数据科学与工程算法
Algorithm Foundations of Data Science and Engineering

Lecture 0: 课程简介

王延昊 副教授

华东师范大学数据科学与工程学院
电子邮件: yhwang@dase.ecnu.edu.cn

2023/2/28

大纲

- 课本与参考资料
- 课程基本要求
- 联系方式
- 课程概述
  - What Is Data Science?
  - Course Schedule
- 建议

课本与参考资料

- 课本 (必须)
  - 高明, 胡卉芪(著). 数据科学与工程算法基础
  - Avrim Blum, John Hopcroft, Ravindran Kannan. Foundations of Data Science
  - Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets

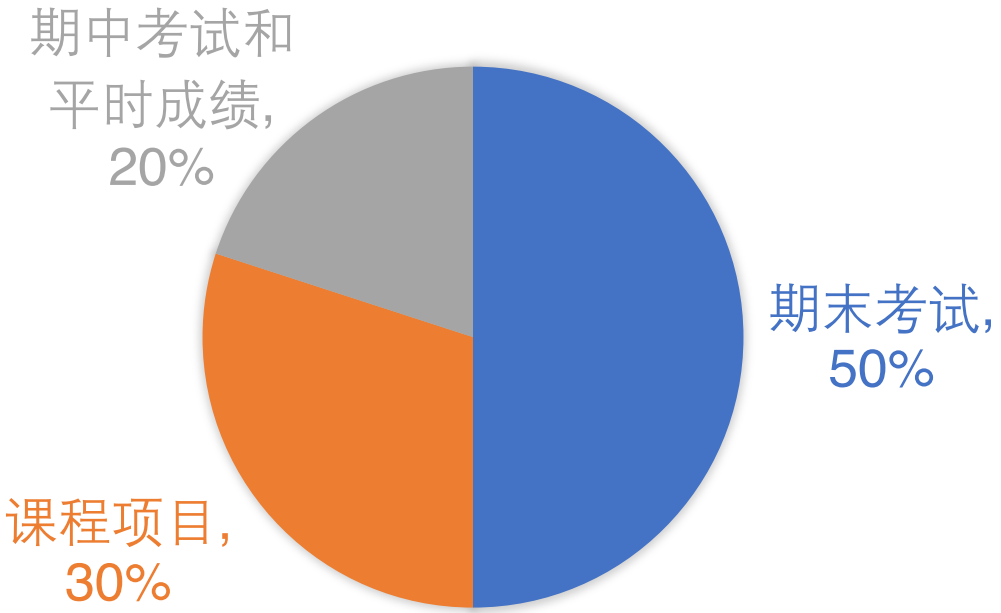
- 参考资料 (可选)
  - Gilbert Strang. Linear Algebra and Its Applications (Fourth Edition)
  - Albert Bifet. Machine Learning for Data Streams with Practical Examples in MOA https://moa.cms.waikato.ac.nz/book/
  - John E. Mitchell. Integer and Combinatorial Optimization https://homepages.rpi.edu/~mitchj/matp6620/

基本要求

1. 中/英文课件将在上课前2天上传至课程网站

2. 学生需要
   - *上课认真听讲
   - *课前预习当节课程内容
   - *下一周上课前完成上周课程的课后练习题
   - 有余力的情况下，完成课外扩展阅读

课程评价



联系方式

- 任课教师：王延昊
  - 办公室: 数学馆东115室
  - 电子邮件: yhwang@dase.ecnu.edu.cn
- 助教: 浦家希 & 李佳
  - 办公室: 数学馆东102室
  - 电子邮件: 51215903021@stu.ecnu.edu.cn & jiali@stu.ecnu.edu.cn

- 课程主页：https://yhwang1990.github.io/ads-2023-spring

- 研究方向
  - Data stream mining
  - Graph mining
  - Privacy-preserving data mining
  - Algorithmic fairness

课程安排

- **理论课**
  - 每周二 上午9:50至11:25 (2023/2/28 – 2023/6/27)
  - 每周五 下午1:00至2:35 (2023/3/3 – 2023/3/31)
  - 地点：教书院223

