



数据挖掘

Data Mining

第一课 数据挖掘绪论

主讲人：丁兆云、周鳌

<https://github.com/zyding1983/datamining>



效果测试

A

能够听到声音，效果正常

B

能够听到声音，偶尔卡顿

C

能够听到声音，卡顿严重

D

听不到声音

提交



手头目前的工具

A**手机****B****电脑****C****笔****D****纸****提交**



学生专业人数了解

- A
- B
- C
- D
- E
- F
- G

- 管理科学与工程专业
- 目标工程专业
- 运筹与任务规划专业
- 大数据工程专业
- 仿真工程专业
- 指挥信息系统工程专业
- 其他专业

提交

课程教辅

< 聊天信息(8) ↗

管理科学与工程专业



d



刘斌



周鳌



朱先强



朱席席



刘恺



徐翔



王庆勇



目标工程专业

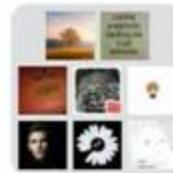
运筹与任务规划专业

大数据工程专业

课程教辅-学生在微信群中的命名规则

专业	命名规则	示例
管理科学与工程	管科+姓名	管科张三
目标工程	目标+姓名	目标王五
运筹与任务规划	规划+姓名	规划李四
大数据工程	数据+姓名	数据孙九

微信二



数据挖掘2020上本科



该二维码7天内(2月19日前)有效，重新进入将更新

zoom会议信息

网址: <https://zoom.com.cn>

下载地址: <https://zoom.com.cn/download>

会议ID: **835-421-5851**

[**https://zoom.com.cn/j/8354215851**](https://zoom.com.cn/j/8354215851)

内容提纲

- I . 数据挖掘由来
- II . 数据挖掘的过程模型
- III. 数据挖掘的主要研究内容
- IV. 主要参考资料
- V. 课程要求

背景

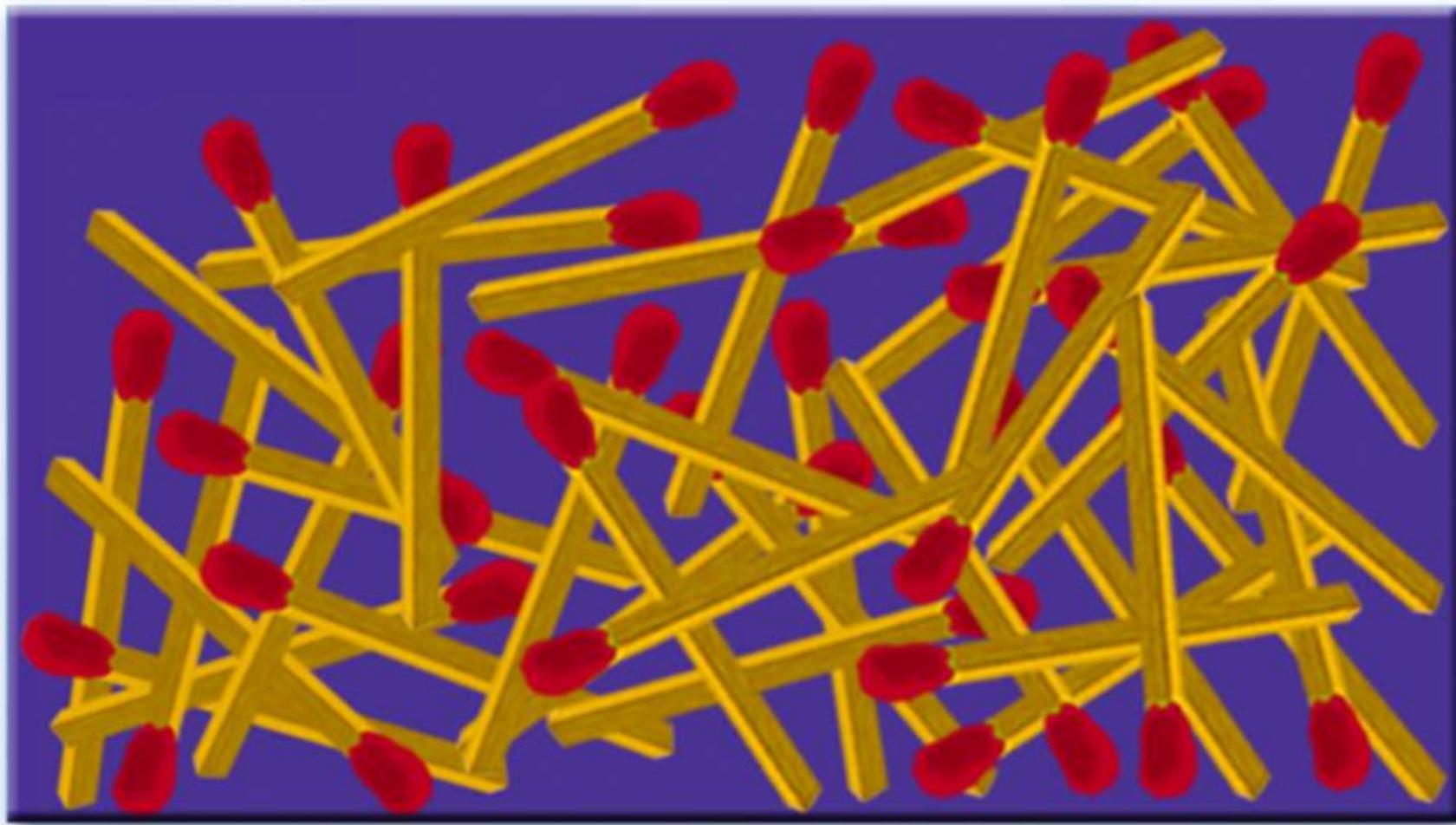


背景

- 随着大数据库的建立和海量数据的不断涌现，必然提出对强有力的数据分析工具的迫切需求。但现实情况往往
是“**数据十分丰富，而信息相当贫乏。**”
- 快速增长的海量数据收集、存放在大型数据库中，没有
强有力的工具，理解它们已经远远超出人的能力。因此，
有人称之为：“**数据坟墓**”。



这里有几根火柴？

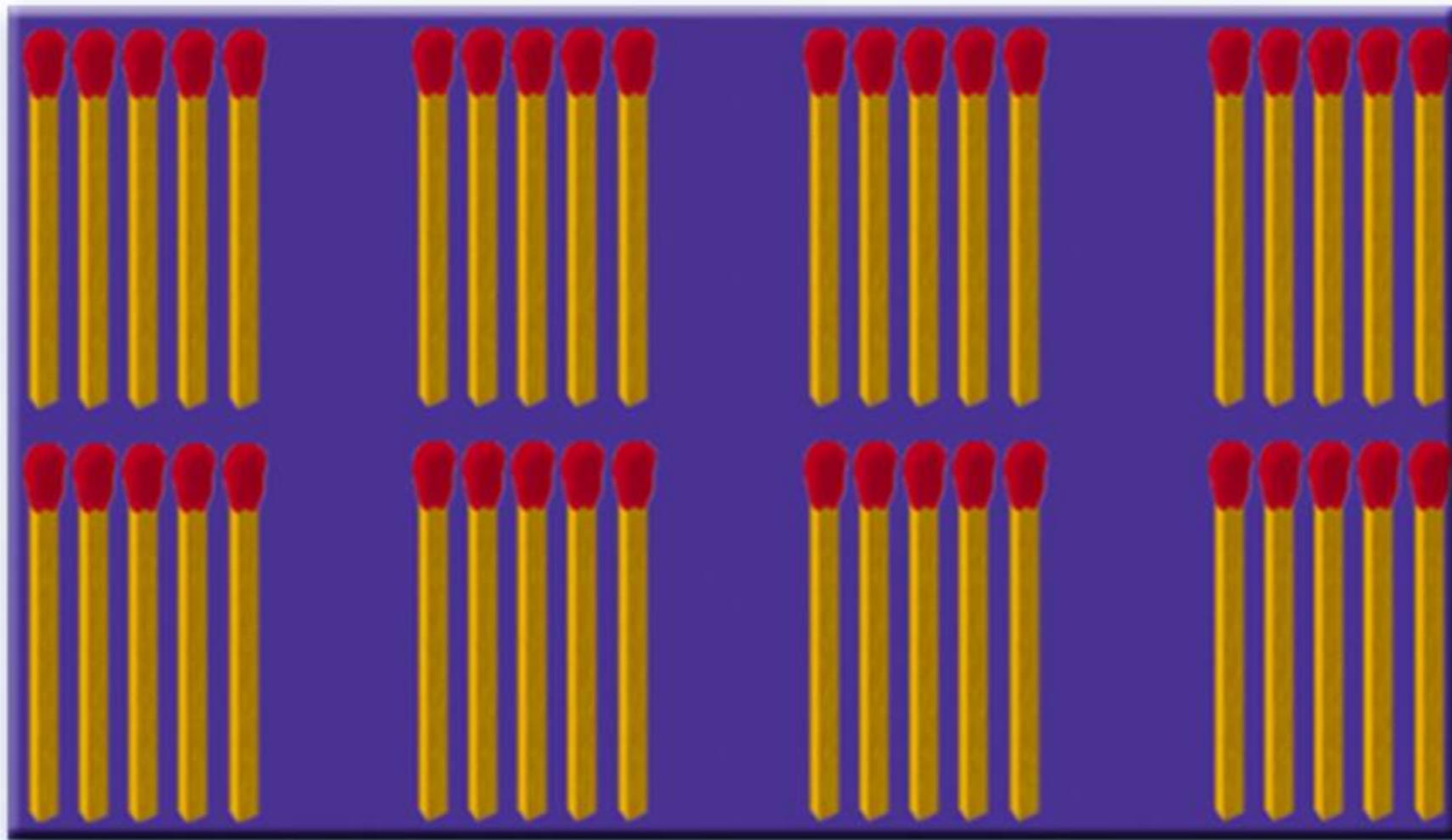


有多少根火柴棒

- A 30
- B 35
- C 40
- D 45

提交

现在呢？



有多少根火柴棒

- A 30
- B 35
- C 40
- D 45

提交

为什么要用数据挖掘（1/3）

■ 数据爆炸但知识贫乏

□ 人们积累的数据越来越多。但是，目前这些数据还仅仅应用在数据的录入、查询、统计等功能，无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，导致了“数据爆炸但知识贫乏”的现象。

为什么要用数据挖掘（2/3）

■ 从商业数据到商业智能的进化

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集 (60年代)	“过去五年中我的总收入是多少？”	计算机、磁带和磁盘	IBM CDC	提供历史性的、静态的数据信息
数据访问 (80年代)	“在新英格兰的分部去年三月的销售额是多少？”	关系数据库(RDBMS) 结构化查询语言(SQL) ODBC	Oracle Sybase Informix IBM Microsoft	在记录级提供历史性的、动态数据信息
数据仓库 决策支持 (90年代)	“在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论？”	联机分析处理(OLAP) 多维数据库 数据仓库	Pilot Comshare Arbor Cognos Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘 (正在流行)	“下个月波士顿的销售会怎么样？为什么？”	高级算法 多处理器计算机 海量数据库	Pilot Lockheed IBM SGI 其他初创公司	提供预测性的信息

为什么要用数据挖掘（3/3）

■ 科学发展范式



KDD的出现

- 基于数据库的知识发现（KDD）一词首次出现在1989年举行的国际人工智能联合大会IJCAI-89 Workshop。
- 1995年在加拿大蒙特利尔召开了第一届KDD国际学术会议（KDD'95）。
- 由Kluwers Publishers出版，1997年创刊的《Knowledge Discovery and Data Mining》是该领域中的第一本学术刊物。

数据挖掘的定义

- 数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- Data Mining is the process of automatically extracting interesting and useful hidden patterns from usually massive, incomplete and noisy data. [Wikipedia]

II. 数据挖掘的过程模型

- 数据、信息、知识
- 数据挖掘的过程
- 数据来源



数据、信息、知识

客户信息表

25	10	2	优
29	12	3	优
32	9	6	良
38	7	12	良
36	18	13	中
30	15	4	优
...

