Topic 9: 机器学习原理及应用

Machine Learning Fundamentals and Applications

刘跃文 博士 Dr. LIU, Yuewen

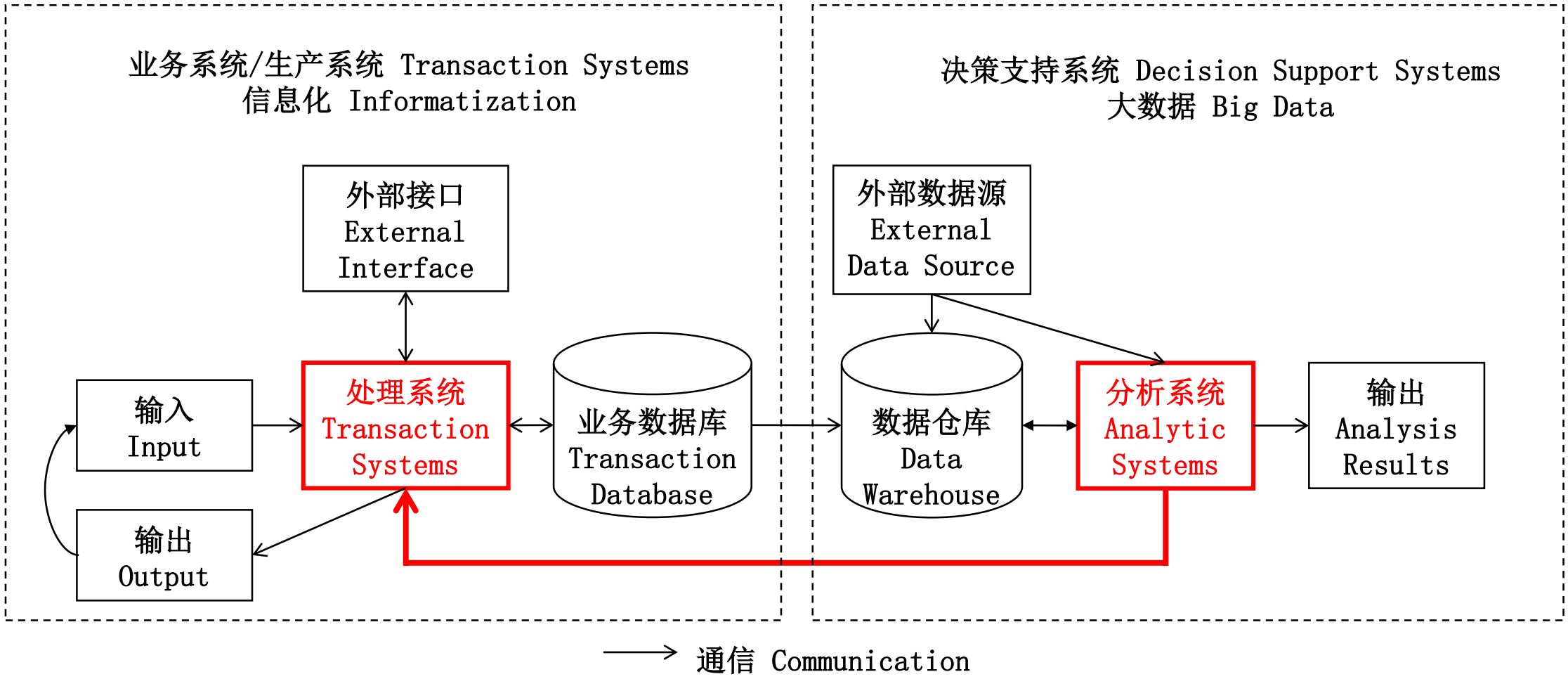
教授、博士生导师 Professor

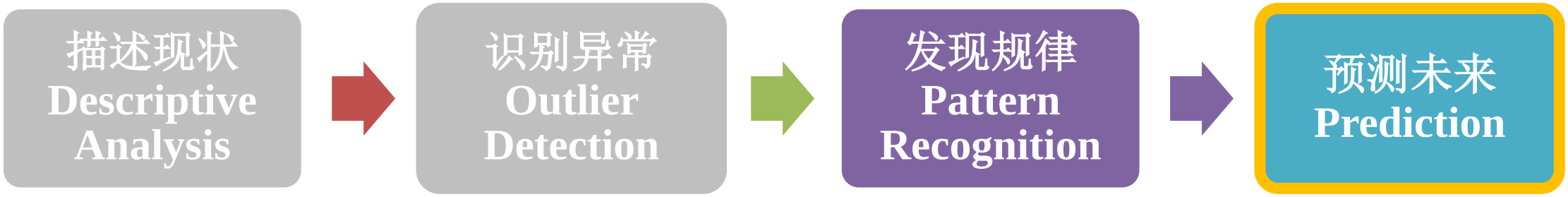
liuyuewen@xjtu.edu.cn

西安交通大学管理学院

School of Management, Xi'an Jiaotong University

V2.0, 2023-Oct





数据挖掘：知识发现
Data Mining: Knowledge Discovery

提纲

1. 机器学习原理
2. 特征表与数据预处理
3. 机器学习模型的分类与应用场景
4. 机器学习模型的评估与弱点
5. 机器学习与商业模式

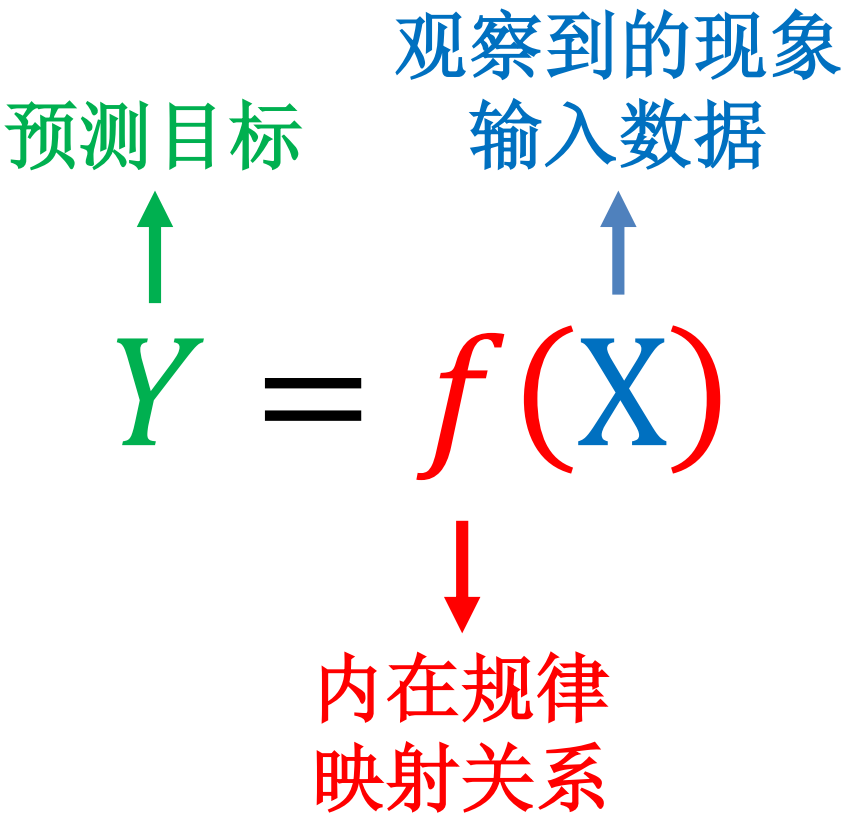
西安交通大学管理学院
THE SCHOOL OF MANAGEMENT
XI'AN JIAOTONG UNIVERSITY