- **实验课**
  - 每周五 下午1:00至2:35 (2023/4/7 – 2023/6/30)
  - 地点：教书院223
  - 项目1：2023/4/7 – 2023/4/28，项目报告2023/5/6前提交
  - 项目2：2023/5/5 – 2023/5/26，项目报告2023/6/3前提交
  - 项目3：2023/6/2 – 2023/6/23，项目报告2023/7/1前提交

课程安排

- **课程背景**
  - 课程概述 (第1周)
- **概率与统计算法**
  - 抽样算法 (第1周)
  - 概率不等式 (第2周)
  - 哈希算法 (第3周)
  - 概要数据结构 (第4周)
  - 数据流算法 (第5周)
  - 马尔科夫链与随机游走 (第6，7周)
- **期中总结与复习 (第8周)**
- **期中考试 (第9周)**

课程安排

- **线性代数**
  - 特征值计算 (第10周)
  - SVD and PCA (第11周)
  - 矩阵因式分解 (第12周)
- **组合优化**
  - 线性规划和整数规划 (第13-14周)
  - 子模函数优化 (第15周)
  - 社区发现 (第16-17周)
- **期末总结与复习 (第18周)**

Data Science and Big Data

- How to understand big data?
  - **Volume:** PBs data daily processed by Baidu and Google; Alibaba and Tencent have data more than 100PB.
  - **Velocity:** Large Hadron Collider generates PB data in seconds; many streaming such as clickstream, log, Twitter. #Trans. is almost 100,000 per second in Taobao during “Double 11”.
  - **Variety:** Structured, semi-structured and non-structured, like text, logs, video, voice, image.
  - **Value:** Interests, behaviors, trustworthiness, and privacy, …
- Fragmentation of information
  - Telecom
  - E-commerce
  - Social media
  - Internet of things (IOT)
  - ……

Birth of Data Science

- **Challenges** of 4V: Volume, Velocity, Variety, Value
- **Modern Hardware:** GPU, FPGA, TB's Memory, GB's Network, …
- **Open-Source Software Platform:** Hadoop, Spark, Storm, TensorFlow, …
- **Applications:** E-Commerce, Sharing Economy, Internet of Things (IoT), Industry 4.0, Smart City, Intelligent Education, …

Data is Important

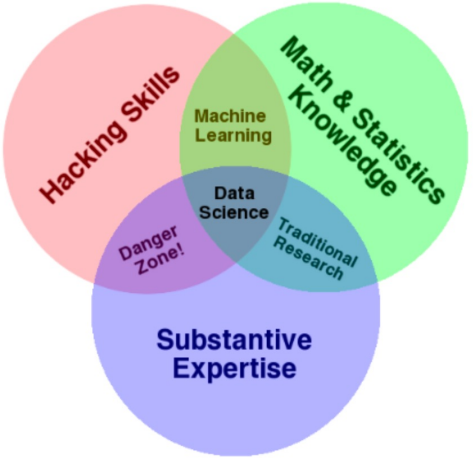
- Data becomes an independent factor of production
  - In 2017, in the age of the Internet economy, data is a new factor of production, a fundamental resource, and a strategic resource
  - On April 9, 2020, data becomes a new factor of production, just like land, labor, capital, and technology

- Data is the foundational resource of the digital economy, facing many challenges such as data silos, digital divide, data privacy and data security.

Data is Power

What is Data Science?

- **Data Science** is an interdisciplinary field, which is a continuation of different data analysis fields such as mathematics, statistics, machine learning, data mining, and parallel computing, similar to Knowledge Discovery in Databases (KDD).



The Goals of Data Science:

- Extract knowledge
- Insight from data in various forms, either structured or unstructured
- Help users understand massive data

DS Co-Evolution

- Data science was mentioned by John W. Tukey in 1962 (“The Future of Data Analysis”)
- Data science was defined by Peter Naur in 1974 (“Concise Survey of Computer Methods”)
- Many data mining methods were proposed in the 1980s of the 20th century
- In 1996, international federation of classification societies issue set up a conference, namely Data Science, Classification and Related Methods
- In June 2009, Nathan Yau published a paper talking about the rising of data science
- Data scientist is the sexiest job in the 21st century (Hal Varian on Sep. 2012)

Types of Data Scientists

- **Data developer**: data acquisition, organization and management.
- **Data researcher**: statisticians, social scientist, computer scientist, etc.
- **Data creative**: experts in machine learning, data mining, and programming, etc., contributor in open-source community,
- **Data businessman**: project manager, Chief Data Officer (CDO)
- **Mixed/Generic type**: deep-understand in business, professional in technology, good at programming, etc.