数据、信息、知识

客户信息表

25	10	2	优
29	12	3	优
32	9	6	良
38	7	12	良
36	18	13	中
30	15	4	优
...

数据
信息

25<年龄<30, 收入>10, 工作时间>2年的消费者是优质顾客

知识

数据：符号、事实和数字

➤ 数据

★ 数据是未经加工和修饰的原料。

★ 数据是可以记录、通信和能识别的符号，它通过有意义的组合来表达现实世界中的某种实体（具体对象、事件、状态或活动）的特征。

年龄	收入(万)	工作时间(年)	顾客类型
25	10	2	优
29	12	3	优
32	9	6	良
38	7	12	良
36	18	13	中
30	15	4	优
...

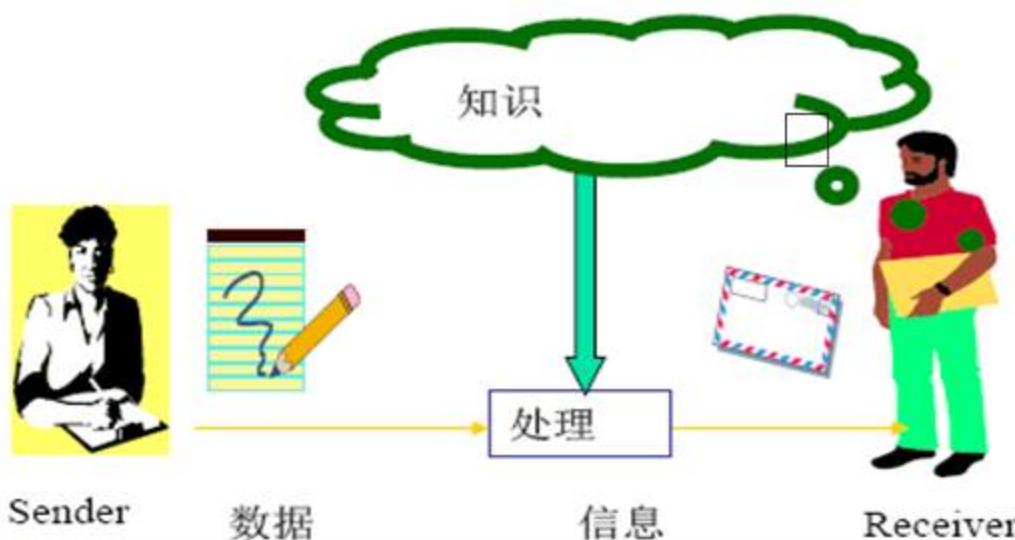
数据



信息：经过提炼、加工和解释的数据

► 信息

★ 美军《野战条令(FM64)》：信息是对数据经过过滤、融合、标准化、对比、翻译、分类、管理等一系列环节处理后得到的。



知识：结构化，有价值的信息

➤ 知识

★ 知识是对信息内容进行提炼、比较、挖掘、分析、概括、判断和推论得到的。



25<年龄<30， 收入>10， 工作时间>2年的消费者是优质顾客



“8,000” 和 “10,000” 表示

- A 数据
- B 信息
- C 知识
- D 智慧

提交

“8,000米是飞机飞行最大高度”与“10,000米的高山”表示

- A 数据
- B 信息
- C 知识
- D 智慧

提交

“飞机无法飞越这座高山” 表示

- A 数据
- B 信息
- C 知识
- D 智慧

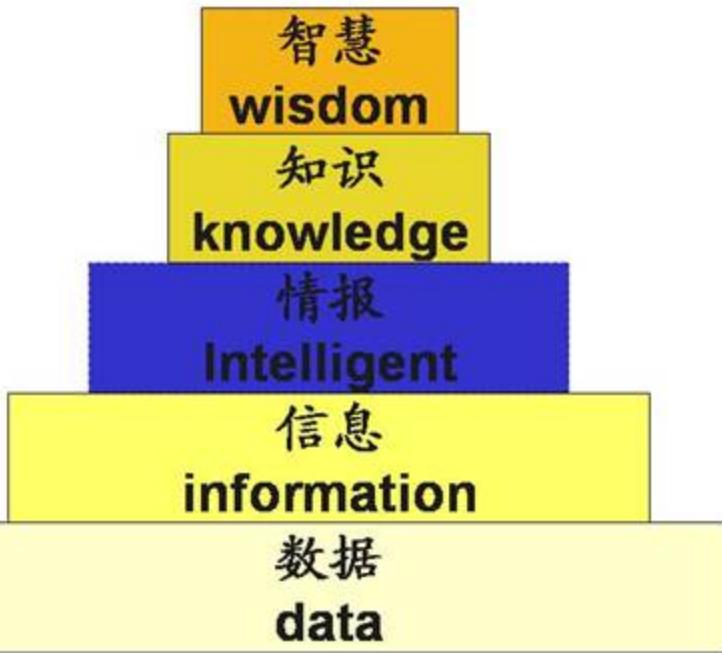
提交

“飞机必须飞得比山高”表示

- A 数据
- B 信息
- C 知识
- D 智慧

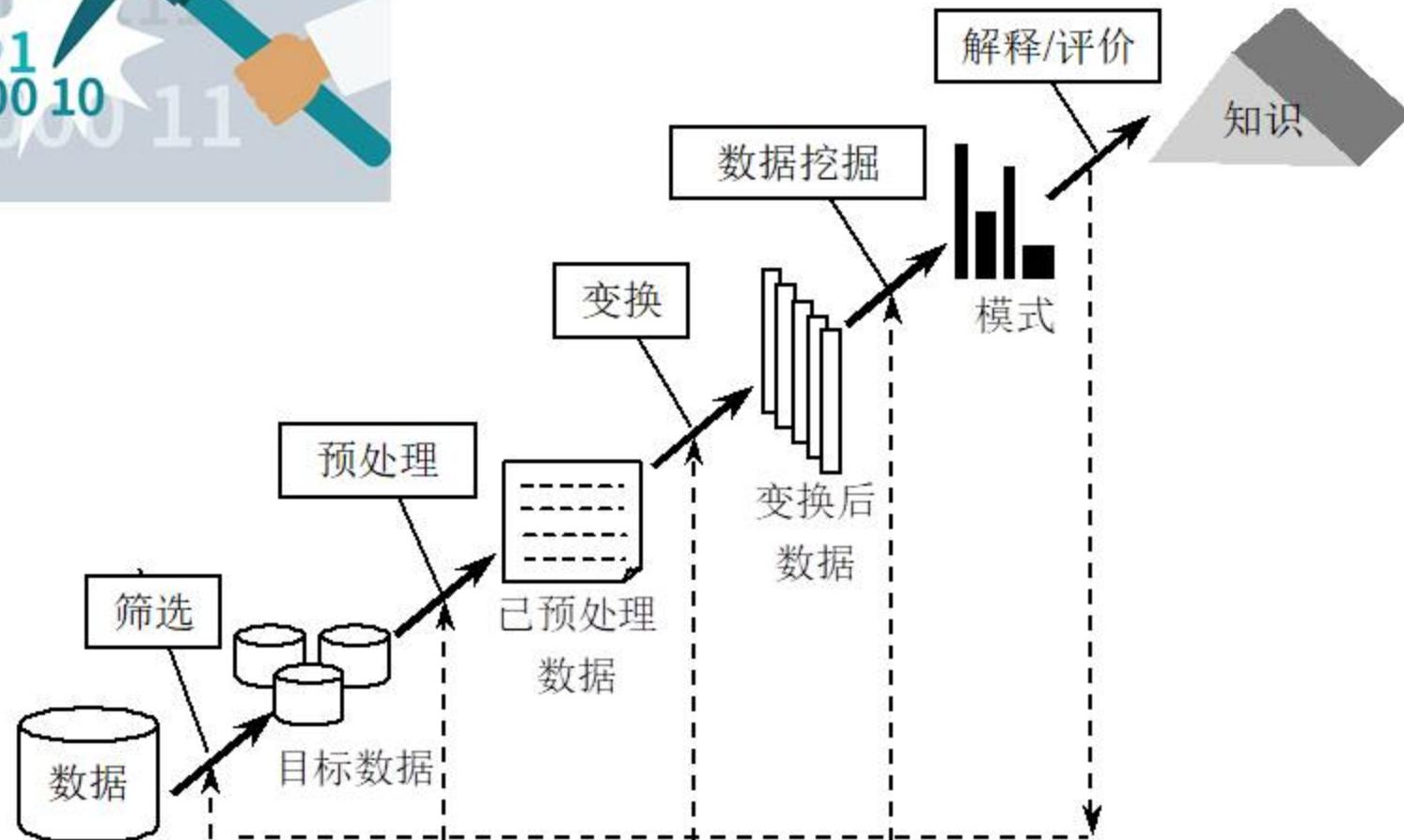
提交

数据、信息、知识



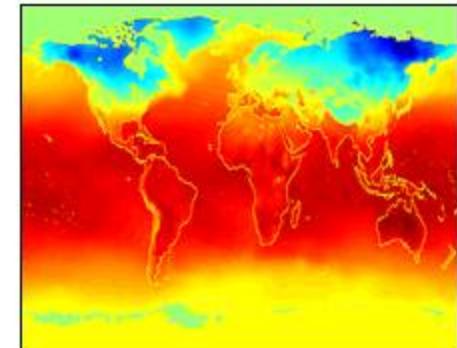
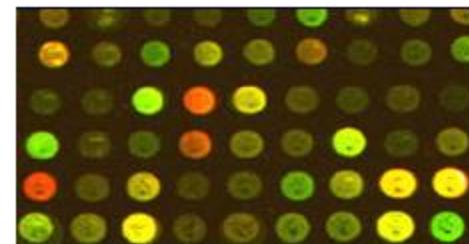
- “8,000”和“10,000”是数据；
- “8,000米是飞机飞行最大高度”与“10,000米的高山”是信息；
- “飞机无法飞越这座高山”是知识；
- “飞机必须飞得比山高”是智慧。