1. 机器学习原理
Machine Learning Fundamentals

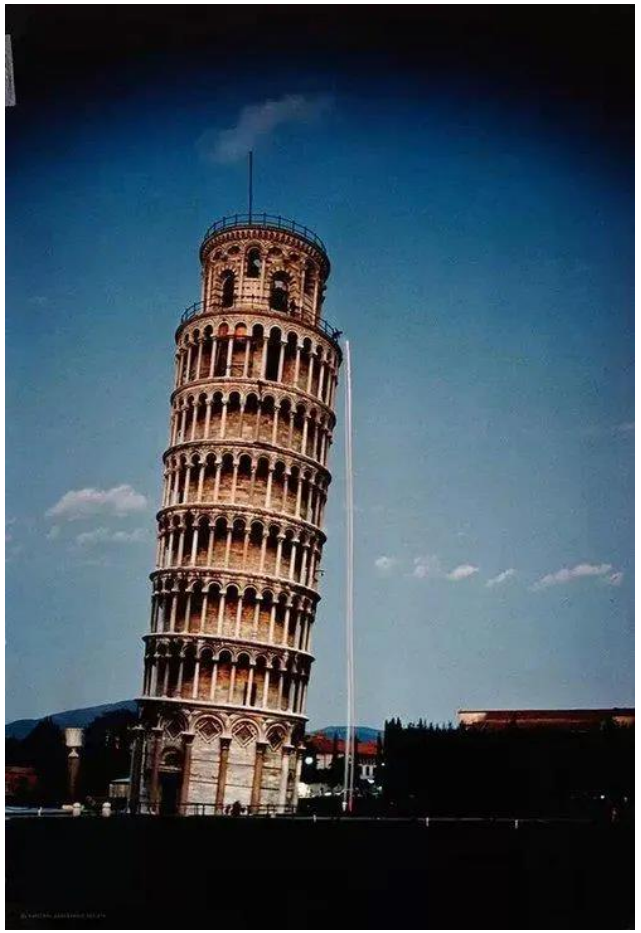
1. 机器学习的本质是找到函数

- 从输入（Input）到输出（Output）的映射关系/计算机制。

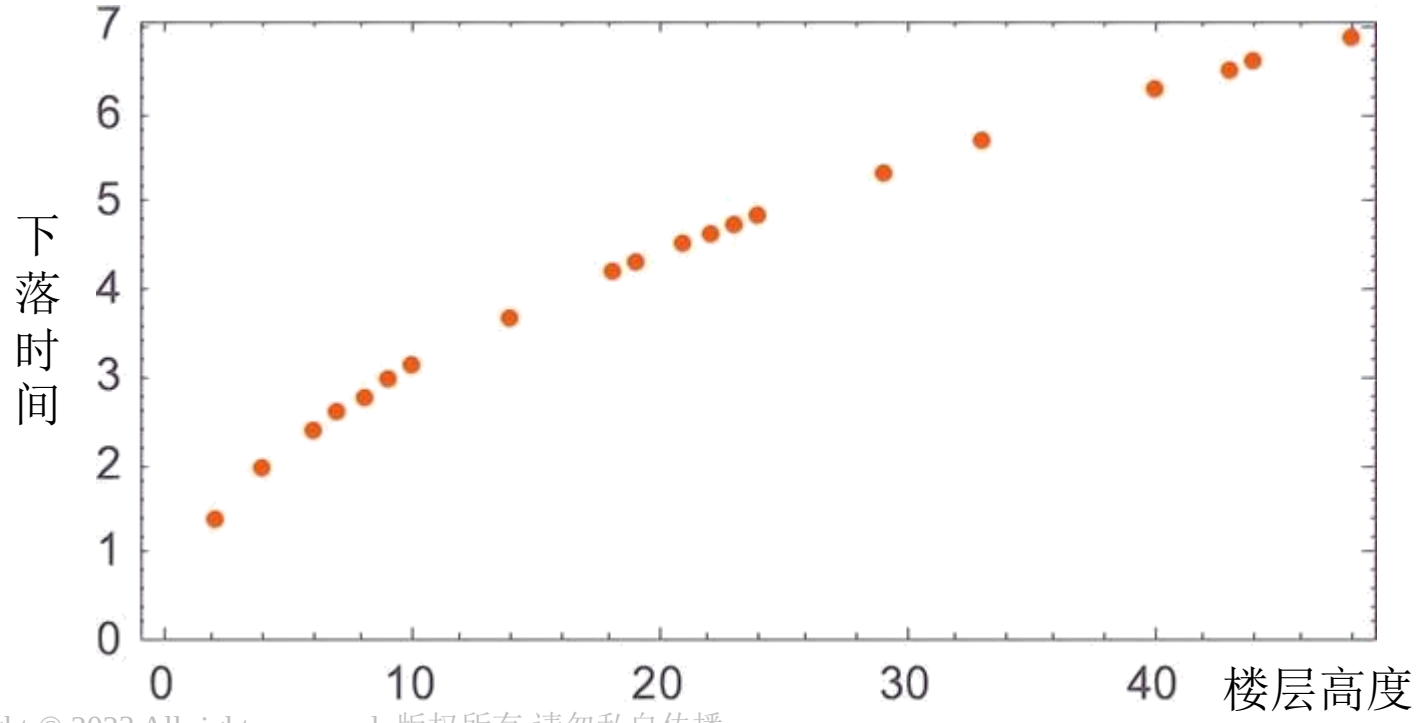
机器学习模型	输入 Input X	$f()$ →	输出 Output Y
客户流失预测	客户历史数据	→	客户流失/非流失
客户价值预测	客户消费数据	→	客户价值
图片识别	动物图片	→	图片标注 猫/狗
人脸识别	人脸图片	→	身份
机器翻译	汉语	→	英语
自动驾驶	雷达数据	→	车辆位置