Why do we need to learn this course?

N	Algorithm	2016	2011	Domain
1	Regression	67%	58%	Statistics
2	Clustering	57%	52%	Data Mining / Statistics
3	Decision Trees	55%	60%	Data Mining
4	Visualization	49%	38%	Visualization
5	K-nearest neighbors	46%	-	Data Mining
6	PCA	43%	-	Statistics
7	Statistics	43%	48%	Statistics
8	Random Forests	38%	-	Data Mining
9	Sequence analysis	37%	30%	Data Mining
10	Text Mining	36%	28%	NLP
11	Ensemble methods	34%	28%	Machine Learning
12	SVM	34%	29%	Machine Learning
13	Boosting	33%	23%	Machine Learning

N	Algorithm	2016	2011	Domain
14	Neural networks	24%	27%	Machine Learning
15	Optimization	24%	-	Optimization
16	Naive Bayes	24%	22%	Machine Learning
17	Data Integration	22%	20%	Data Management
18	Anomaly detection	20%	16%	Data Mining
19	Deep Learning	19%	-	Machine Learning
20	SVD	16%	-	Algebraic
21	Association rules	15%	29%	Data Mining
22	Graph Mining	15%	14%	Data Mining
23	Bayesian networks	13%	-	Machine Learning
24	Genetic algorithms	8.8%	9.3%	Machine Learning
25	Survival Analysis	7.9%	9.3%	Statistics
26	EM	6.6%	-	Statistics

Why do we need to learn this course?

Remarks

1. Most popular among new options added in 2016 are K-nearest neighbors, PCA, Random Forests, Optimization, Neural networks, Deep Learning, and Singular Value Decomposition
2. The biggest declines are Association rules, Statistics, and Decision Trees

Course Features

N	Domain	Count
1	Data Mining	9
2	Machine Learning	8
3	Statistics	4
4	Visualization	1
5	NLP	1
6	Data Management	1
7	Optimization	1
8	Algebra	1

N	Data Model	Type
1	Relational Database	Structured
2	Time Series	Semi-Structured
3	Graph	Semi-Structured
4	Text	Non-Structured
5	Image	Non-Structured
6	Video	Non-Structured
7	Audio	Non-Structured

Features

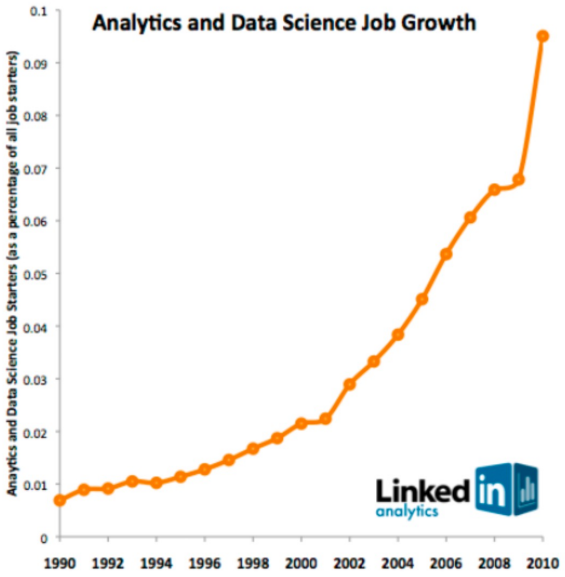
1. Algorithms for data science involve in many disciplines, such as data mining, machine learning, statistics, visualization, NLP, data management, optimization, and algebra, etc.
2. Tasks in data science problems are various in data types.

Four Paradigms of Scientific Research

- Experimental science
- Theoretical science
- Computational science

- Data science?
  - It was firstly proposed by Jim Gray (a database researcher) in 2009.
  - The Forth Paradigm: Data-Intensive Scientific Discovery was written by Tony Hey (vice president of Microsoft) et al. in 2009.
  - Thus, the capability for big data processing is important to scientific researchers.

The Shortage of Data Scientists



Take-Aways

- Advices to learning algorithm foundations of data science and engineering
  - Not a reading course
  - More than a programming course, though it is project-heavy
  - No standard answers