数据挖掘过程

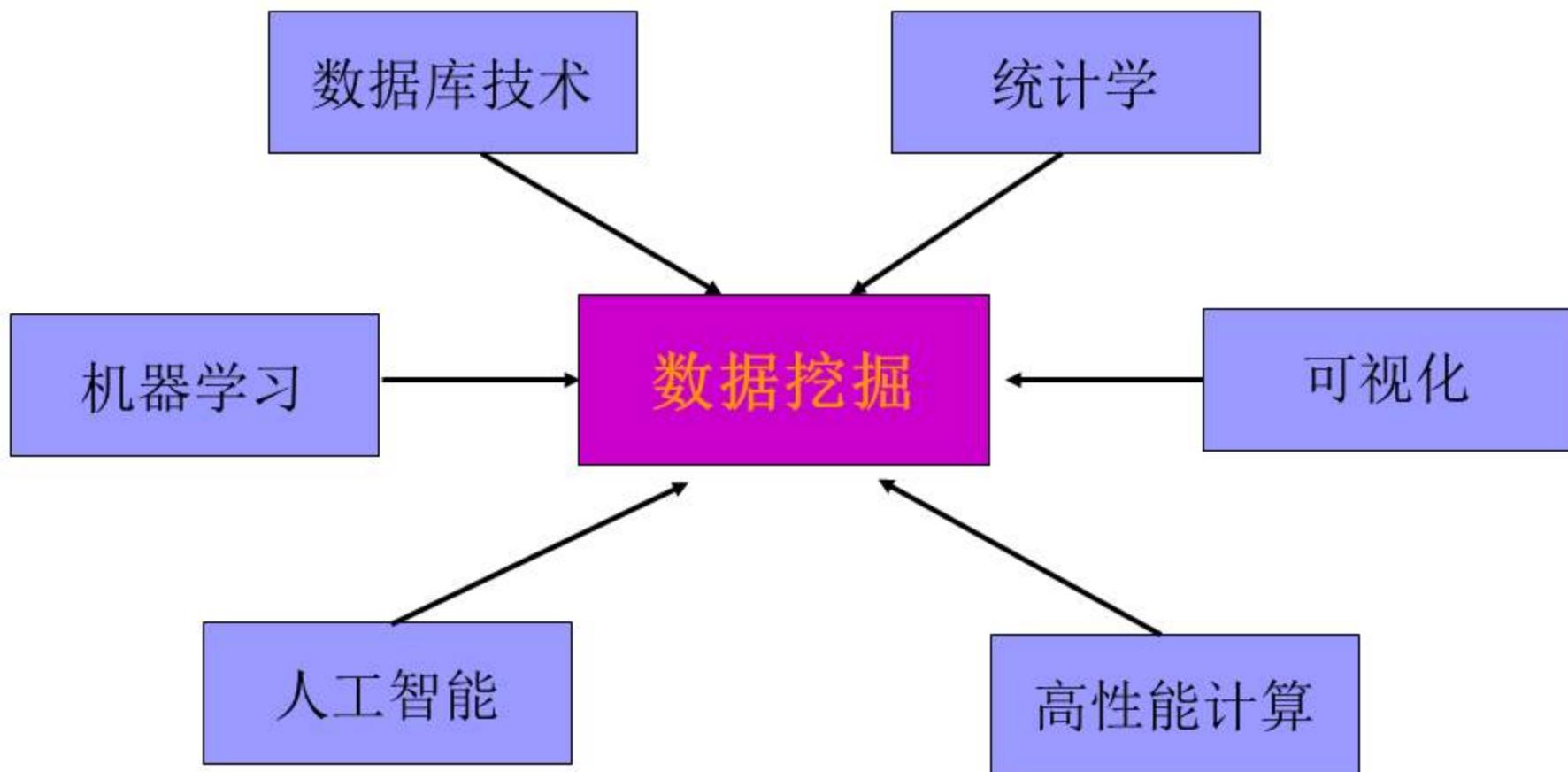


数据来源

- 关系数据库
- 数据仓库
- 事务数据库
- 空间、时间数据库
- 文本和多媒体数据（异构的）
- 各种结构化、半结构化的数据
- 互联网、移动互联网数据源
-



数据挖掘是多学科交叉的产物





学过课程调查

A

统计学

B

数据库

C

机器学习

D

Python编程

E

算法复杂性

提交

III. 数据挖掘的主要研究内容

- 关联规则挖掘
- 非监督式机器学习-聚类
- 监督式机器学习
 - ✓ 离散标签预测-标签分类
 - ✓ 连续标签预测-数值预测
- 回归





基础掌握程度了解

A

关联规则挖掘概念前期已掌握

B

非监督式机器学习概念前期已掌握

C

监督式机器学习概念前期已掌握

D

回归概念前期已掌握

提交



是否听说过啤酒与尿布的故事

A

是

B

否

提交

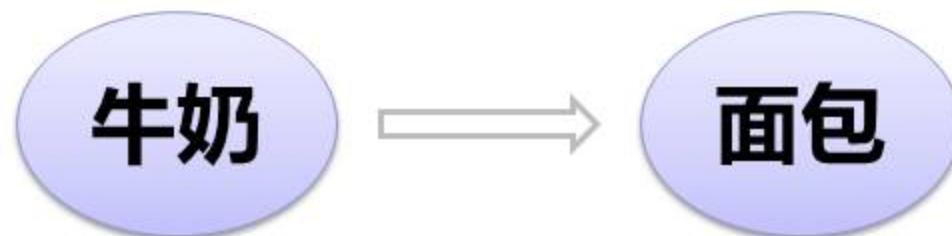
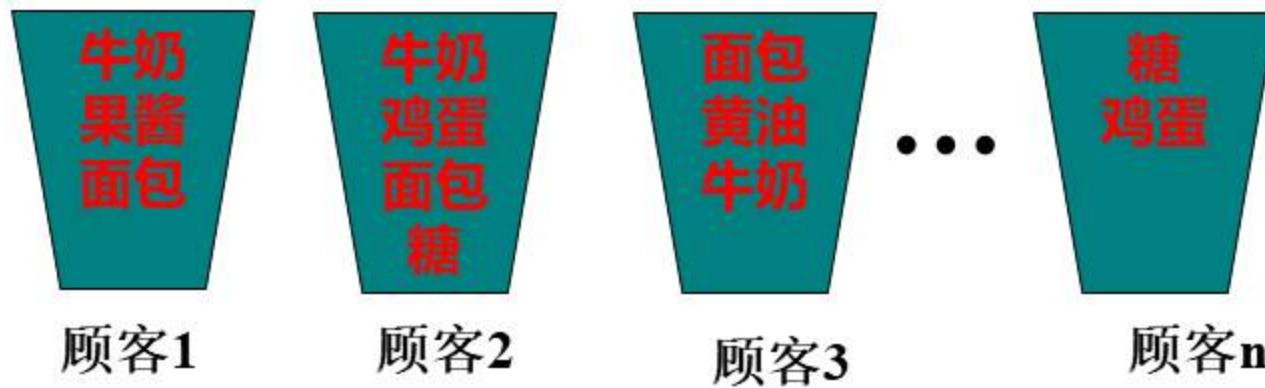
关联规则挖掘

[Descriptive]



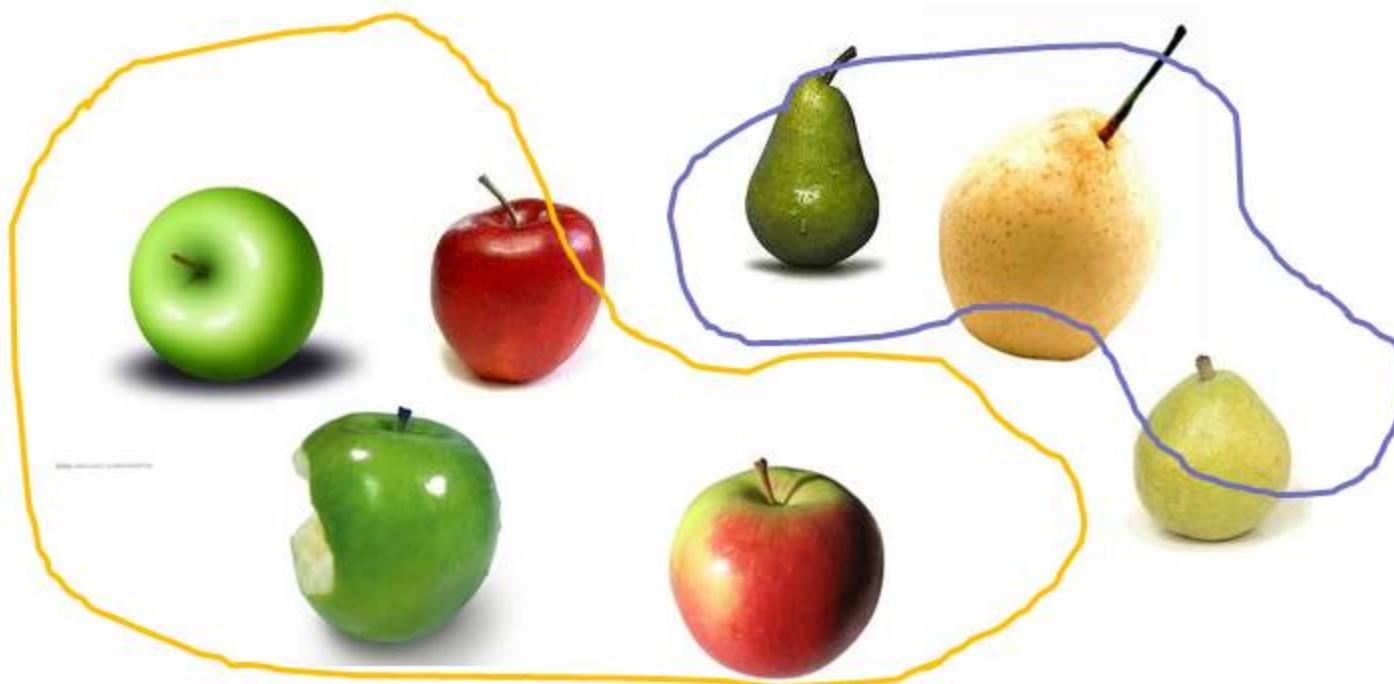
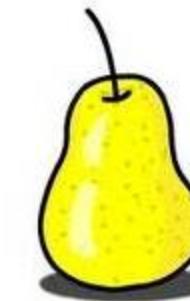
关联规则挖掘

[Descriptive]