2. 预测铁球落地时间



假设你在16世纪晚期，想知道从比萨斜塔的每层楼上掉下来的炮弹需要多长时间才能触地。你可以在每种情况下测量它并制作一个结果图表。

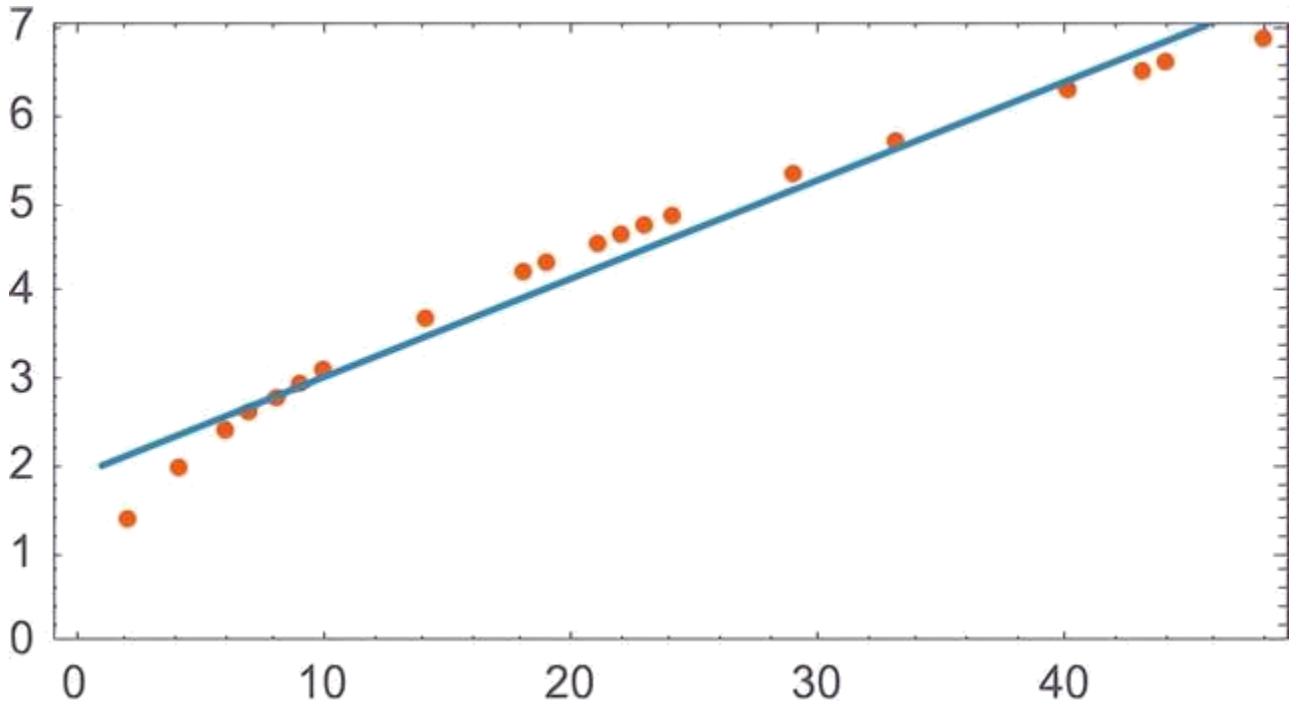


预测目标
下落时间

输入数据
下落高度

$$Y = f(X)$$

第一步：找到一个合适的函数形式；
第二步：基于数据去估计（学习）其中的参数大小。



预测目标 Y 铁球下落时间

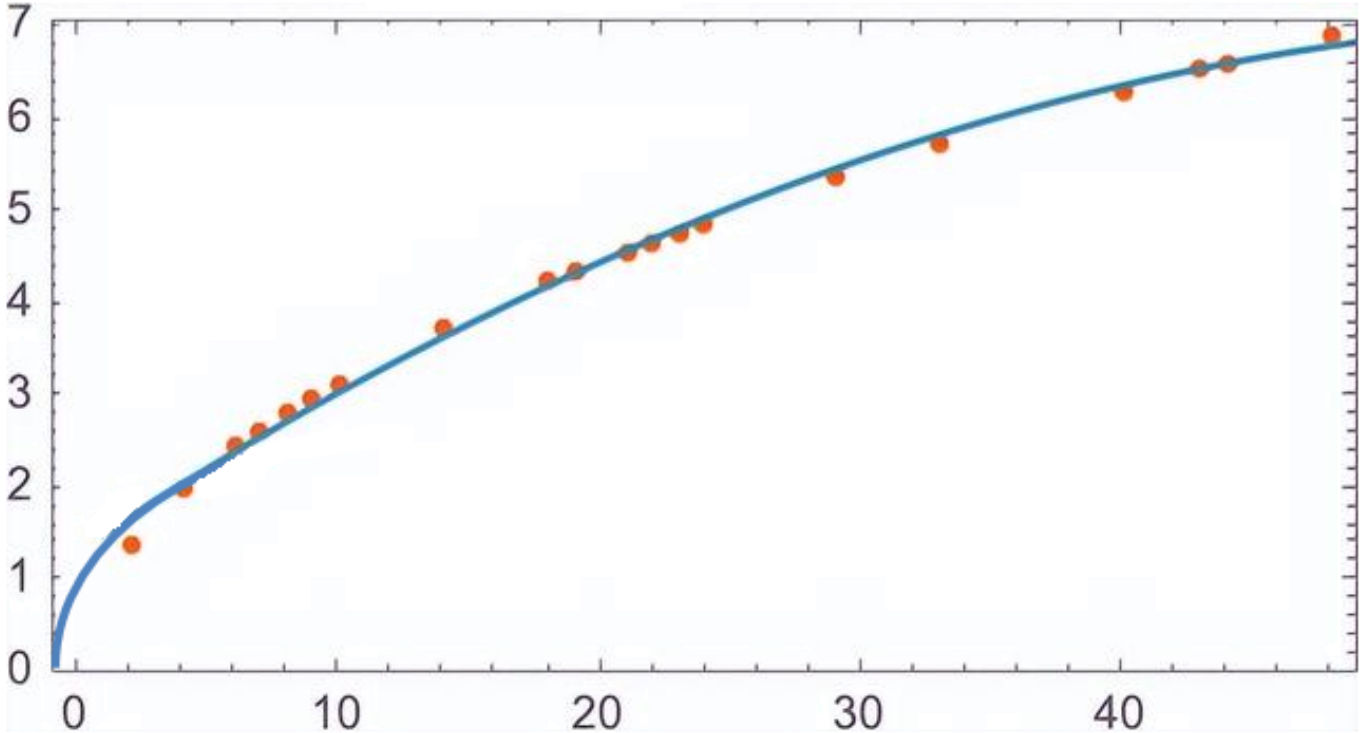
输入变量 X铁球下落高度

函数形式 $Y = aX + b$

拟合函数 $Y = 0.1X + 2$

学习参数 $a = 0.1, \mathrm{b} = 2$

需要选择“正确”的函数形式；
并理解函数形式背后的理论/原理。

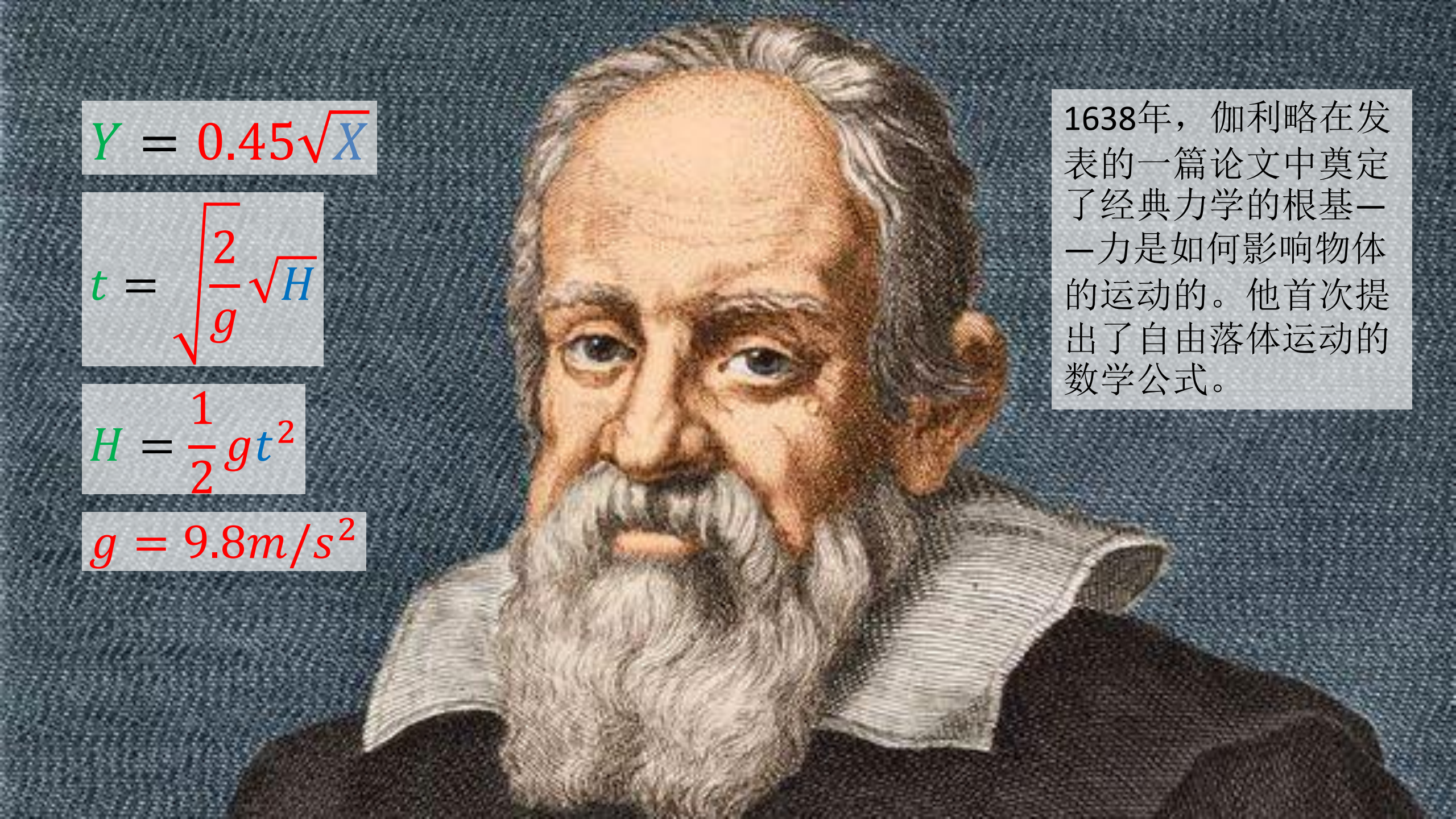


函数形式

$$Y = a\sqrt{X} + b$$

拟合函数 $Y = 0.45\sqrt{X}$

学习参数 $a = 0.45, \mathrm{b} = 0$



$$Y = 0.45\sqrt{X}$$

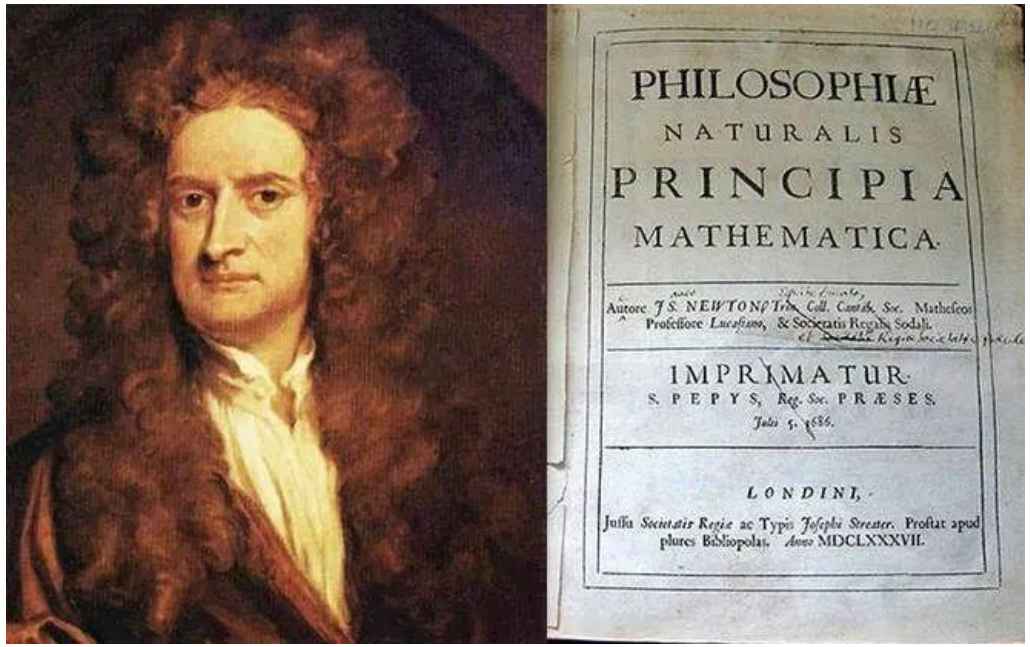
$$t = \sqrt{\frac{2}{g}}\sqrt{H}$$

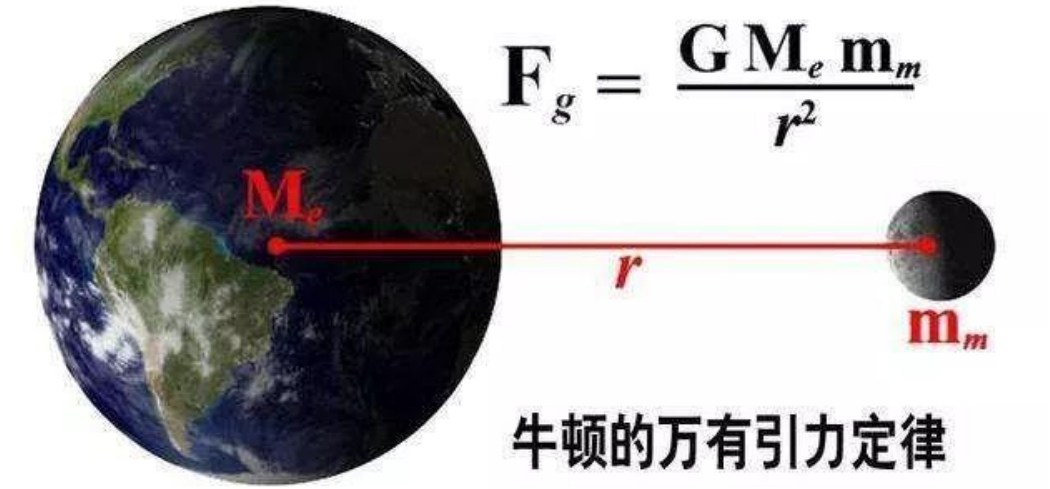
$$H = \frac{1}{2}gt^2$$

$$g = 9.8m/s^2$$

1638年，伽利略在发表的一篇论文中奠定了经典力学的根基——力是如何影响物体的运动的。他首次提出了自由落体运动的数学公式。

https://www.sohu.com/a/324878830_100028727





1687年（清康熙二十六年），英国物理学家艾萨克.牛顿发表了他的巨著《自然哲学的数学原理》，这本书的出版预示着科学时代的到来。

牛顿在《自然哲学的数学原理》第三卷中写道：“最后，如果由实验和天文学观测，普遍显示出地球周围的一切天体被地球重力所吸引，并且其重力与它们各自含有的物质之量成比例，则月球同样按照物质之量被地球重力所吸引。另一方面，它显示出，我们的海洋被月球重力所吸引；并且一切行星相互被重力所吸引，彗星同样被太阳的重力所吸引。由于这个规则，我们必须普遍承认，一切物体，不论是什么，都被赋与了相互的引力（gravitation）的原理。因为根据这个表象所得出的一切物体的万有引力（universal gravitation）的论证……”

3. 预测/预警客户流失

- Customer churn prediction
- 发展一个新客户是需要一定成本的，一旦客户流失，将会对商家造成损失，所以对客户流失的预测显得尤为重要。
- 预测客户流失的作用有：
  - 预测哪些客户有可能称为流失客户，在他流失之前使用触达策略挽留客户；
  - 分析客户流失的原因，寻找先行指标来提升留存率，完善产品。
  - 对于已经流失的客户，改变产品运营策略拉回客户，促进回流。

https://www.jianshu.com/p/143782bc15e4
https://zhuanlan.zhihu.com/p/40197660
https://zhuanlan.zhihu.com/p/68397317

