分布、二项分布等等

1. 问题与应用（能力要求）

会用Breeze包创建各种分布。

1. 思考与实践

Breeze包中创建的各种分布的命令与R、Python软件的区别

1. 教学方法与手段

课堂讲授

1. **随机模拟和统计推断**
2. 目的与要求

掌握随机数的产生方法和统计推断的技巧可

1. 教学内容

第一节 随机数的产生

1. 主要内容

用逆累积分布函数法、拒绝法产生随机数

1. 基本概念和知识点

逆累积分布函数法、拒绝法

1. 问题与应用（能力要求）

掌握逆累积分布函数法、拒绝法产生随机数

第二节 EM优化

1. 主要内容

EM算法，收敛性分析

1. 基本概念和知识点

EM算法

1. 问题与应用（能力要求）

掌握分布式EM算法。

1. 思考与实践

思考对于混合高斯模型如何运用分布式EM算法。

1. 教学方法与手段

课堂讲授

1. **马尔科夫链蒙特卡罗**
2. 目的与要求

理解和掌握利用马尔科夫链产生具有渐进分布相近的随机数的方法

1. 教学内容

第一节 Metropolis Hastings算法

1. 主要内容

Metropolis Hastings算法产生随机数

1. 基本概念和知识点

Metropolis Hastings算法

1. 问题与应用（能力要求）

能够利用 Metropolis Hastings算法产生随机数

第二节 Slice取样法

1. 主要内容

Slice取样法产生随机数

1. 基本概念和知识点

Slice取样

1. 问题与应用（能力要求）

能够利用 Slice取样算法产生随机数

第三节 Gibbs取样法

1. 主要内容

Gibbs取样法产生随机数

1. 基本概念和知识点

Gibbs取样法

1. 问题与应用（能力要求）

能够利用 Slice取样算法产生随机数

1. 思考与实践

思考Metropolis Hastings算法，Slice取样法， Gibbs取样法产生随机数的差异。

1. 教学方法与手段

课堂讲授

1. **优化方法**
2. 目的与要求

介绍最优化中的各种算法：交替方向乘子法，线性化Bregman迭代方法，最速下降法，牛顿法，拟牛顿法，共轭梯度法，BFGS算法等方法

1. 教学内容

第一节 交替方向乘子法，线性化Bregman迭代方法

1. 主要内容

介绍针对不同优化问题的交替方向乘子法，线性化Bregman迭代方法

1. 基本概念和知识点

交替方向乘子法、线性化Bregman迭代法

1. 问题与应用（能力要求）

掌握交替方向乘子法，线性化Bregman迭代方法

第二节 最速下降法，牛顿法

1. 主要内容

介绍最速下降法的计算法，锯齿现象，介绍牛顿法

1. 基本概念和知识点

最速下降方向，锯齿现象，牛顿方向

1. 问题与应用（能力要求）

掌握最速下降法、牛顿法的计算方法，优缺点，锯齿现象的产生原因。

第三节 拟牛顿法

1. 主要内容

介绍拟牛顿法的迭代思想

1. 基本概念和知识点

拟牛顿法

1. 问题与应用（能力要求）

掌握拟牛顿法的计算方法，优缺点。

第四节 共轭梯度法，BFGS算法

1. 主要内容

介绍共轭梯度法，BFGS算法的迭代思想

1. 基本概念和知识点

共轭方向，拟牛顿条件，DFP算法，BFGS算法，拟牛顿条件

1. 问题与应用（能力要求）

共轭梯度法，BFGS算法

1. 思考与实践

思考各种优化算法的优缺点。

1. 教学方法与手段

课堂讲授

1. **自举法**
2. 目的与要求

介绍约束规划问题的最优性条件，可行方向法，罚函数法，梯度投影法，Frank-Wolf方法，起作用集方法。

1. 教学内容

第一节 自由自举法

1. 主要内容

介绍自有自举法的基本思想

1. 基本概念和知识点

自有自举法。

1. 问题与应用（能力要求）

掌握自有自举法的计算步骤

第二节 子集合自举法

1. 主要内容

介绍子集合自举法的基本思想

1. 基本概念和知识点

子集合自举法

1. 问题与应用（能力要求）

掌握子集合自举法的计算步骤

1. 思考与实践

思考并掌握自有自举法和子集合自举法的计算步骤和原理。

1. 教学方法与手段

课堂讲授

1. **常用大数据统计方法**
2. 目的与要求

介绍KNN, K近邻法，Logistic回归，线性支持向量机，线性判别分析，决策树，一元回归，多元回归，l1算法。

1. 教学内容

第一节 KNN, K近邻

1. 主要内容