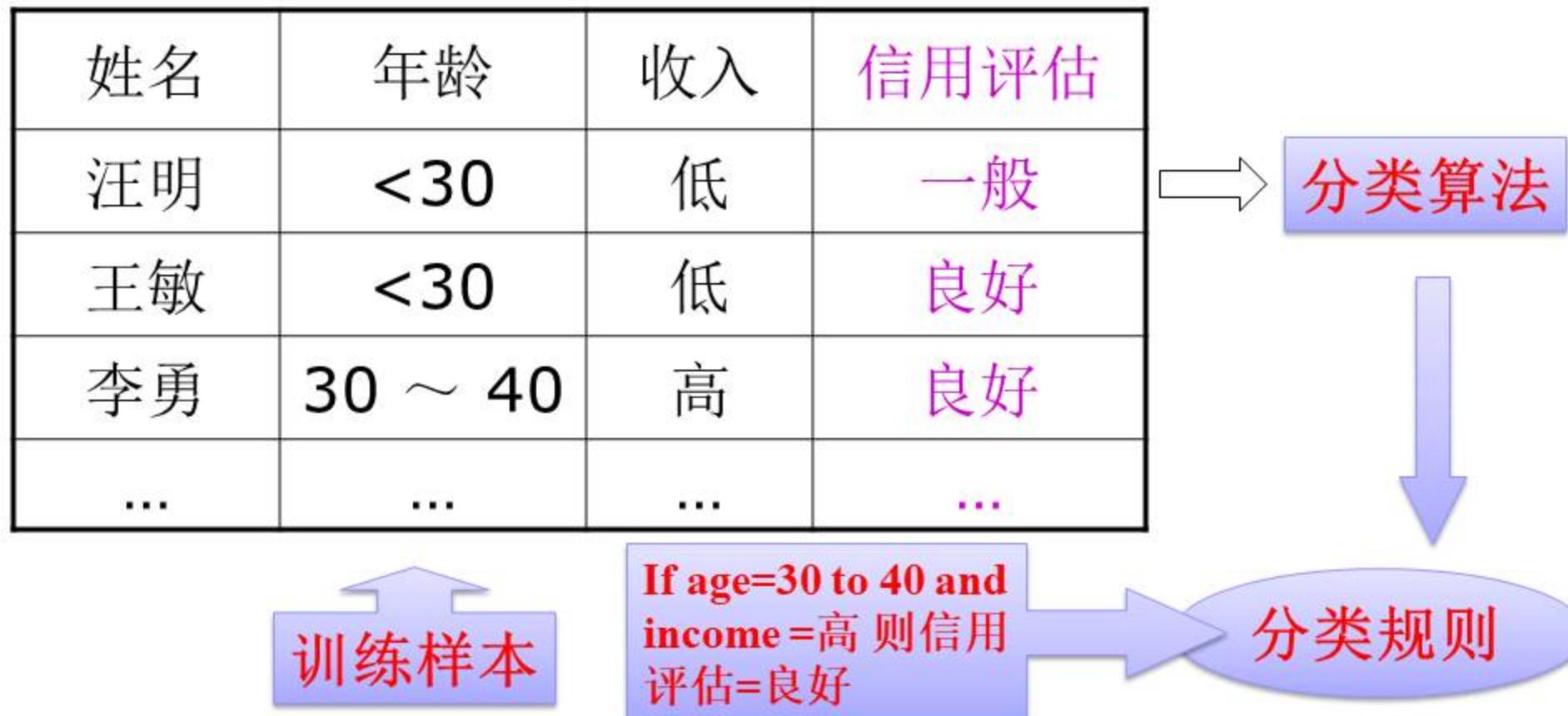


分类

[Predictive]



分类分析——第一步：学习建模



分类分析——第二步：分类测试

姓名	年龄	收入	信用评估
张丰	>40	高	?
王敏	<30	低	?
李勇	30 ~ 40	高	?
...

测试数据

良

分类
规则

新数据：李勇，
30 ~ 40，收入
高，信用评估如
何？

数值预测——第一步：学习建模

姓名	年龄	收入	信用值
汪明	<30	低	65
王敏	<30	低	74
李勇	30 ~ 40	高	78
...

训练样本

数值预测——第二步：预测测试

姓名	年龄	收入	信用值
张丰	>40	高	?
王敏	<30	低	?
李勇	30 ~ 40	高	?
...

测试数据

分类与数值预测案例1（1/2）

■ 员工离职预测

- 从给定的影响员工离职的因素和员工是否离职的记录，建立一个模型预测有可能离职的员工。

属性	说 明
Age	年龄
Attrition	是否已经离职， 0：离职， 1：未离职
BusinessTravel	商务差旅频率
DistanceFromHome	员工所在部门
Education	员工的教育程度，从1到5，5表示教育程度最高
.....
YearsWithCurrManager	跟目前的管理者共事年数

分类与数值预测案例2 (2/2)

■ 房价预测

□ 房屋销售价格以及房屋的基本信息建立模型，来预测在此期间其他房屋的销售价格

- 销售日期、销售价格、卧室数、浴室数、房屋面积、停车面积、楼层数、房屋评分、建筑面积、地下室面积、建筑年份、修复年份、纬度、经度

20150302	545000	3	2.25	1670	6240	1	8	1240	430	1974	0	47.6413	-122.113
20150211	785000	4	2.5	3300	10514	2	10	3300	0	1984	0	47.6323	-122.036
20150107	765000	3	3.25	3190	5283	2	9	3190	0	2007	0	47.5534	-122.002
20141103	720000	5	2.5	2900	9525	2	9	2900	0	1989	0	47.5442	-122.138
20140603	449500	5	2.75	2040	7488	1	7	1200	840	1969	0	47.7289	-122.172
20150506	248500	2	1	780	10064	1	7	780	0	1958	0	47.4913	-122.318
20150305	675000	4	2.5	1770	9858	1	8	1770	0	1971	0	47.7382	-122.287
20140701	730000	2	2.25	2130	4920	1.5	7	1530	600	1941	0	47.573	-122.409
20140807	311000	2	1	860	3300	1	6	860	0	1903	0	47.5496	-122.279
20141204	660000	2	1	960	6263	1	6	960	0	1942	0	47.6646	-122.202
20150227	435000	2	1	990	5643	1	7	870	120	1947	0	47.6802	-122.298
20140904	350000	3	1	1240	10800	1	7	1240	0	1959	0	47.5233	-122.185
20140902	385000	3	2.25	1630	1598	3	8	1630	0	2008	0	47.6904	-122.347
20150413	235000	2	1	930	10505	1	6	930	0	1930	0	47.4337	-122.329
20140930	350000	3	1	1300	10236	1	6	1300	0	1971	0	47.5028	-121.77
20150507	1350000	4	1.75	2000	3728	1.5	9	1820	180	1926	0	47.643	-122.299
20140530	459900	3	1.75	2580	11000	1	7	1290	1290	1951	0	47.5646	-122.181
20140723	430000	6	3	2630	8800	1	7	1610	1020	1959	0	47.7166	-122.293
20141003	718000	5	2.75	2930	7663	2	9	2930	0	2013	0	47.5308	-122.184

案例1和案例2分别是什么问题

- A 案例1：标签类别预测，案例2：标签类别预测
- B 案例1：数值预测，案例2：数值预测
- C 案例1：标签类别预测，案例2：数值预测
- D 案例1：数值预测，案例2：标签类别预测

提交

分类与数值预测案例1（1/2）

■ 员工离职预测

- 从给定的影响员工离职的因素和员工是否离职的记录，建立一个模型预测有可能离职的员工。

属性	说 明
Age	年龄
Attrition	是否已经离职，0：离职，1：未离职
BusinessTravel	商务差旅频率
DistanceFromHome	员工所在部门
Education	员工的教育程度，从1到5，5表示教育程度最高
.....
YearsWithCurrManager	跟目前的管理者共事年数

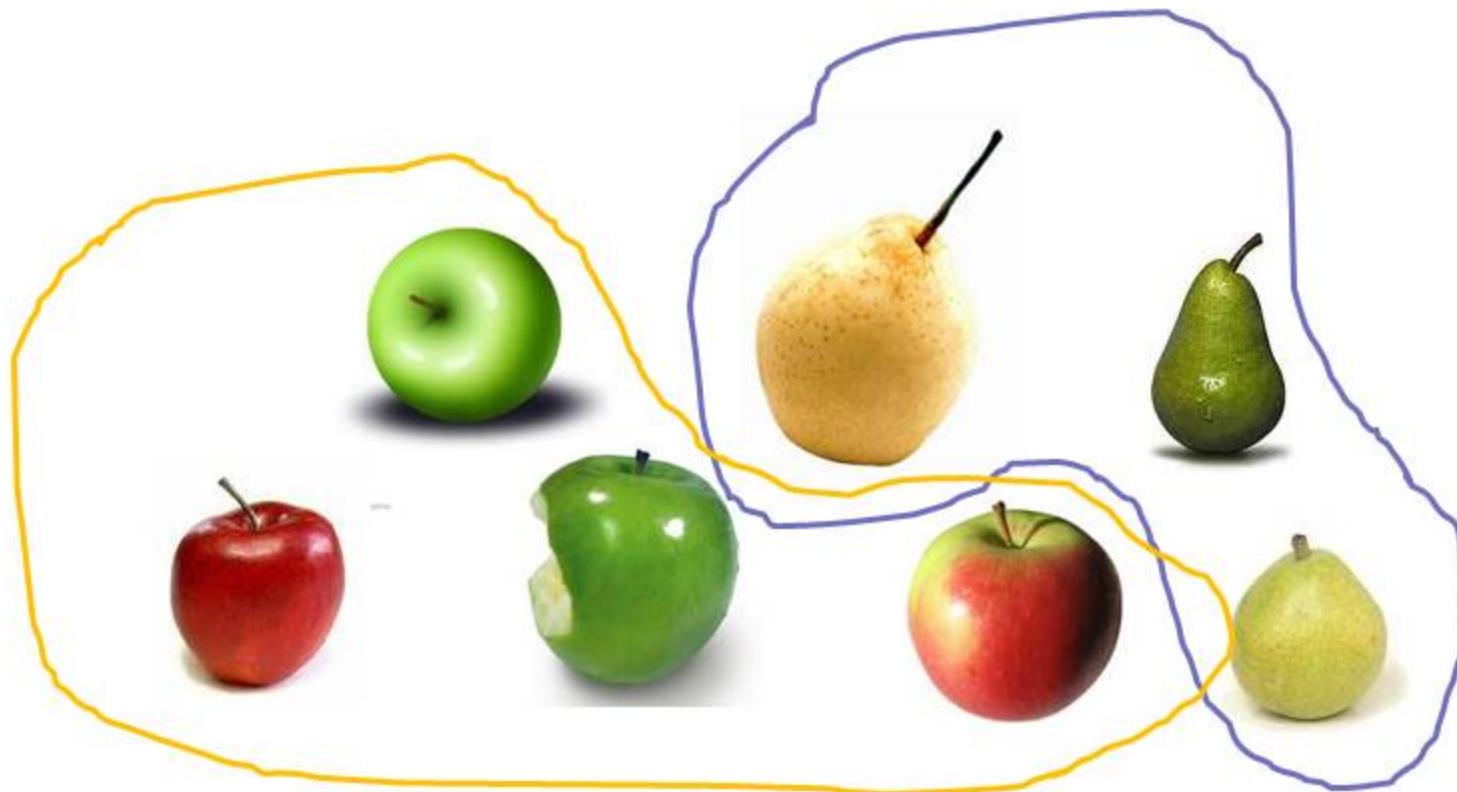
分类与数值预测案例2 (2/2)

■ 房价预测

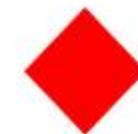
□ 房屋销售价格以及房屋的基本信息建立模型，来预测在此期间其他房屋的销售价格

- 销售日期、销售价格、卧室数、浴室数、房屋面积、停车面积、楼层数、房屋评分、建筑面积、地下室面积、建筑年份、修复年份、纬度、经度

20150302	545000	3	2.25	1670	6240	1	8	1240	430	1974	0	47.6413	-122.113
20150211	785000	4	2.5	3300	10514	2	10	3300	0	1984	0	47.6323	-122.036
20150107	765000	3	3.25	3190	5283	2	9	3190	0	2007	0	47.5534	-122.002
20141103	720000	5	2.5	2900	9525	2	9	2900	0	1989	0	47.5442	-122.138
20140603	449500	5	2.75	2040	7488	1	7	1200	840	1969	0	47.7289	-122.172
20150506	248500	2	1	780	10064	1	7	780	0	1958	0	47.4913	-122.318
20150305	675000	4	2.5	1770	9858	1	8	1770	0	1971	0	47.7382	-122.287
20140701	730000	2	2.25	2130	4920	1.5	7	1530	600	1941	0	47.573	-122.409
20140807	311000	2	1	860	3300	1	6	860	0	1903	0	47.5496	-122.279
20141204	660000	2	1	960	6263	1	6	960	0	1942	0	47.6646	-122.202
20150227	435000	2	1	990	5643	1	7	870	120	1947	0	47.6802	-122.298
20140904	350000	3	1	1240	10800	1	7	1240	0	1959	0	47.5233	-122.185
20140902	385000	3	2.25	1630	1598	3	8	1630	0	2008	0	47.6904	-122.347
20150413	235000	2	1	930	10505	1	6	930	0	1930	0	47.4337	-122.329
20140930	350000	3	1	1300	10236	1	6	1300	0	1971	0	47.5028	-121.77
20150507	1350000	4	1.75	2000	3728	1.5	9	1820	180	1926	0	47.643	-122.299
20140530	459900	3	1.75	2580	11000	1	7	1290	1290	1951	0	47.5646	-122.181
20140723	430000	6	3	2630	8800	1	7	1610	1020	1959	0	47.7166	-122.293
20141003	718000	5	2.75	2930	7663	2	9	2930	0	2013	0	47.5308	-122.184