假如你是银行/运营商客户经理，假设你想知道哪些客户在未来几个月会离开，以便进行客户挽留。你可以通过企业的信息系统获取大量的客户资料。

	A	B	C	D	E	U
1	customerID	gender	SeniorCitizen	Partner	Dependents	Churn
2	7590-VHVEG	Female	0	Yes	No	No
3	5575-GNVDE	Male	0	No	No	No
4	3668-QPYBK	Male	0	No	No	Yes
5	7795-CFOCW	Male	0	No	No	No
6	9237-HQITU	Female	0	No	No	Yes
7	9305-CDSKC	Female	0	No	No	Yes
8	1452-KIOVK	Male	0	No	Yes	No
9	6713-OKOMC	Female	0	No	No	No
10	7892-POOKP	Female	0	Yes	No	Yes
11	6388-TABGU	Male	0	No	Yes	No
12	9763-GRSKD	Male	0	Yes	Yes	No
13	7469-LKBCI	Male	0	No	No	No
14	8091-TTVAX	Male	0	Yes	No	No
15	0280-XJGEX	Male	0	No	No	Yes

https://zhuanlan.zhihu.com/p/60142926?from_voters_page=true

预测目标
是否流失

$$Y = f(X)$$

变量/属性/特征

- gender 性别（male/female）
- SeniorCitizen（老年人与否1/0）
- Partner（有无合作伙伴）
- Dependents（有无家属）
- tenure（用户入网月数/留存月数）
- PhoneService（用户是否有电话服务）
- MultipleLines（用户是否有多线）
- InternetService (互联网服务提供商：DSL, Fiber optic, No)
- OnlineSecurity (在线安全Yes, No, No internet service)
- OnlineBackup(在线备份Yes, No, No internet service)
- DeviceProtection (设备检测Yes, No, No internet service)
- TechSupport (技术支持Yes, No, No internet service)
- StreamingTV (流媒体电视Yes, No, No internet service)
- StreamingMovie (流媒体电影Yes, No, No internet service)
- Contract（客户的合同期限：Month-to-month, One year, Two year）
- PaperlessBilling(无纸化账单：Yes, No)
- PaymentMethod (支付方式：Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges（月费用）
- TotalCharges（总费用）

朴素的业务经验总结及表达

- 朴素的业务经验总结：流失客户有哪些特征？
  - 例如：男性、中年、每月通话数量0-60次、每月通话时长0-300分钟

- 朴素的业务经验表达/应用方式：
  - **条件筛选**：筛选30-50岁男性、月通话次数0-60次且通话时长0-300分钟的用户作为潜在的流失客户，进行客户关怀。
  - **积分预警**：男性[20分]、女性[0分]；30-50岁[20分]、20-30岁[10分]、50-60岁[10分]、其它[0分]；每月通话数量0-60次[20分]、60-150次[10分]、其它[0分]；每月通话时长0-300分钟[20分]、300-500分钟[10分]、其它[0分]；……最终累积得分越高，风险越高，如果分数超过100分则作为潜在流失客户进行预警。

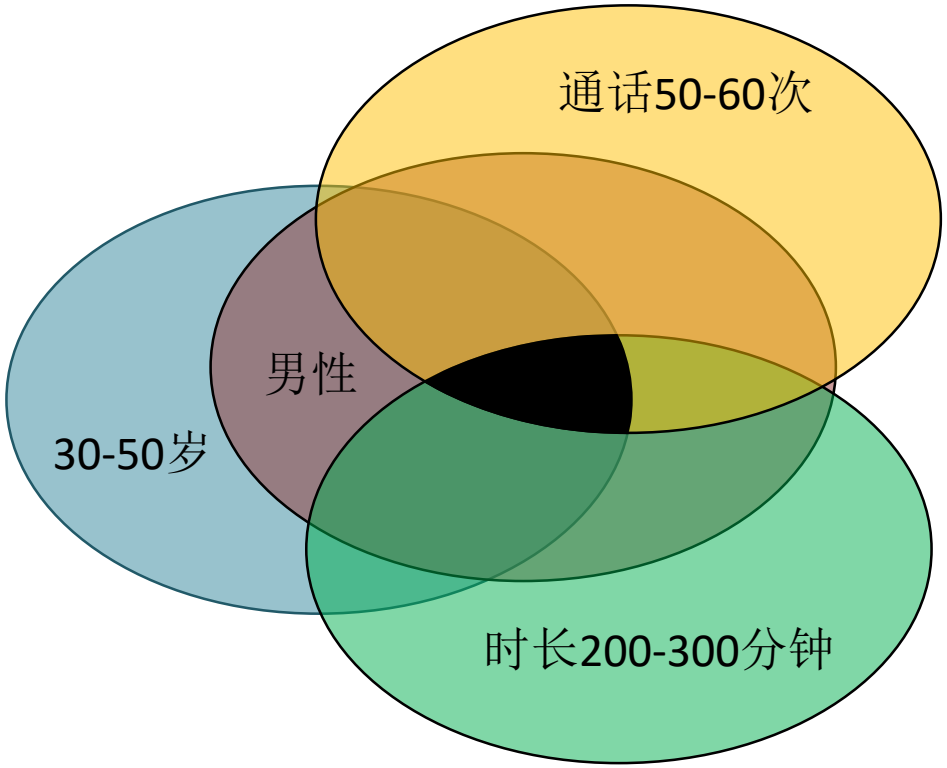
条件筛选方法

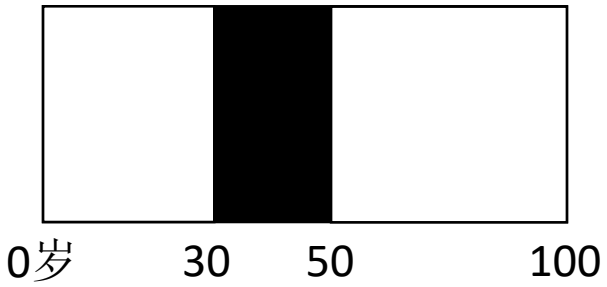
$$Y=\begin{cases}1 & if\ gender=1\ and\ age\geq 30\\ & and\ age\leq 50\ and\ calls\leq 60\\ & and\ callLength\leq 300;\\ 0 & otherwise.\end{cases}$$

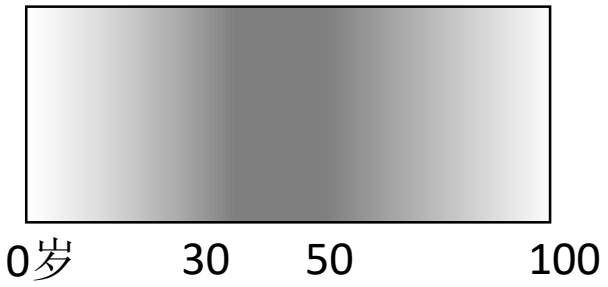
问题：参数是拍脑袋得到的；阈值也是拍脑袋决定的；分数大小很难控制。

条件筛选方法的缺陷

- 每一个条件都是硬边界条件（为什么是30-50岁，29-51岁是否可以？）事实上，很少有条件与预测目标有绝对的关系。
- 多个条件不能配合使用（如年龄51岁，但其它所有条件都极其符合要求的条件，是否可能流失？）
- 结果没有排序，操作难度较大。

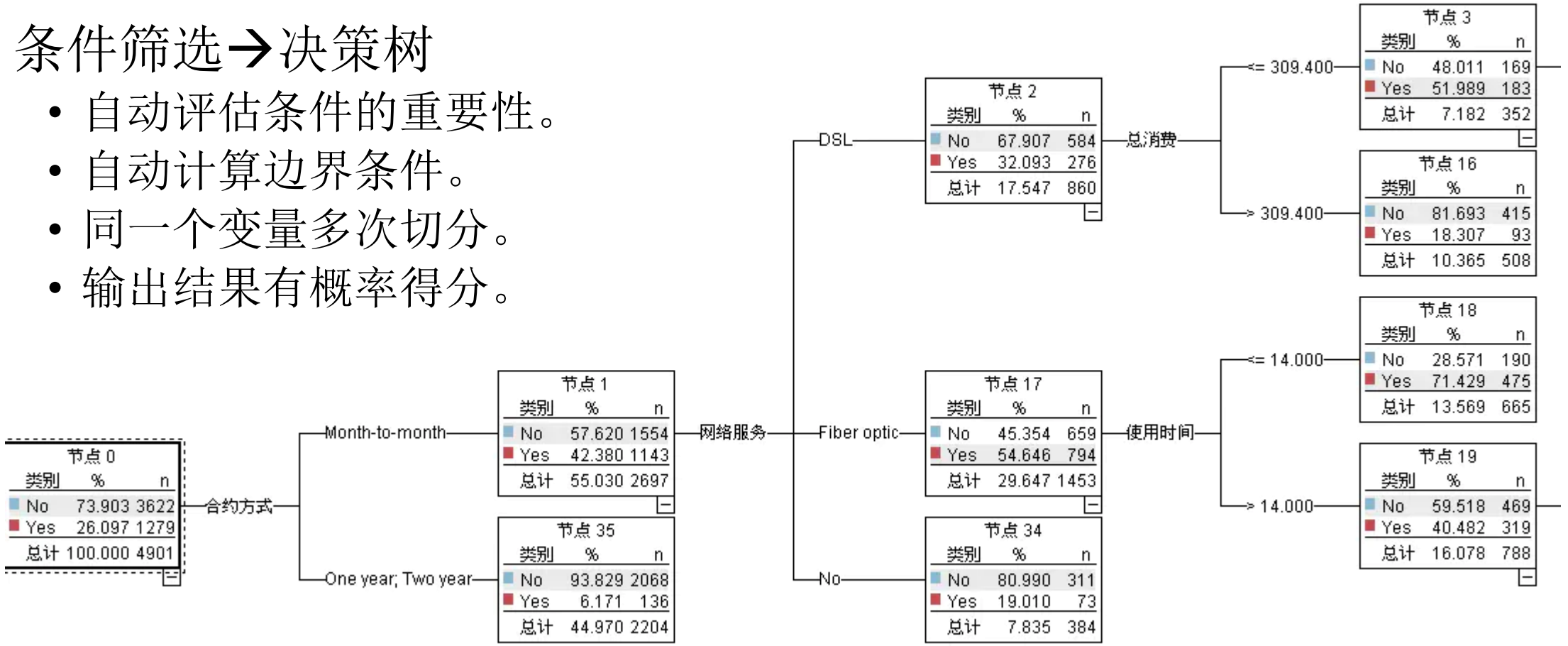






能不能让计算机来决定边界？

- 条件筛选→决策树
  - 自动评估条件的重要性。
  - 自动计算边界条件。
  - 同一个变量多次切分。
  - 输出结果有概率得分。



积分预警方法

Z β_0 β_1 X_1 β_2 X_2 β_3 X_4 β_4 X_4

客户得分 = 100 + 5*男性 - 5*通话人数 - 2*通话时长 + 10*掉话次数

如果客户得分>80分，则预警客户流失。

$$Y = \begin{cases} 1 & if\ Z > 80 \\ 0 & otherwise \end{cases} \qquad Z = \sum_j \beta_j X_j$$