介绍KNN和K近邻算法的基本思想，区别与联系

1. 基本概念和知识点

KNN，K近邻

1. 问题与应用（能力要求）

掌握KNN和K近邻算法的计算要领

第二节 Logistic回归，线性支持向量机

1. 主要内容

介绍Logistic回归，线性支持向量机的原理和计算技巧

1. 基本概念和知识点

Logistic回归，线性支持向量机，最优间隔超平面

1. 问题与应用（能力要求）

掌握Logistic回归，线性支持向量机的计算技巧

第三节 线性判别分析，决策树

1. 主要内容

介绍线性判别分析，决策树的应用领域、计算技巧

1. 基本概念和知识点

线性判别分析，类类距离，类间距离，决策树

1. 问题与应用（能力要求）

掌握线性判别分析，决策树的应用领域

第四节 一元回归，多元回归，l1算法

1. 主要内容

介绍一元回归和多元回归的原理，计算步骤以及岭回归，lasso方法

1. 基本概念和知识点

最小二乘回归，岭回归，lasso

1. 问题与应用（能力要求）

掌握一元回归，多元回归的区别于联系，会运用岭回归，l1算法来求解实际问题

1. 思考与实践

思考KNN, K近邻，Logistic回归，线性支持向量机，线性判别分析，决策树，一元回归，多元回归，l1算法等方法的计算步骤的应用领域。

1. 教学方法与手段

课堂讲授

1. **数据降维**
2. 目的与要求

介绍主成分分析、奇异值分解、典型相关分析的基本概念和计算思想，应用领域。

1. 教学内容

第一节 主成分分析

1. 主要内容

介绍主成分分析的基本原理

1. 基本概念和知识点

主成分，协方差矩阵

1. 问题与应用（能力要求）

掌握主成分分析的求解原理

第二节 奇异值分解

1. 主要内容

介绍奇异值分解的原理，掌握奇异值分解和特征值分解的区别与联系

1. 基本概念和知识点

奇异值分解

1. 问题与应用（能力要求）

掌握奇异值分解的计算步骤

第三节 典型相关分析

1. 主要内容

介绍典型相关分析的基本原理

1. 基本概念和知识点

典型相关分析，投影向量

1. 问题与应用（能力要求）

掌握典型相关分析的计算步骤，应用领域，会计算典型载荷

第四节 卷积神经网络CNN,循环神经网络RNN及LSTM

1. 主要内容

介绍卷积神经网络CNN,循环神经网络RNN及LSTM基本的原理和计算思想

1. 基本概念和知识点

卷积，上采样，下采样，门原理

1. 问题与应用（能力要求）

掌握卷积神经网络CNN,循环神经网络RNN及LSTM基本的原理和计算思想

1. 思考与实践

思考主成分分析与典型相关分析两者之间的联系，思考循环神经网络RNN及LSTM的区别。

1. 教学方法与手段

课堂讲授

五、各教学环节学时分配

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **教学环节**  **教学时数**  **课程内容** | **讲**  **课** | **习**  **题**  **课** | **讨**  **论**  **课** | **实验** | **其他教学环节** | **小**  **计** |
| 第一章 | 4 | 0 | 随堂讨论 |  |  |  |
| 第二章 | 4 | 0 | 随堂讨论 | 2 |  |  |
| 第三章 | 2 | 0 | 随堂讨论 | 2 |  |  |
| 第四章 | 2 | 0 | 随堂讨论 | 2 |  |  |
| 第五章 | 6 | 0 | 随堂讨论 | 2 |  |  |
| 第六章 | 4 | 0 | 随堂讨论 | 2 |  |  |
| 第七章 | 6 | 0 | 2 | 2 |  |  |
| 第八章 | 4 | 0 | 2 | 0 |  |  |
| 合计 | 32 | 0 | 4 | 12 |  |  |

六、课程考核

（一）考核方式：考试

（二）成绩构成

平时成绩占比：40% 期末考试占比：60%

（三）成绩考核标准

课堂考勤：准时到达课堂，不迟到，不早退，占10%；课堂表现：遵守课堂秩序，认真听讲，积极参与课堂互动，占5%；平时作业：按时，高质量完成平时作业，占25%；期末考试：占60%。

六、推荐教材和教学参考资源

1. 冯兴东，分布式统计计算. 上海财经大学出版社，2018.

八、其他说明

无

大纲修订人： 蔡佳 修订日期：2021年11月13日

大纲修订人： 张赞波 修订日期：2022年08月30日

大纲审定人： 审定日期：