聚类分析原理介绍



A

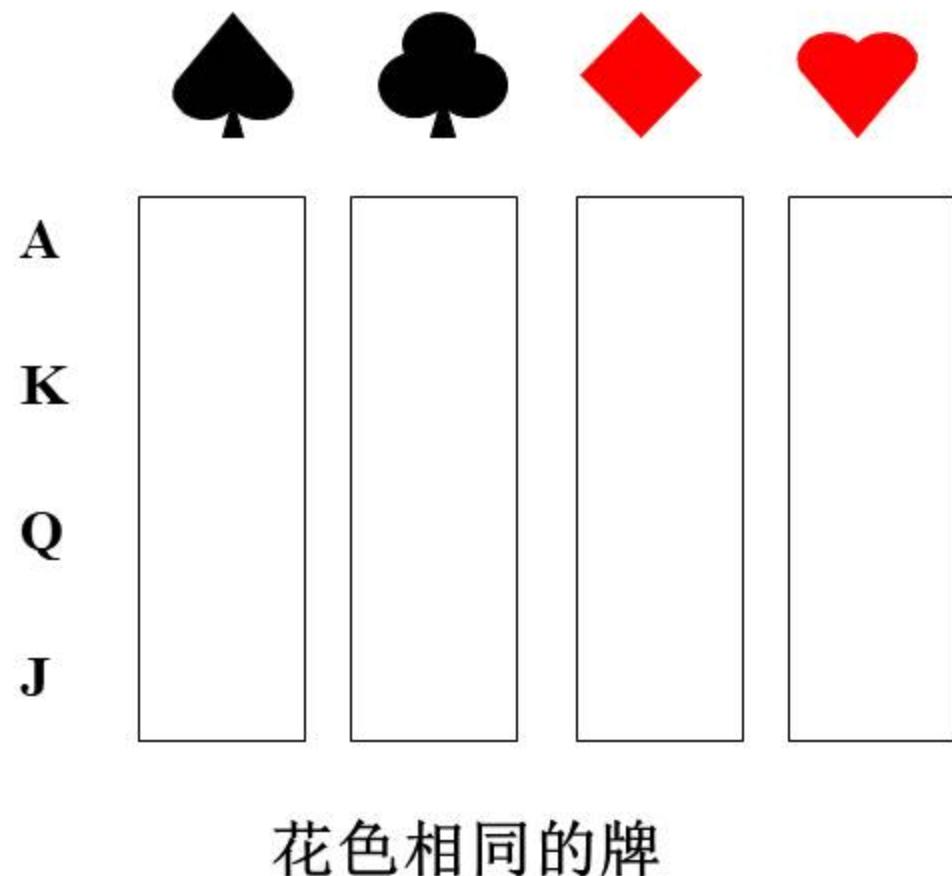
K

Q

J

聚类分析原理介绍

- 分成四组
- 每组里花色相同
- 组与组之间花色相异



聚类分析原理介绍

- 分成四组
- 符号相同的牌为一组



A

K

Q

J

符号相同的牌

人从出生到长大的过程中，是如何认识事物的？

- A 先分类，后聚类
- B 先聚类，后分类

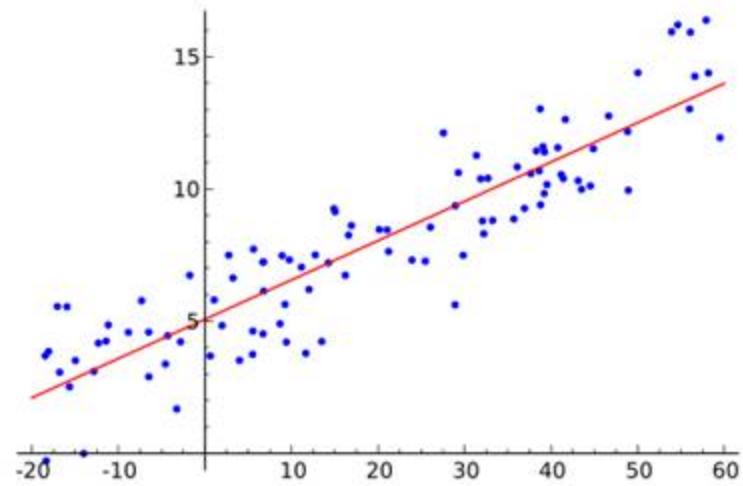
提交

$$Y = f(X, \beta)$$

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

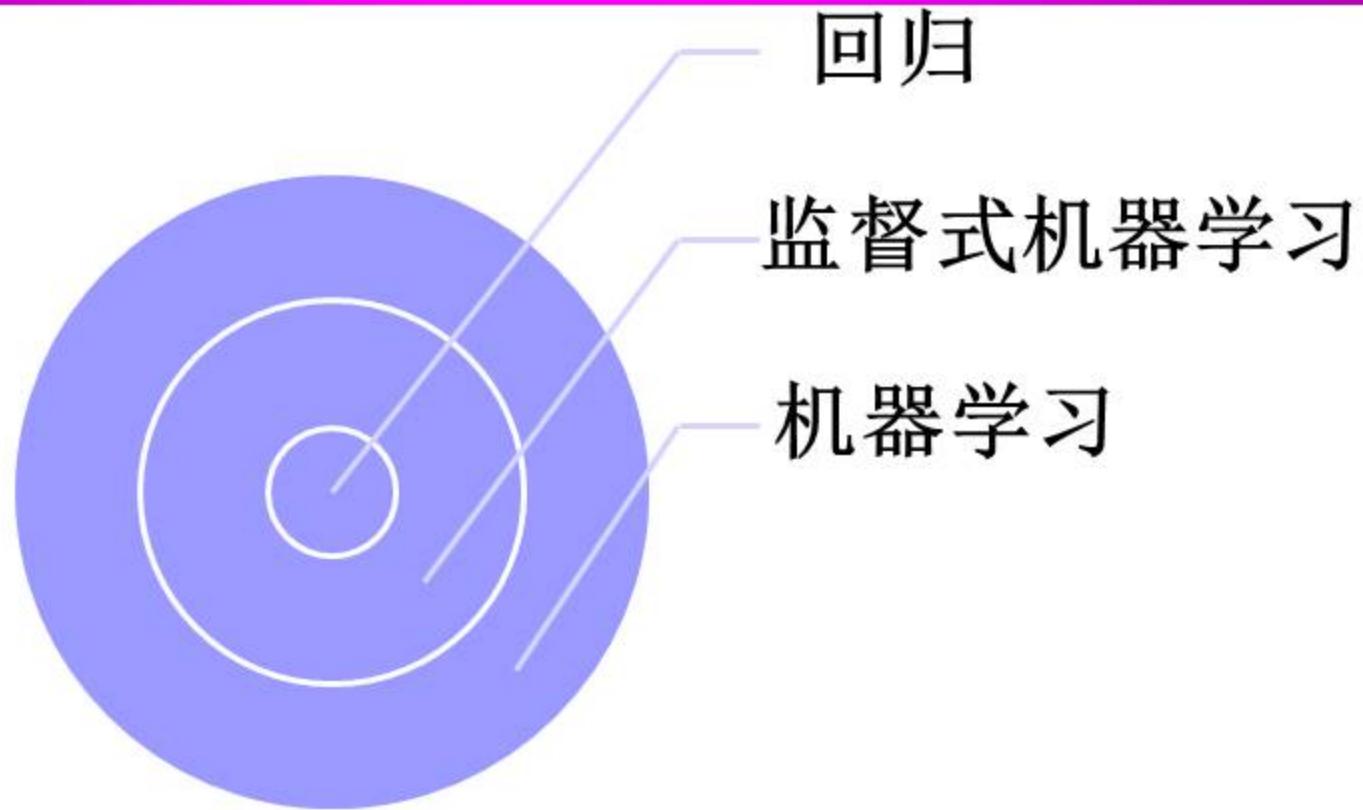


$$y = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

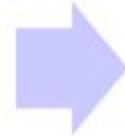
回归分析：建立多个变量之间的定量关系

回归

[Predictive]



回归



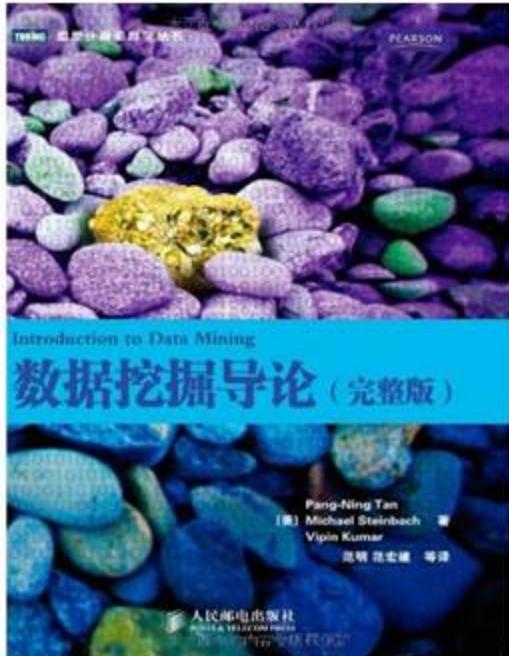
数值预测

III. 数据挖掘的主要研究内容

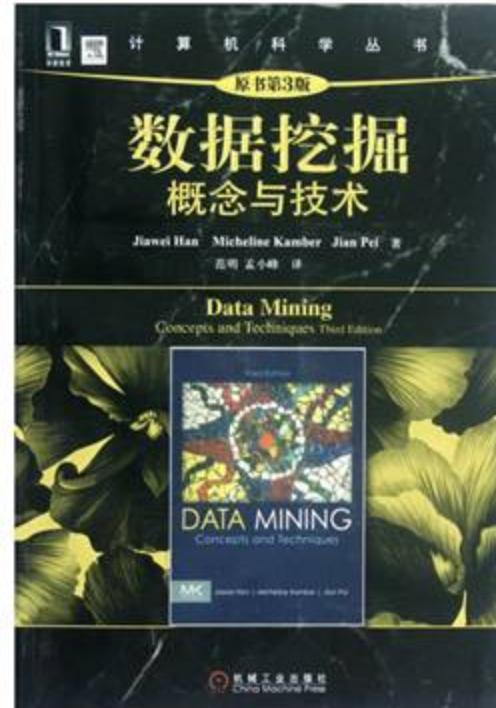
- 关联规则挖掘
- 非监督式机器学习-聚类
- 监督式机器学习
 - ✓ 离散标签预测-标签分类
 - ✓ 连续标签预测-数值预测
- 回归



IV. 主要参考资料-教材及参考书



教材



参考书

IV. 课堂内容安排

第一章 数据挖掘绪论	第1-2节
第二章 认识数据	第3-6节
第三章 数据预处理	第7-12节
第四章 关联规则挖掘	第13-16节
第五章 聚类基础	第17-22节
第一次学生作业汇报	第23-24节
第六章 分类基础	第25-30节
第七章 支持向量机	第31-32节
第八章 回归分类	第33-34节
第九章 神经网络分类	第35-36节
第十章 集成学习	第37-38节
第十一章 离群点检测	第39-40节
学生实验与汇报	第41-48节

IV. 主要参考资料-我的GitHub资源

The screenshot shows a GitHub repository page for 'zyding1983 / datamining'. The top navigation bar includes links for Watch (0), Star (0), Code, Issues (0), Pull requests (0), Actions, Projects (0), Wiki, Security, Insights, and Settings. The main content area shows a commit from 'zyding1983' adding 'ne11' to several files: 'books', 'codes', 'datasets', 'ppt', and '.DS_Store'. The commit message is 'add ne11'. The repository path is 'datamining / 2020A /'. There are buttons for Create new file, Upload files, and Find.