问题：参数是拍脑袋得到的；阈值也是拍脑袋决定的；分数大小很难控制。

积分预警方法的缺陷

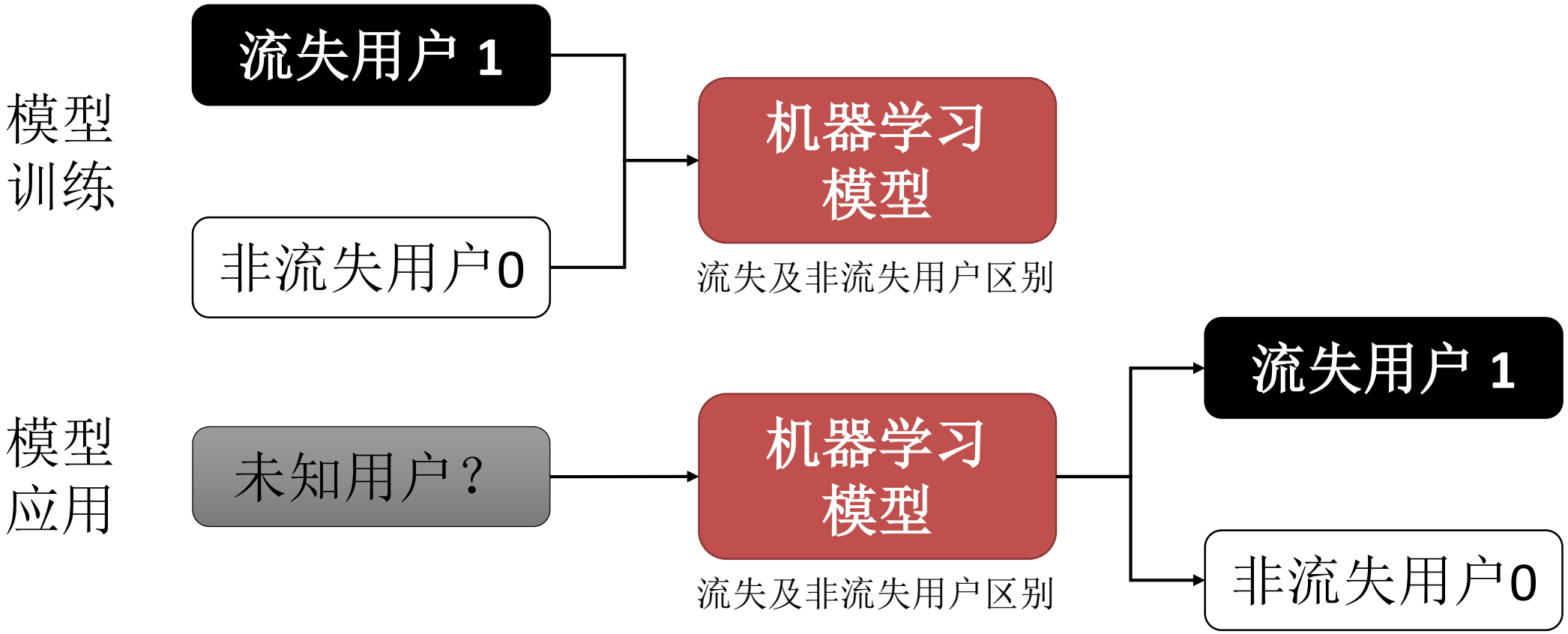
- 权重全靠拍脑袋。
- 分数的权重不具有科学性（为什么男性加20分，为什么不是19分）？
- 分数结果没有上限。
- 权重的非科学性，导致分数之间不具备可比性（例如：男性、48岁、65次、320分钟，共计60分；与女性、45岁、1次、15分钟，共计60分；相比较而言，哪个更容易流失？）
- 因为权重是人为设置的，分数的颗粒度往往不够（例如：男性、48岁、65次、320分钟，共计60分；男性、45岁、89次、480分钟，也是60分；相比较而言，显然前者更容易流失）

能不能让计算机来决定权重？

- 积分预警→逻辑回归
  - 权重由计算机来决定。
  - 计算分数在0-1之间。
  - 分数为概率值，具有可比性。

	coef	std err
Intercept	4.7601	0.470
duration	-0.2917	0.015
feton	-1.4144	0.128
gender	1.4394	0.131
call_10086	-0.9040	0.126
peakMinDiff	-0.0024	0.001
edu_class	0.5434	0.078
AGE	-0.0195	0.005
prom	2.3925	0.693
nrProm	-0.7430	0.248
posTrend	-1.5598	0.416
negTrend	-1.3003	0.414
peakMinAv	0.0011	0.000
posPlanChange	-1.0211	0.624

4. 机器学习模型训练及应用



如何选择样本？

- 目标样本（正样本）：需要筛选出的用于工作的目标对象。
- 对照样本（负样本）：除了目标对象之外，其它的普通对象。

- 通常而言，目标样本和对照样本要相对平衡。
- 机器学习的目标是，找到目标样本和对照样本的区别（模式）。

如何描述样本：属性提取/特征工程

- 静态表（每人1条数据）：
  - 用户信息表（profile）：上一月份的用户资料，上一月份的用户资料信息中哪些用户离网已经有了标记；
- 动态表（每人数据条数不确定）：
  - 话单表（cdr_call）：提供了最近六个月的通话数据
  - 数据表（cdr_mms）：提供了最近六个月的数据详单
  - 短信表（cdr_sms）：提供了最近六个月的短信详单
  - 掉话表（cdr_kill）：提供了最近六个月掉话的情况
- 基于动态表，提取属性表，每人1条数据
  - 近6个月通话数、对端数、异常掉线数、近3个月短信条数……

属性表

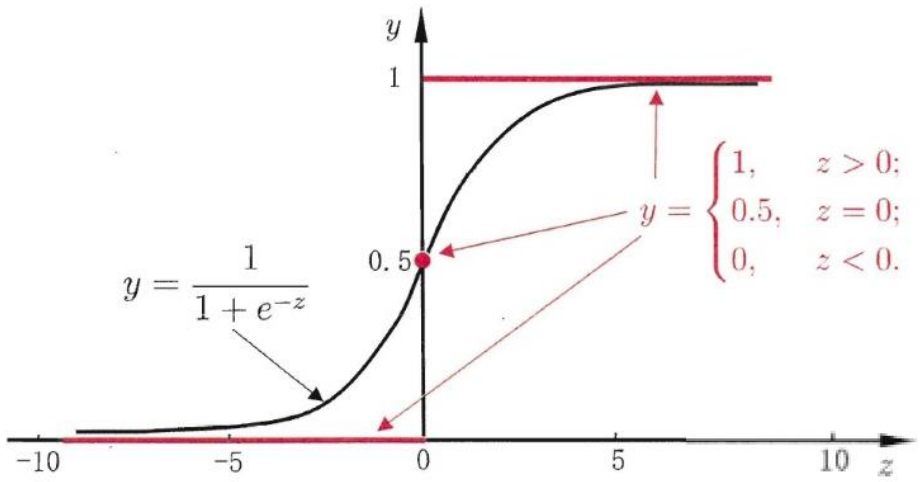
手机号	号码注册时间	性别	年龄	……	近3月通话次数	近3月通话时长	近1月短信条数	……	是否流失
13…6678	2016-7	男	28		67	208	34		否
13…2345	2015-3	男	45		122	466	8		否
13…9854	2020-6	女	23		89	29	23		是

	性别	年龄	近3月通话时长	近1月短信条数
流失用户 1	男性占比74% 女性占比26%	平均28岁	平均37分钟	平均13条
非流失用户0	男性占比72% 女性占比28%	平均36岁	平均460分钟	平均15条
	属性	特征	特征	属性

深入理解“机器学习模型”

模型	特征	参数
选择合适的模型 例如：决策树、逻辑回归、支持向量机、神经网络、……	使用哪些特征？ 特征数量 特征种类 例如：逻辑回归、决策树中的参数	利用样本，训练模型中的参数 例如：逻辑回归中的系数、决策树中的决策顺序、阈值等
人工设定		机器计算

