# 《数据科学基础》课程大纲

# 一、 课程基本信息

| | | | | | |
|---|---|---|---|---|---|
| **课程编号** | 06113220 | **学分** | 2 | **开课学期** | □秋 ■ 春 |
| **课程名称** | **（中文）**数据科学基础 | | | | |
| | **（英文）**Foundations of Data Science | | | | |
| **课程类别** | 专业理论课 | | | | |
| **课程学时** | **32** | | | | |
| **教学方式** | 课堂讲授为主 ，结合讨论 | | | | |
| **考核方式**<br>**评分依据** | 大作业<br>平时成绩（课堂参与）10%，大作业成绩 90% | | | | |
| **先修课程**<br>**预备知识** | 概率论，数理统计，线性代数 | | | | |
| **教材与**<br>**参考文献** | **教材：**Avrim Blum, John Hopcroft and Ravindran Kannan， Foundations of Data Science， 2016.1<br>**参考文献：**<br>（1）Jiawei Han ,Micheline Kamber, Jian Pei ,Data Mining: Concepts and Techniques，3rd Edition<br>（2）Rachel Schutt, et.al.，Doing Data Science，OREILLY ，2013.10<br>(3) Gareth James, et al.，An Introduction to Statistical Learning，Springer ，2013.9 | | | | |

# 二、 教学目标

本课程以数据为主体，将大规模数据有效计算作为数据科学的基本问题，通过讲授数据科学中数据表示与主流数据算法的数学基础与基本原理，使学生掌握数据分析处理所依据的理论基础与思维方法，从而提高研究生的专业能力并促进学科发展。

# 三、 课程简介

数据科学是通过数据计算揭示自然界及人类行为和规律的科学，包含了数据处理的理论、方法与技术。从 20 世纪 60 年代提出，已逐渐被人们认识并成为当今计算机学科的热点领域。本课程将大规模数据有效计算作为数据科学的基本问题，讲述计算空间中数据的数学表示、基本变换方法，大规模数据的表达结构及其操作分析的基础算法，机器学习基础理论以及数据流等不同形态数据的处理与计算模型等。课程使学生能够掌握大规模数据计算的基础理论与基本方法，为进一步开展相关研究或开发奠定坚实的理论基础。

# 四、 课程教学内容

**第 1 章 概述 （2 学时）**

1. 什么是数据科学
2. 数据科学的任务构成
3. 数据科学的理论基础
4. 数据科学基础内涵

**第 2 章 高维空间(High-Dimensional Space) （6 学时）**

1. 随机变量与高维空间
2. 高维空间概述
3. 大数定律
4. 高维空间的几何特性
5. 高维空间中的高斯分布
6. 中心极限定理
7. 高维数据的降维问题

**第 3 章 最佳子空间和奇异值分解 （4 学时）**

1. 概述
2. 数据中心化
3. 奇异向量
4. 奇异值分解 SVD
5. 秩为 k 的最佳近似子空间
6. 左奇异向量
7. 奇异值分解计算中的幂方法
8. 奇异向量和特征向量

**第 4 章 随机图 （6 学时）**

1. 随机图模型 G(n,p)的定义
2. 随机图相变及相变阈值
3. 相变阈值判定方法
4. 图的巨大分支
5. 圈（cycle）与连通分支
6. 分支过程
7. 图渐增特性的相变
8. 非均匀随机图
9. 图的成长模型

**第 5 章 随机游走与马尔可夫链 （6 学时）**

1. 概述
2. 马尔可夫链的平稳分布
3. 马尔科夫链蒙特卡罗（MCMC）方法原理
4. MCMC 中的马尔可夫链生成方法

5. 随机游走在无向图上的收敛
6. 电子网络与逃逸概率
7. 单位边权重无向图上的随机游走
8. 马尔可夫链在 Web 中的应用
9. 隐马尔可夫链

**第 6 章 机器学习　　（6 学时）**

1. 概述
2. 过拟合和一致收敛
3. 奥卡姆剃刀—学习的简单性原理
4. 正则化
5. VC-维
6. 在线学习与感知机算法
7. 核函数
8. 支持向量机
9. 强学习与弱学习—Boosting 方法
10. 优化方法-随机梯度下降法

**第 7 章 海量数据算法　　（2 学时）**

1. 概述
2. 数据流的频率矩及其应用
3. 基于抽样的大矩阵计算方法

# 四、课程实践环节

以大作业形式组织。分组进行，2 人一组，选择与课程相关的一个主题，作业内容分理论部分和应用与实验部分。对于理论部分需要调研文献，清楚阐述所包含模型或算法的原理；对于应用与实验部分，自行确定数据集，利用题目理论部分的模型/算法做实验，整理与说明实验结果。

撰写实践报告，并以 PPT 形式进行课堂讲解与讨论。