Branch: master ▾ [datamining / 2020A /](#) Create new file Upload files Find

zyding1983 add ne11 Latest commit 305e61

..

books add new classes

codes add ne11

datasets add ne11

ppt add ne11

.DS_Store add ne11

<https://github.com/zyding1983/datamining>

IV. 主要参考资料-学生众包提问反馈

dm学生提问反馈区 自动保存成功

编辑 插入 格式 公式 数据 视图 表单 帮助

常规 字体

A	B	C	D	E	F	G	H	I	J
序号	问题	提问学生学号	提问学生专业	回复问题	回复学生学号	纠正问题	纠正学生学号	教辅回复	老师回复
1	1 有什么问题，可以	111111	管科	这里也可以插入	1111		1111	问题很好	不错
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									

在里面提1问，上课答题成绩中计1分

在里面回复1个问题，上课答题成绩中计2分

在里面纠正1个回复问题，上课答题成绩中计5分

IV. 主要参考资料-网络课程资源 (1/2)

The screenshot shows the Xuetangx website interface. At the top, there is a navigation bar with the logo '学堂在线 xuetaongx.com', followed by links for '首页', '课程', '院校', '微学位', '学堂云', '雨课堂', a search bar containing '请输入课程、老师、学校', a magnifying glass icon, and buttons for 'APP下载', '注册', and '登录'. Below the header, the course title '数据挖掘：理论与算法（自主模式）' is displayed, along with '自主模式' and '国家级精品' badges. It indicates the course is from Tsinghua University and belongs to the Computer category (667). A video thumbnail for the course is shown, featuring a male professor in a white shirt and glasses standing in front of a digital background with numbers and text. The video player shows a play button, a timestamp of '0:00 / 2:00', and a volume icon. To the right of the video, there is a '课程描述' section with the text '最有趣的理论+最有用的算法=不得不学的数据科学'. Below this are several course details: '开课时间: 2017.01.18 08:00', '结课时间: 永久开课'; '学习时长: 8小时/周', '课程进度: 连续至第12讲'; '报名人数: 7.4万人', '先修知识: 概率统计、线性代数、数据...'. Two buttons are present: a white '免费学习' button and a blue '认证学习' button. Below these buttons is a question '什么是认证证书?'. At the bottom of the page, there are five navigation links: '课程内容' (underlined), '授课教师', '精华笔记', '常见问题', and '相关课程'.

<http://www.xuetangx.com/courses/course-v1:TsinghuaX+80240372X+sp/about>

IV. 主要参考资料-网络课程资源 (2/2)

MOOC中国

搜寻课程



数据挖掘 专项课程

Data Mining

Analyze Text, Discover Patterns, Visualize Data. Solve real-world data mining challenges.

伊利诺伊大学香槟分校 Coursera

计算机

专项课程

数据科学

黑五

普通 (中级)

6 个月

课程概况

The Data Mining Specialization teaches data mining techniques for both structured data which conform to a clearly defined schema, and unstructured data which exist in the form of natural language text. Specific course topics include pattern discovery, clustering, text retrieval, text mining and analytics, and data visualization. The Capstone project task is to solve real-world data mining challenges using a restaurant review data set from Yelp.

Courses 2 – 5 of this Specialization form the lecture component of courses in the online Master of Computer Science Degree in Data Science. You can apply to the degree program either before or after you begin the Specialization.

<https://www.mooc.cn/course/6784.html>

IV. 主要参考资料-实践网络资源educoder

https://www.educoder.net/search?value=数据挖掘_NUIDT_2020本科



点击翻转课堂

共找到相关结果1个

数据挖掘_NUIDT_2020本科

zyding 国防科技大学 成员数: 1

- (1) 需要注册
- (2) 注册登录后点击搜到的课程
- (3) 加入课程需要邀请码: **63UM4**
- (4) 以后课堂、课后编程作业都利用该平台
- (5) 请同学们**2月20日**之前完成注册, 各专业课代表统计注册情况反馈给我

IV. 主要参考资料-数据挖掘python代码资源

<https://scikit-learn.org>

The screenshot shows the official website for scikit-learn. At the top, there's a navigation bar with links for 'Install', 'User Guide', 'API', 'Examples', and 'More'. Below the header, the main title 'scikit-learn' is displayed in large letters, followed by the subtitle 'Machine Learning in Python'. A horizontal menu bar at the bottom includes 'Getting Started', 'What's New in 0.22.1', and 'GitHub'. On the right side of the page, there's a list of bullet points highlighting the library's features:

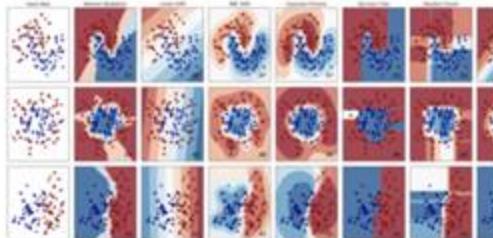
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...

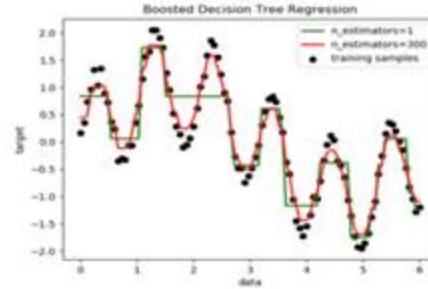


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...





Python编程基础了解

A

了解Anaconda，且动手安装过

B

了解pycharm，且动手安装过

C

有python编程基础

提交

IV. 主要参考资料-Anaconda下载安装

<https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/>

Anaconda3-5.3.1-Linux-x86.sh	527.3 MiB	2018-11-20 04:00
Anaconda3-5.3.1-Linux-x86_64.sh	637.0 MiB	2018-11-20 04:00
Anaconda3-5.3.1-MacOSX-x86_64.pkg	634.0 MiB	2018-11-20 04:00
Anaconda3-5.3.1-MacOSX-x86_64.sh	543.7 MiB	2018-11-20 04:01
Anaconda3-5.3.1-Windows-x86.exe	509.5 MiB	2018-11-20 04:04
Anaconda3-5.3.1-Windows-x86_64.exe	632.5 MiB	2018-11-20 04:04

根据自己电脑操作系统下载对应版本安装

IV. 主要参考资料-pycharm下载安装

<https://www.jetbrains.com/pycharm/download>



Version: 2019.3.3

Build: 193.6494.30

7 February 2020

[System requirements](#)

[Installation Instructions](#)

[Other versions](#)

Download PyCharm

[Windows](#) [Mac](#) [Linux](#)

Professional

For both Scientific and Web Python development. With HTML, JS, and SQL support.

[Download](#)

[Free trial](#)

Community

For pure Python development

[Download](#)

[Free, open-source](#)

注意下载**Community**版本

IV. 主要参考资料-pycharm环境配置



https://blog.csdn.net/ling_mochen/article/details/79314118

以前没有安装过**python**环境的同学请课后完成，有问题微信群中咨询教辅

没有**python**编程基础的请课后自学简单的**python**基础

IV. 主要参考资料-第1次课后作业

- 1、安装且配置好python环境
- 2、每个同学从网上下载“新型肺炎”至少3个省份（湖北、家乡、湖南）每天（从1月20日起）确诊病例、新增病例、疑似病例、疑似新增病例、治愈病例、治愈新增病例、死亡病例、死亡新增病例（要求提交数据excel文件给教辅负责人，**平时课后成绩计40分**）
- 3、利用python分别画出3个省份的8个字段的变化趋势【3个省份对比图、确诊和疑似的对比图、等等】（提交带图的word文件给教辅负责人，**平时课后成绩计40分**）
- 4、利用回归方法、神经网络等方法拟合3个省份的确诊病例变化趋势，预测未来7天的确诊病例数目（提交代码、预测图给教辅负责人，**平时课后成绩计20分**）

2月20日之前提交作业给对应教辅负责人

<https://ncov.dxy.cn/ncov/h5/view/pneumonia?scene=2&clicktime=1579582238&enterid=1579582238&from=singlemessage&isappinstalled=0>

IV. 主要参考资料-数据挖掘工具

<https://rapidminer.com>



WHY RAPIDMINER INDUSTRIES PRODUCTS LEARN RESOURCES PARTNERS COMPANY

RapidMiner
Exceeds Your
Expectations

Gartner

RapidMiner a Leader in the 2019 Gartner Magic Quadrant for Data Science and Machine Learning Platforms for the sixth year in a row

[Read the Report](#)

FORRESTER®

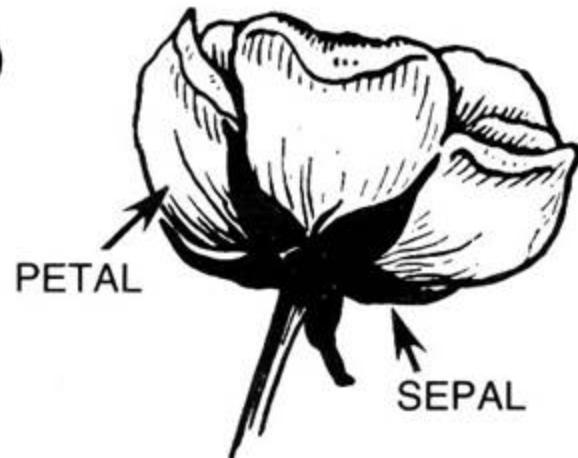
RapidMiner a Leader in the 2018 Forrester Wave Multimodel Predictive Analytics & Machine Learning Solutions for the second year in a row

rapidminer

IV. 主要参考资料-实验基础数据集

■ IRIS (sepal:萼片,petal:花瓣)

- 1. sepal length in cm
- 2. sepal width in cm
- 3. petal length in cm
- 4. petal width in cm
- 5. class:
 - -- Iris Setosa
 - -- Iris Versicolour
 - -- Iris Virginica



```
6.7,3.0,5.2,2.3,Iris-virginica  
6.3,2.5,5.0,1.9,Iris-virginica  
6.5,3.0,5.2,2.0,Iris-virginica  
6.2,3.4,5.4,2.3,Iris-virginica  
5.9,3.0,5.1,1.8,Iris-virginica  
5.1,3.8,1.6,0.2,Iris-setosa  
4.6,3.2,1.4,0.2,Iris-setosa  
5.3,3.7,1.5,0.2,Iris-setosa  
5.0,3.3,1.4,0.2,Iris-setosa  
7.0,3.2,4.7,1.4,Iris-versicolor  
6.4,3.2,4.5,1.5,Iris-versicolor  
6.9,3.1,4.9,1.5,Iris-versicolor  
5.5,2.3,4.0,1.3,Iris-versicolor
```