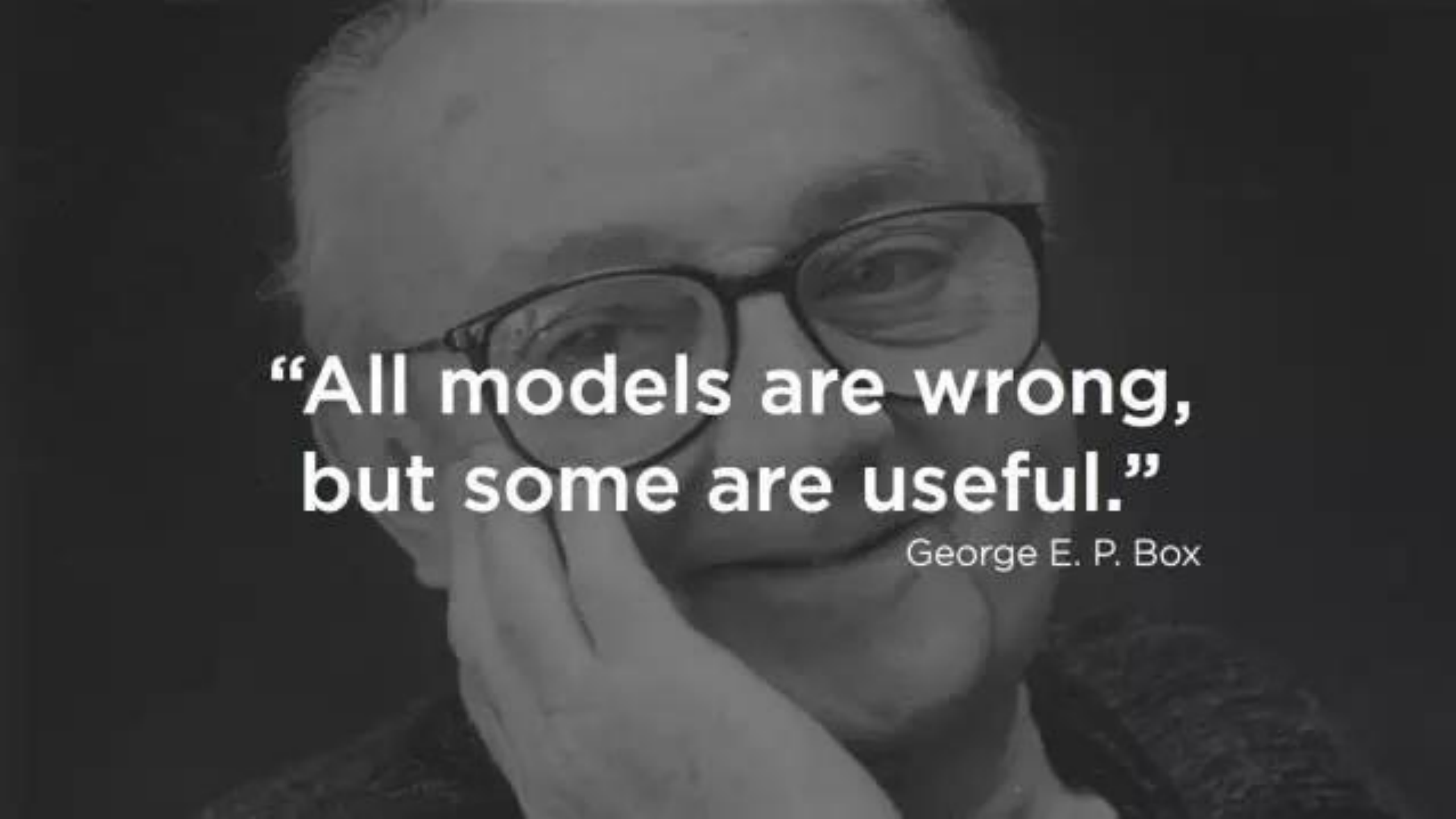
- 模型选择：逻辑回归 $y=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\cdots)}}$
- 特征选择：
  - x_1：年龄
  - x_2：近3个月通话时长
  - x_3：近1个月短信数
  - x_4：近3个月掉话数
  - ……
- 参数训练：
  - $\beta_0=0.356$，$\beta_1=2.403$，$\beta_2=3.567$，……
- 最终模型：$y=\frac{1}{1+e^{-(0.356+2.403*\text{age}+3.567*call+\cdots)}}$



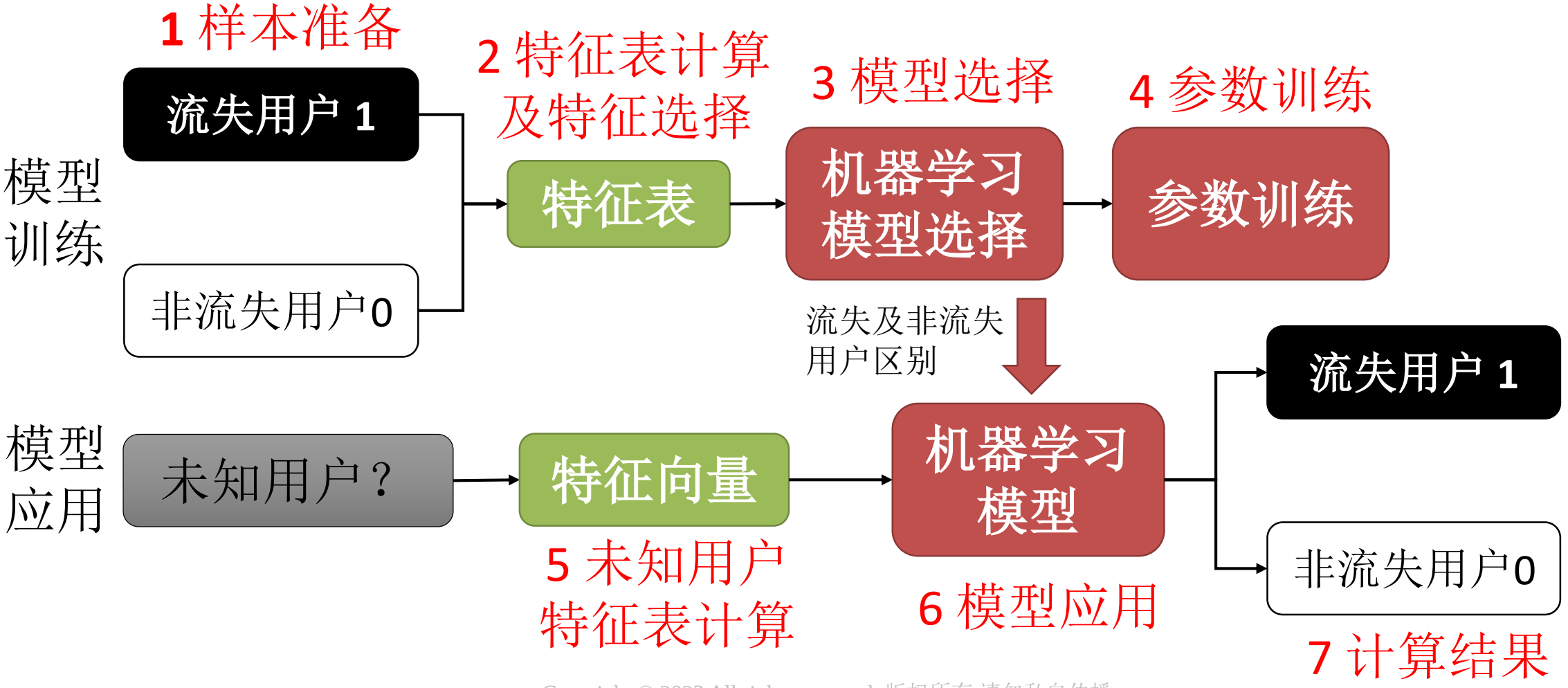
$$Y = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-(8.51+0.73*\text{男性}-0.25*\text{通话人数}-0.08*\text{通话时长}+6.42*\text{掉话次数})}}$$

ID	性别	通话人数	通话时长	掉话次数	Z	Y	预测Label
185…7203	男	8	15	2	-8.13	1.000	流失
186…3204	女	28	360	4	12.4	0.167	不流失

问题：这个函数是“真实”的函数吗？
这个函数形式未必是最佳的那个函数形式；
所采用的指标也不能涵盖所有的指标；
准确率也不可能达到100%
人类可以理解，这个函数也比较管用。


“All models are wrong,
but some are useful.”
George E. P. Box

机器学习的完整过程



2. 数据预处理及特征工程
Machine Learning Fundamentals

1、数据预处理的目标：特征表/宽表

- 数据预处理的核心对象：动态表/明细表。
- 处理的目标：另一张表/特征表/宽表。

- 动态表本来不就是一张表吗？
- **为什么要处理成另一张表**/特征表/宽表
  - （1）动态表不能直接用于建模
  - （2）**多源数据融合的必选方法**

（1）从一张表（动态表/明细表）到另一张表（特征表/宽表）

日期	时间	Pos	ID	金额
18-11-23	09:55:13	3	74	3.5
18-11-23	09:56:14	8	76	2.5
18-11-23	09:58:00	11	35	10.0
18-11-23	09:58:11	7	74	5.0
18-11-23	09:59:21	2	32	7.5
18-11-23	09:59:55	3	77	6.0

ID	早餐次数	平均早餐时间	早退次数	平均早餐金额	成绩
32	18	8:55	8	5.4	75
35	95	7:05	0	5.6	91
74	65	7:40	0	2.7	80
76	34	8:20	2	3.2	69
77	78	6:59	3	4.5	86
79	99	6:55	0	3.5	87

有什么变化？　　行的变化、列的变化

数据挖掘真正的基础

特征/属性
Attributes

- 另一张表
- 业界常称为 特征表 或者 宽表

分析对象
UoA
Unit of Analysis

ID	早餐次数	平均早餐时间	早退次数	平均早餐金额	成绩
32	18	8:55	8	5.4	75
35	95	7:05	0	5.6	91
74	65	7:40	0	2.7	80
76	34	8:20	2	3.2	69
77	78	6:59	3	4.5	86
79	99	6:55	0	3.5	87

ID　　输入　　目标

（2）多来源数据的融合

ID	早餐次数	平均早餐时间	早退次数	平均早餐金额	超市购物金额	图书馆借阅数	洗澡时间熵	性别	成绩
32	18	8:55	8	5.4	453	6	1.5	男	75
35	95	7:05	0	5.6	24	34	2.0	女	91
74	65	7:40	0	2.7	88	7	1.1	男	80
76	34	8:20	2	3.2	789	23	3.5	女	69
77	78	6:59	3	4.5	43	77	0.8	男	86
79	99	6:55	0	3.5	86	0	3.1	男	87