请同学们2月20日之前下载该数据集，我的GitHub仓库中

V. 课程要求-成绩组成

● 成绩组成

■ 考试成绩：60

■ 平时成绩：40

✓1、上课答题成绩（主要来自共享文档提问反馈和
上课答题）：5

✓2、课后作业成绩（主要来自educoder）：5

✓3、综合大作业：30

平时成绩评定方法（1、2单项）：全班平时成绩每单项的最高分作为5分，其他学生成绩分数以每项最高分作为分母，在此基础上，求分值比例r（每个学生成绩除以最高分），然后平时成绩单项分值以 $5 \times r$ 来衡量

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台

✓1、阿里的天池大数据竞赛【朱席席】

✓2、kaggle平台竞赛【刘凯】

✓3、DataCastle大数据竞赛平台

<http://www.pkbidata.com/> 【徐翔】

✓4、数泉竞赛平台(datafountain.cn)【王庆勇】

✓5、科赛 (<https://www.kesci.com/apps/home/competition>)
【王庆勇】

✓6、GitHub数据挖掘类源代码分析

V. 课程要求-综合大作业-阿里天池大数据竞赛

Active 算法大赛 新品实验室 NEW 创新应用大赛 程序设计大赛 新人赛 诸神之战

【入门】Docker练习场 新人赛

赛事简介：这是为Docker学习者开设的练习场。Docker技术，有人说是“集装箱”，有人说是“造梦空间”，有人说是“胶囊式公寓”，TA到底是什么呢？

主办方：Alibaba Cloud TIANCHI 天池

奖金 ￥0 团队 651 赛季 1 2020-12-31

【长期赛】安全AI挑战者计划第一期 - 人脸识别对抗 新人赛

赛事简介：AI安全性有诸多挑战，为了抵御未来AI面临的安全风险，阿里安全联合清华大学，以对抗样本为核心，假想未来作为安全AI防守者的身份，结合内容安全场景，从文字、图像、视频、声音...

主办方：清华大学 阿里安全

奖金 ￥0 团队 1148 赛季 2 2020-03-31

【追风少年】台风图像时间序列预测 新人赛

赛事简介：我们为什么要预测台风？因为台风就在那里。通过气象卫星图像预测台风的发展强度、行进轨迹、乃至于降水在各地的实时详细分布，为台风预警与防灾贡献一份力量。

主办方：TIANCHI 天池

奖金 ￥0 团队 924 赛季 3 2020-03-31

<https://tianchi.aliyun.com/competition/gameList/coupleList>
比赛数目：18（新人赛）

V. 课程要求-综合大作业-kaggle竞赛

11 Active Competitions			
	Deepfake Detection Challenge Identify videos with facial or voice manipulations <small>Featured · Code Competition · 2 months to go · 📹 video data, online video</small>	\$1,000,000 1,579 teams	
	Google QUEST Q&A Labeling Improving automated understanding of complex question answer content <small>Featured · Code Competition · a day to go · 📖 text data, nlp</small>	\$25,000 1,552 teams	
	Real or Not? NLP with Disaster Tweets Predict which Tweets are about real disasters and which ones are not <small>Getting Started · Ongoing · 📖 text data, binary classification</small>	\$10,000 2,629 teams	
	Bengali.AI Handwritten Grapheme Classification Classify the components of handwritten Bengali <small>Research · Code Competition · a month to go · 📷 multiclass classification, image data</small>	\$10,000 1,178 teams	
	Digit Recognizer Learn computer vision fundamentals with the famous MNIST data <small>Getting Started · Ongoing · 📖 tabular data, multiclass classification, image data, object identification</small>	Knowledge 2,385 teams	
	Titanic: Machine Learning from Disaster Start here! Predict survival on the Titanic and get familiar with ML basics <small>Getting Started · Ongoing · 📖 binary classification, tabular data, tutorial</small>	Knowledge 15,973 teams	

<https://www.kaggle.com/competitions>

比赛数目： 11（活跃赛）

V. 课程要求-综合大作业-DataCastle竞赛

进行中



城市交通流量时空预测

算法竞赛

山东省大数据局、青岛市大数据发展管理局

为深入贯彻落实习近平总书记视察山东重要讲话、重要指示批示精神，加快实施《数字山东发展规划（2018—2022年）》，推进数字山东建设，培育富有活力的数字经济，由山东省大数据局主办，青岛市大数据发展管理....

人工智能

时间：2019/09/23–2020/06/20

参赛人数：851



进行中



识别失信企业大赛

算法竞赛

山东省大数据局、青岛市大数据发展管理局

为深入贯彻落实习近平总书记视察山东重要讲话、重要指示批示精神，加快实施《数字山东发展规划（2018—2022年）》，推进数字山东建设，培育富有活力的数字经济，由山东省大数据局主办，青岛市大数据发展管理....

人工智能

时间：2019/09/23–2020/06/20

参赛人数：1423



进行中



山东省（青岛）数据创新创业应用赛

创意竞赛

山东省大数据局、青岛市大数据发展管理局

为深入贯彻落实习近平总书记视察山东重要讲话、重要指示批示精神，加快实施《数字山东发展规划（2018—2022年）》，推进数字山东建设，培育富有活力的数字经济，由山东省大数据局主办，青岛市大数据发展管理....

人工智能

时间：2019/09/20–2020/06/20

参赛人数：595



进行中



ARVR精品应用大赛

创意竞赛



https://www.dcjingpai.com/static_page/cmpList.html

比赛数目：9（活跃赛+训练赛）

V. 课程要求-综合大作业-数泉竞赛



文本实体识别及关系抽取

中国计算机学会

奖励

0

可报名

队伍

1,659

③ 2019-08-17 ~ 2020-03-31

阿尔茨海默症的识别

中国计算机学会

奖励

0

可报名

队伍

2,264

③ 2019-08-17 ~ 2020-03-31

O2O商铺食品安全相关评论发现

中国计算机学会

奖励

0

可报名

队伍

3,505

③ 2019-08-17 ~ 2020-03-31

https://www.datafountain.cn/competitions?state=in_service
比赛数目：3

V. 课程要求-综合大作业-科赛

和鲸 Kesci

K-Lab 项目 数据集 比赛 **任务**

Datathon 训练营

未报名 Datathon训练营 训练营

「Datathon训练营」是由和鲸社区主办发起的医疗数据分析入门训练营，旨在帮助零基础的医学领域人才和对医疗数据感兴趣的新手入门数据分析。

参赛人数 321 参赛团队 299 2019/10/21 - 2020/10/21

首届“全国人工智能大赛”(AI+4K HDR赛项)

未报名 首届“全国人工智能大赛”(AI+4K HDR赛项) ¥2,680,000

2019年8月，深圳市人民政府决定专门设立人工智能领域权威赛事——全国人工智能大赛（以下简称大赛）。大赛将立足国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展。

参赛人数 1601 参赛团队 1082 2019/10/17 - N/A

首届“全国人工智能大赛”(行人重识别 Person ReID 赛项)

未报名 首届“全国人工智能大赛”(行人重识别 Person ReID 赛项) ¥2,680,000

2019年8月，深圳市人民政府决定专门设立人工智能领域权威赛事——全国人工智能大赛（以下简称大赛）。大赛将立足国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展。

参赛人数 2652 参赛团队 1935 2019/10/17 - N/A

迁移学习提供「借贷风险评估」解决方案 练习赛

未报名 迁移学习提供「借贷风险评估」解决方案 练习赛

金融场景是算法落地的重要场景。本次练习赛，我们聚焦于「借贷风险评估」问题，探索机器学习细分领域——迁移学习，在金融场景的更多可能性，以及其实践落地。|本练习赛长期开放报名：2019年04月01日 - 2020年03月28日

参赛人数 583 参赛团队 529 2019/04/01 - 2020/03/28

<https://www.kesci.com/home/competition>

比赛数目：5

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台

✓1、阿里的天池大数据竞赛【18（新人赛）朱席席】

✓2、kaggle平台竞赛【11（活跃赛）刘凯】

✓3、DataCastle大数据竞赛平台

<http://www.pkbidata.com/> 【9（活跃赛+训练赛）徐翔】

✓4、数泉竞赛平台(datafountain.cn)【3王庆勇】

✓5、科赛 (<https://www.kesci.com/apps/home/competition>)
【5王庆勇】

✓6、GitHub数据挖掘类源代码分析

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台【前5项作业要求】

- ✓ 1、单人选择题目，不允许重复，先报名先得，微信发给对应的教辅
- ✓ 2、只能够选择比赛截至时间3月20日之后的题目
- ✓ 3、注册账号命名规范：NUDT丁兆云DM2020A+姓名字母首写（如张三命名为“NUDT丁兆云DM2020AZS”，命名不规范不计成绩）
- ✓ 4、成绩评价方法（30分）：
 - ✓ 排名成绩20（排名计量： \log_2 竞赛平台中的排名/ \log_2 竞赛平台中的所有队伍数目，排名从高到底，前4名20分，第5-9名19分，后面以此类推减去1分，减到0不计分）
 - ✓ 提交ppt文档：PPT质量10（题目来源0.5、题目内容0.5、题目数据0.5、求解思路2、详细过程5、实验结果1、排名结果截图0.5）

V. 课程要求-综合大作业-

● 平时成绩

■ 综合大作业平台

- ✓1、阿里的天池大数据竞赛【18（训练赛）朱席席】
- ✓2、kaggle平台竞赛【11（活跃赛）刘凯】
- ✓3、DataCastle大数据竞赛平台
<http://www.pkbidata.com/> 【9（活跃赛+训练赛）徐翔】
- ✓4、数泉竞赛平台(datafountain.cn)【3王庆勇】
- ✓5、科赛 (<https://www.kesci.com/apps/home/competition>)
【5王庆勇】
- ✓6、GitHub数据挖掘类源代码分析

	网址	负责人
1	https://github.com/abdulfatir/twitter-sentiment-analysis	丁兆云
2	https://github.com/AlanConstantine/SinglePass	丁兆云
3	https://github.com/liuhuanyong/TopicCluster	丁兆云
4	https://github.com/YcheCourseProject/CommunityDetection	丁兆云
5	https://github.com/letiantian/TextRank4ZH	丁兆云
6	https://github.com/timothyasp/PageRank	丁兆云
7	https://github.com/klyc0k/EDSFilter	丁兆云
8	https://github.com/raviekambaram/forecasting_civil_unrest	丁兆云
9	https://github.com/Nhrkr/Predicting-Social-Unrest	丁兆云
10	https://github.com/Rvl101/Predictive-Analysis-of-Social-Unrest-Events	丁兆云
11	https://github.com/aashish-jain/Social-unrest-prediction	丁兆云
12	https://github.com/Vikramjeet-Singh/SocialUnrestTwitterAnalysis	丁兆云
13	https://github.com/rutvikbhavsar/SVM_Classification_Model_for_civil_u_nrest_relevant_tweets	丁兆云
14	https://github.com/Venkteshkavi/Takaval-Advanced-Information-Retrieval	丁兆云

V. 课程要求-综合大作业-github源代码

	网址	负责人
15	https://github.com/lzha97/hk_news	丁兆云
16	https://github.com/mathemakitten/china-misinformation	丁兆云
17	https://github.com/ryankhaleghi/HongKongProtest-Tweet-NLP	丁兆云
18	https://github.com/SecondDim/crawler-news	丁兆云
19	https://github.com/rex-chien/taiwan-2020-election-hustings	丁兆云
20	https://github.com/pynayzr/pynayzr	丁兆云
21	https://github.com/olala7846/twnews	丁兆云
22	https://github.com/lincht/PTT	丁兆云
23	https://github.com/YYYYMao/rssCrawler	丁兆云
24	https://github.com/shihs/taiwan-company-database	丁兆云
25	https://github.com/TaiwanStat/Taiwan-news-crawlers	丁兆云
26	https://github.com/matchawu/OpenData	丁兆云
27	https://github.com/g0v/twly_crawler	丁兆云
28	https://github.com/g0v/addressbook.parser	丁兆云

V. 课程要求-综合大作业-github源代码

	网址	负责人
29	https://github.com/johnb30/gdelt_download	丁兆云
30	https://github.com/erbrown33/elk-gdelt-tutorial	丁兆云
31	https://github.com/choltz95/Evis	丁兆云
32	https://github.com/code-jon/GDELT_Predict	丁兆云
33	https://github.com/gdelt-analysis/downloader	丁兆云
34	https://github.com/olivierdupuis/gdelt_mining	丁兆云
35	https://github.com/flyrightsister/gdelt_heatmap	丁兆云
36	https://github.com/jeremieperes/MongoDB-Gdelt	丁兆云
37	https://github.com/AustinTSchaffer/GDELT-Event-Data-Processor	丁兆云
38	https://github.com/attacc/networkgdelt	丁兆云
39	https://github.com/ahazeemi/RevDet	丁兆云
40	https://github.com/crconline/MMSS-webmining	丁兆云
41	https://github.com/Kali-Dev/policy-recommendation_data_visualization_omdena_app	丁兆云
42	https://github.com/kenneth-zhou/News-scrap	丁兆云

V. 课程要求-综合大作业-github源代码

	网址	负责人
43	https://github.com/Ebiquity/CASIE	刘凯
44	https://github.com/mallabiisc/RESIDE	刘凯
45	https://github.com/google-research/bert	刘凯
46	https://github.com/facebookresearch/SpanBERT	刘凯
47	https://github.com/autoliuweijie/K-BERT	刘凯
48	链接: https://pan.baidu.com/s/1S6MI8rQyQ4U7dLszyb73Yw 提取码: gc6f	刘凯
49	https://github.com/DataScienceNigeria/ERNIE-2.0-from-Baidu-Inc.	刘凯
51	https://github.com/Alasd/noise_fairlearn	王庆勇
52	https://github.com/IBM/AIF360	王庆勇
53	https://github.com/dddoss/tensorflow-socher-ntn	朱席席
54	https://github.com/nddsg/ProjE	朱席席
55	https://github.com/AbdullahAshfaq/DNN_commonsense-reasoning	朱席席

V. 课程要求-综合大作业-github源代码

	网址	负责人
56	https://github.com/xrb92/DKRL	朱席席
57	https://github.com/bxshi/ConMask	朱席席
58	https://github.com/TimDettmers/ConvE	朱席席
59	https://rajarshd.github.io/ChainsofReasoning/	朱席席
60	https://github.com/nju-websoft/DSKG	朱席席
61	https://github.com/xwhan/DeepPath	朱席席
62	https://github.com/IBCNServices/pyRDF2Vec	朱席席

V. 课程要求-综合大作业-

● 平时成绩

■ 综合大作业平台

- ✓1、阿里的天池大数据竞赛【18（训练赛）朱席席】
- ✓2、kaggle平台竞赛【11（活跃赛）刘凯】
- ✓3、DataCastle大数据竞赛平台
<http://www.pkbidata.com/> 【9（活跃赛+训练赛）徐翔】
- ✓4、数泉竞赛平台(datafountain.cn)【3王庆勇】
- ✓5、科赛 (<https://www.kesci.com/apps/home/competition>)
【5王庆勇】
- ✓6、GitHub数据挖掘类源代码分析

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台【GitHub源代码分析作业要求】

- ✓ 1、单人选择一个代码题目，不允许重复，先报名先得，在微信群里发布，以微信群里发布的顺序为主，每个专业课代表以excel统计后（excel字段包括：学号、姓名、专业、源代码链接、负责老师或者助教、负责课代表姓名），所有专业一起合并后把整体情况发给我
- ✓ 2、成绩评价方法（30分）

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台【GitHub源代码分析作业要求】

✓ 2、成绩评价方法（30分）：

✓ 跑通源代码（到负责人处验收）计10分

✓ 提交word文档：文档质量成绩20分

- ✓ 源代码理论原理分析2分【提炼出了源代码对应的理论算法（比如贝叶斯分类算法、支持向量机算法等），且详细阐述理论算法】
- ✓ 源代码的输入数据分析5分【有数据基本介绍计1分、有数据字段介绍计1分、有数据的基本统计分析计1分（数据规模、每个字段的均值、方差等统计特性）、有数据的基本可视化分析计1分（直方图、折线图等）、有其他额外数据基本分析计1分（比如词频统计等）】
- ✓ 源代码输出数据分析2分【输出数据基本介绍计1分、输出数据字段介绍计1分】
- ✓ 源代码过程分析5分【预处理方法分析计1分【数据清洗、数据集成、数据规约、数据变换】、数据输入与预处理接口分析计1分、预处理与核心算法接口分析计1分、核心算法与输出接口分析计1分、整体交互流程计1分】
- ✓ 源代码实验结果分析3分【有实验结果输出文件计1分、实验结果可解释性阐述计1分、实验结果可视化分析计1分（如准确率、召回率等的柱状图、直方图分析等）】
- ✓ 源代码进一步分析3分【在源代码基础上，具有其他更进一步的功能实现及分析】

V. 课程要求-综合大作业

● 平时成绩

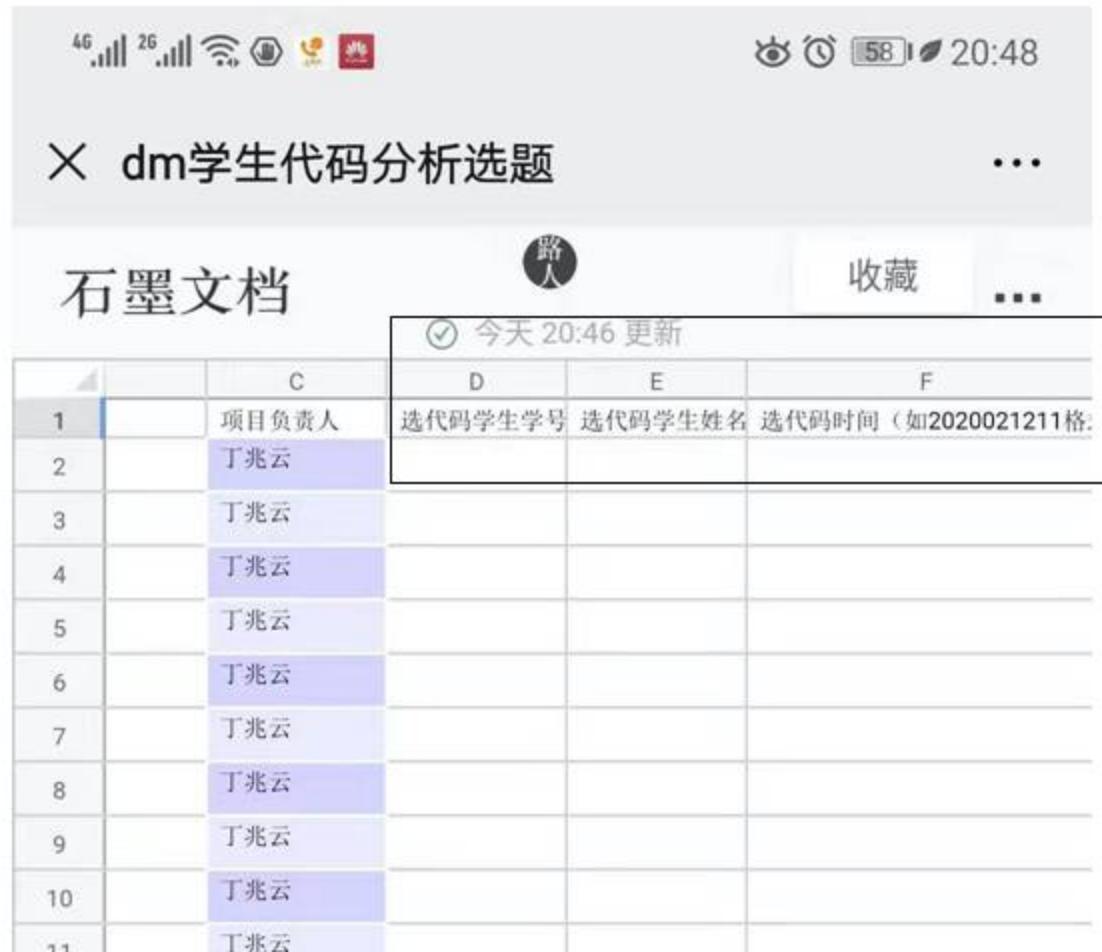
■ 综合大作业平台【GitHub源代码分析作业要求】

- ✓ 成绩评价方法（30分）：
- ✓ 跑通源代码（到负责人处验收）计10分
- ✓ 提交word文档目录
 - ✓ 1、源代码理论原理分析
 - ✓ 2、源代码的输入数据分析
 - ✓ 2.1数据基本介绍；2.2有数据字段；2.3数据的基本统计分析；2.4数据的基本可视化分析；2.5其他额外数据基本分析
 - ✓ 3、源代码输出数据分析
 - ✓ 3.1输出数据基本介绍；3.2输出数据字段介绍
 - ✓ 4、源代码过程分析
 - ✓ 4.1预处理方法分析；4.2数据输入与预处理接口；4.3预处理与核心算法接口分析；4.4核心算法与输出接口分析；4.5整体交互流程
 - ✓ 5、源代码实验结果分析
 - ✓ 5.1实验结果输出文件介绍；5.2实验结果可解释性阐述；5.3实验结果可视化分析
 - ✓ 6、源代码进一步分析

V. 课程要求-综合大作业

● 平时成绩

■ 综合大作业平台【GitHub源代码分析作业要求】



选好题目
后通过微
信小程序
：填写D
、E、F、
G、H列

V. 课程要求-综合大作业

要求：综合大作业在**2月20日**之前所有同学完成选题，各位同学在后期的练习和作业中**以自己的大作业为基础**来完成，这样既能够得平时作业分，也能够为综合大作业完成做好铺垫

V. 第二次课后作业

- 第二次课后作业-在**educoder**平台上完成作业
 - <https://www.educoder.net/shixuns/3b68gcuy/challenges>
 - <https://www.educoder.net/shixuns/relpnffc/challenges>
 - <https://www.educoder.net/shixuns/iotbujuv8/challenges>
 - <https://www.educoder.net/shixuns/zyr6upbs/challenges>
 - <https://www.educoder.net/shixuns/8ozahglc/challenges>
 - <https://www.educoder.net/shixuns/fej8xwbm/challenges>

提交作业截至时间：2020年2月20日

Any Questions?

谢谢！