就餐数据 500→1　购物 50→1　借阅 n→1　洗澡 10→1　基本信息 1　成绩 1

直接的数据关联会导致局部笛卡尔积

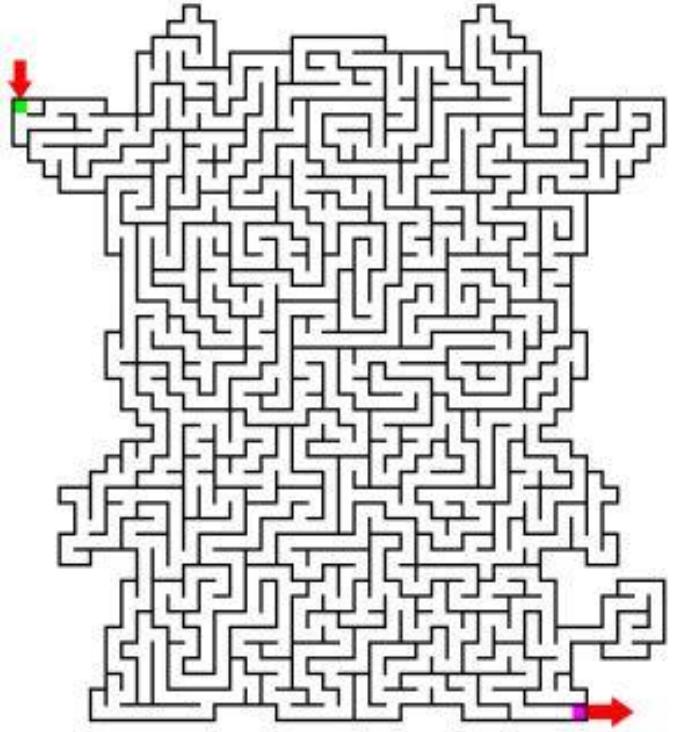
ID	性别	成绩
32	男	75

ID	时间	金额
32	男	75
32	女	91
32	男	80
32	女	69
32	男	86
32	男	87

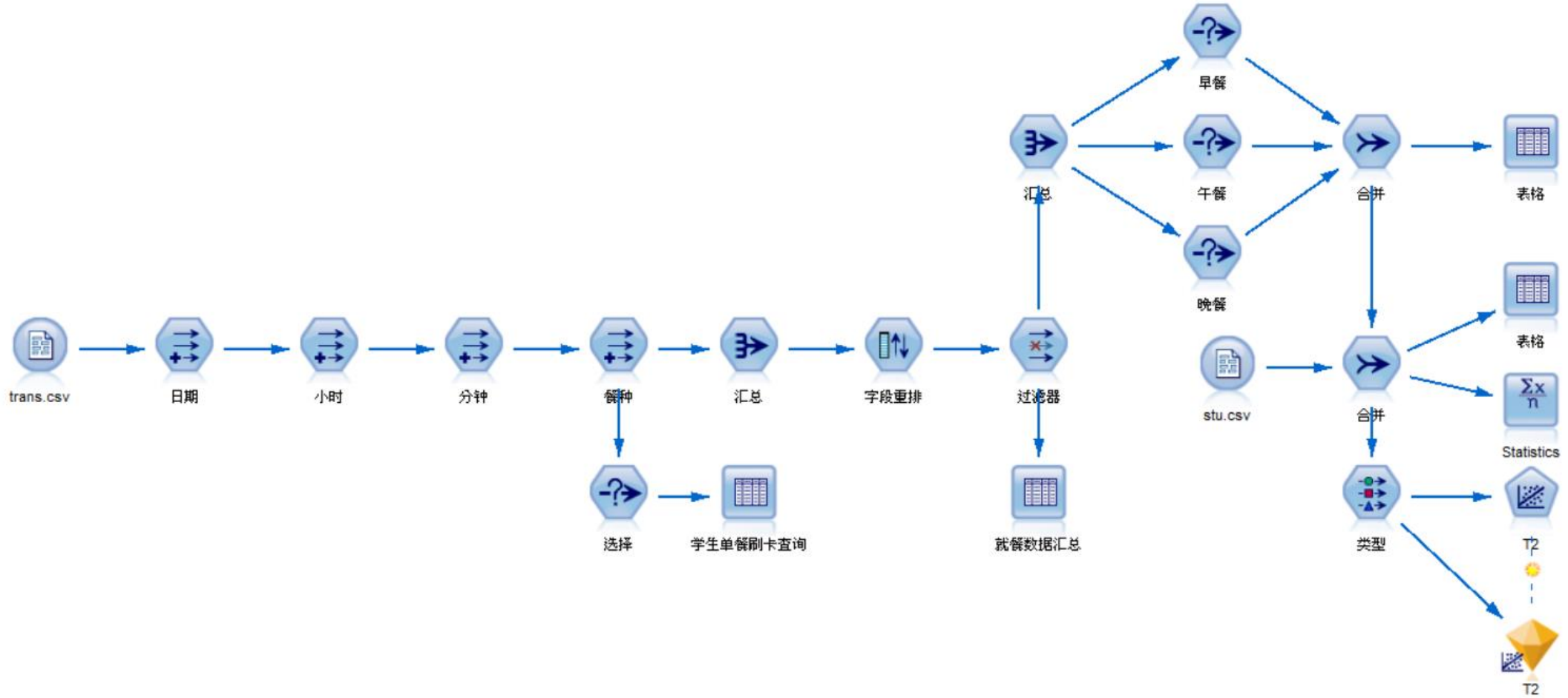
ID	性别	成绩
32	男	75
35	女	91
74	男	80
76	女	69
77	男	86
79	男	87

2. 构建特征表：业务主导、谋事在前

- 第一步：明确预测目标
- 第二步：明确分析对象UoA
- 第三步：思考需要哪些特征（业务主导）
- 第四步：思考计算过程
  （可以先拿简单的例子手算）
- 第五步：动手开始做



3. 案例：学生就餐行为特征表



1. 详细的就餐数据：1人1天1餐多次刷卡多条数据，共计：9.3万条

stuid	campus	canteen	pos	transtime	transvalue
32	北	2	112	2014-10-31 07:34:31	2.000
32	北	2	100	2014-10-31 11:49:55	1.000
35	北	2	112	2014-10-31 16:21:01	2.100
46	北	2	65	2014-10-31 16:48:18	8.000
52	北	2	65	2014-10-31 16:48:15	8.000
54	北	2	65	2014-10-31 16:48:53	8.000
55	北	2	100	2014-10-31 18:36:40	1.000
56	北	2	112	2014-10-31 16:20:47	2.800
61	北	2	112	2014-10-31 07:44:49	2.000
72	北	2	89	2014-10-30 11:40:52	6.000
73	北	2	112	2014-10-30 07:38:38	1.500

1餐多次刷卡合并为1条

2. 压缩就餐数据：1人1天1餐1条数据，共计：4.9万条

stuid	日期	餐种	刷卡次数	transvalue_Sum	分钟_Min
32	2014-10-31	早餐	2	3.600	454
32	2014-10-31	午餐	2	8.000	708
35	2014-10-31	晚餐	3	8.100	971
46	2014-10-31	晚餐	2	10.000	1008
52	2014-10-31	晚餐	3	11.400	1004
54	2014-10-31	晚餐	4	17.400	1005
55	2014-10-31	晚餐	2	5.000	1116
56	2014-10-31	晚餐	2	12.800	969
61	2014-10-31	早餐	2	5.000	464
72	2014-10-30	午餐	2	7.500	700
73	2014-10-30	早餐	4	7.500	449

1人1种餐合并为1条

3. 压缩就餐数据：1人1餐1条数据，共计：107*3条

ID	餐种	就餐次数	平均就餐金额	平均就餐时间	平均刷卡次数
1	早餐	146	3.241	451.164	1.055
1	午餐	200	10.247	702.995	1.985
1	晚餐	176	9.221	1064.358	2.034
2	早餐	2	3.400	435.500	1.500
2	午餐	42	13.586	711.452	2.714
2	晚餐	34	10.191	1078.029	2.412
3	早餐	158	2.131	469.804	1.234
3	午餐	243	5.223	707.872	1.362
3	晚餐	233	5.840	1076.670	2.197
4	早餐	238	2.597	470.017	1.168
4	午餐	272	6.455	718.298	1.294
4	晚餐	252	6.284	1069.536	1.389

1人3种餐合并为1条

4. 合并就餐数据：1人1条数据，共计：107条

ID	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额
1	146	3.241	200	10.247	176	9.221
2	2	3.400	42	13.586	34	10.191
3	158	2.131	243	5.223	233	5.840
4	238	2.597	272	6.455	252	6.284
5	126	2.448	192	7.487	198	8.192
6	135	2.216	228	6.120	223	5.853
7	111	3.622	184	7.829	186	8.657
8	56	4.470	132	10.364	115	10.607
9	109	1.932	149	7.476	140	7.393
10	193	1.399	232	4.440	233	3.842

合并就餐数据：1人1条数据，共计：107条

ID	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额
1	146	3.241	200	10.247	176	9.221
2	2	3.400	42	13.586	34	10.191
3	158	2.131	243	5.223	233	5.840
4	238	2.597	272	6.455	252	6.284
5	126	2.448	192	7.487	198	8.192
6	135	2.216	228	6.120	223	5.853
7	111	3.622	184	7.829	186	8.657
8	56	4.470	132	10.364	115	10.607
9	109	1.932	149	7.476	140	7.393
10	193	1.399	232	4.440	233	3.842

个人基础信息：1人1条数据，共计：107条

ID	Gender	T1	T2	Class	Group	Prov	City	BirthDay
1	1	88	83	1	1	福建	龙岩市	1995-02-01
2	1	76	69	1	1	福建	龙岩市	1996-05-01
3	1	75	74	1	1	福建	龙岩市	1996-10-01
4	1	90	91	2	1	北京	朝阳区	1995-11-12
5	1	86	86	2	1	江西	景德镇市	1996-01-13
6	1	88	91	3	1	内...	呼和浩特市	1996-02-21
7	1	82	75	3	1	广东	湛江市	1995-09-21
8	1	76	83	3	1	辽宁	大连市	1997-10-21
9	1	79	83	3	1	辽宁	大连市	1996-02-22
10	1	74	90	3	1	浙江	杭州市	1995-09-22

最终特征表：1人1条数据，共计：107条

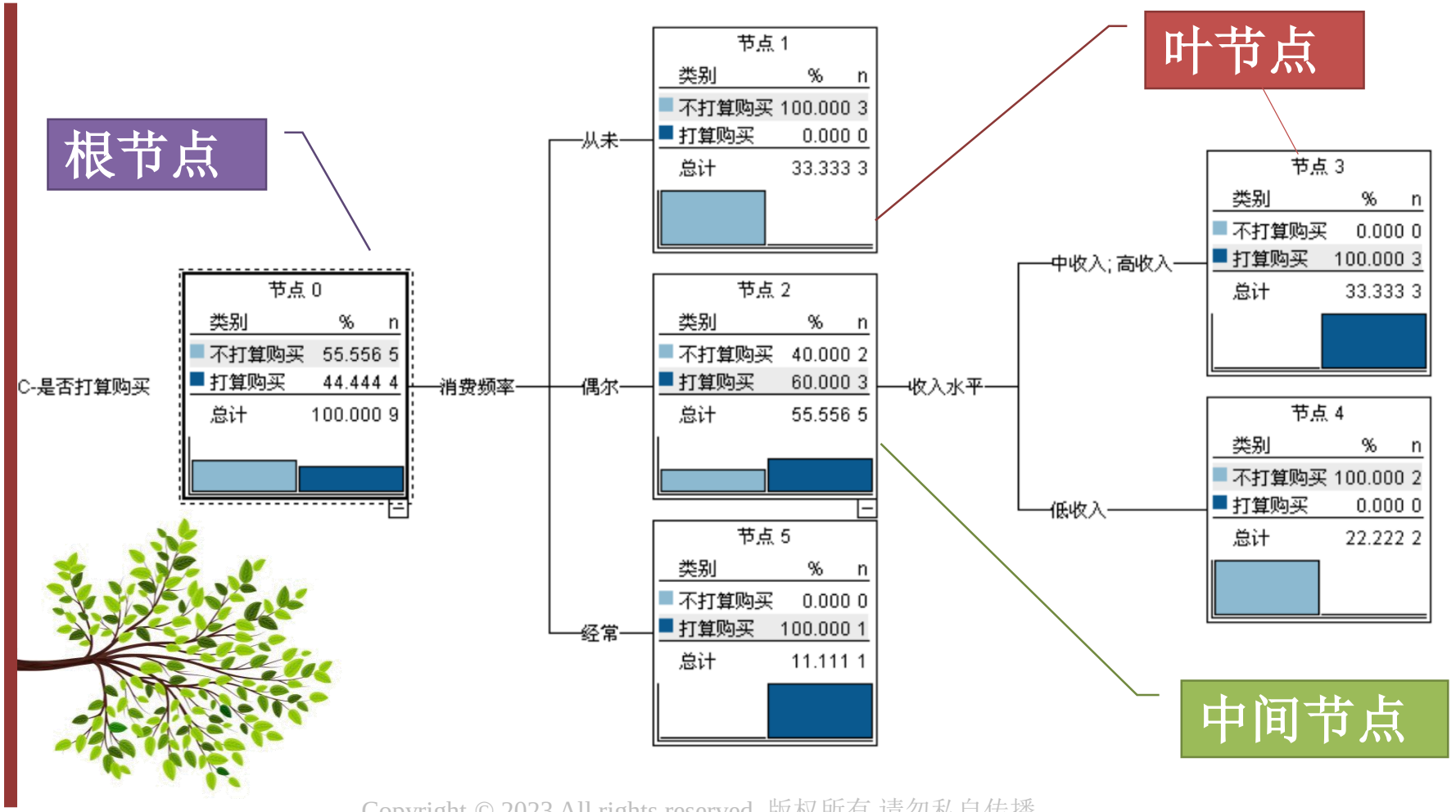
ID	BirthDay	Gender	Prov	City	Class	Group	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额	T1	T2
1	1995-02-01	1	福建	龙岩市	1	1	146	3.241	200	10.247	176	9.221	88	83
2	1996-05-01	1	福建	龙岩市	1	1	2	3.400	42	13.586	34	10.191	76	69
3	1996-10-01	1	福建	龙岩市	1	1	158	2.131	243	5.223	233	5.840	75	74
4	1995-11-12	1	北京	朝阳区	2	1	238	2.597	272	6.455	252	6.284	90	91
5	1996-01-13	1	江西	景德...	2	1	126	2.448	192	7.487	198	8.192	86	86
6	1996-02-21	1	内...	呼和...	3	1	135	2.216	228	6.120	223	5.853	88	91
7	1995-09-21	1	广东	湛江市	3	1	111	3.622	184	7.829	186	8.657	82	75
8	1997-10-21	1	辽宁	大连市	3	1	56	4.470	132	10.364	115	10.607	76	83
9	1996-02-22	1	辽宁	大连市	3	1	109	1.932	149	7.476	140	7.393	79	83
10	1995-09-22	1	浙江	杭州市	3	1	193	1.399	232	4.440	233	3.842	74	90

3. 机器学习模型的种类及应用场景
Machine Learning Fundamentals

1. 传统机器学习模型

- 有监督学习
  - 分类（预测标签）
    - **逻辑回归**
    - **决策树 Decision Tree**、随机森林 Random Forest
    - 支持向量机SVM
  - 回归（预测数值）
    - **线性回归**
    - 支持向量回归、分类回归树
- 无监督学习
  - 对象之间的关系
    - **聚类**
    - 异常值检测
  - 变量之间的关系
    - 关联规则
    - 降维

2. 决策树



- 其分析结论的展示方式类似一棵倒置的树
- 体现了对样本数据的不断分组过程
- 决策树分为2叉树和多叉树、分类树和回归树

- 体现了输入变量和输出变量取值的逻辑关系
- 逻辑比较形式表述的是一种推理规则
- 每个叶节点都对应一条推理规则

- **对新数据对象的分类预测**

- 常用分类器
- 应用非常广泛
  - 欺诈检测
  - 疾病诊断
  - 营销
  - 事故分析

相似文献

决策树模型与logistic回归模型在胃癌高危人群干预效果影响因素分析中的应用

目的 采用决策树模型与logistic回归模型分析影响农村胃癌高危人群干预效果的影响因素.方法 根据胃癌高危人群干预效果及其相关因素,分别建立决策树模型和logistic回归...

刘兵，李苹，朱玫烨，... -《中国卫生统计》

被引量: 7　发表: 2018年

Logistic回归、决策树和神经网络在预测2型糖尿病并发末梢神经病变中的性能比较

近年来,数学方法和计算机技术的发展使复杂的模型预测成为可能.目前能够建立预测模型的方法主要有统计学方法和数据挖掘方法,基于这两类方法的预测技术已逐渐被应用在生...

李长平 -《中国人民解放军军事医学科学院》

被引量: 7　发表: 2009年

决策树模型与logistic回归在中学生尝试吸烟影响因素中的应用

目的 了解江苏省中学生吸烟的分布情况及影响因素,为针对性地开展中学生控烟干预提供科学依据.方法 本研究数据来源于2013年江苏省青少年烟草调查,采用多阶段分层整群随...

曲晨，覃玉，毛涛，... -《中国慢性病预防与控制》 2020年28卷4期 264-269页 ISTIC PKU CA

被引量: 0　发表: 2020年

Logistic回归和决策树在数据库营销响应中的应用

与传统营销方式相比,数据库营销不但能提高营销效益,而且是客户关系管理的基础,是企业从以产品为中心向以客户为中心的经营体系转型的杠杆.由于强大的数据存储和挖掘功...

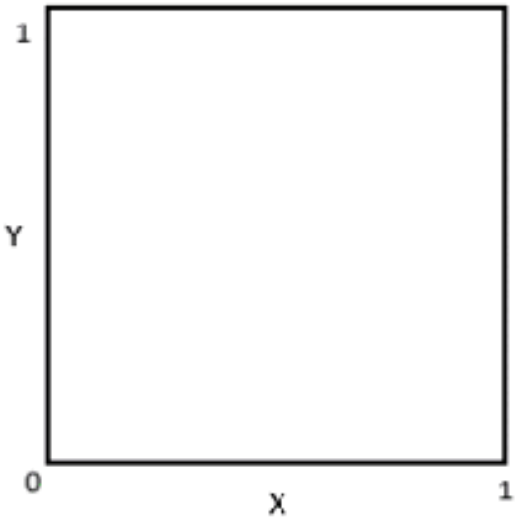
冯伟 -《兰州财经大学》

决策树算法概述

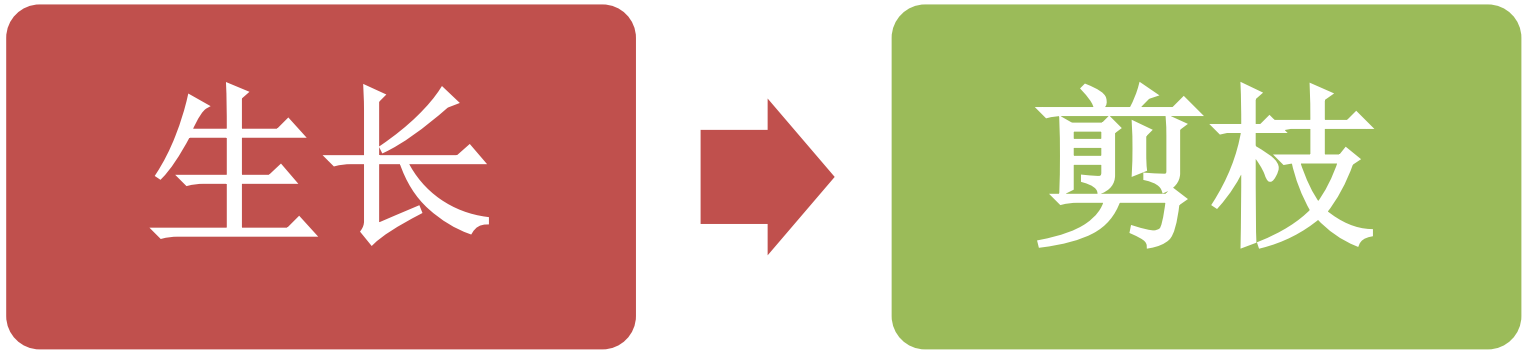
- 决策树建立的过程就是决策树各个分枝依次形成的过程
- 决策树的每个分枝在一定规则下完成对n维特征空间的区域划分
- 决策树建立好后，n维特征空间会被划分成若干个小的边界平行或垂直于坐标轴的矩形区域

决策树的生成逻辑

- 生成/生长
- 确定每一步特征空间划分标准时，都同时兼顾由此将形成的两个区域，希望划分形成的两个区域所包含的样本点尽可能同时“纯正”。



决策树步骤



利用**训练样本集**
完成决策树的建立过程

利用**测试样本集**
对所形成的决策树进行精简

训练阶段表现越完美的模型
应用阶段表现可能会越糟糕

模型扩展：集成学习

- 三个臭皮匠，顶个诸葛亮
- 偏差和方差的存在，使建立在一组训练样本集上的一个模型，所给出的预测往往缺乏稳健性
- 数据挖掘中的策略
- Bagging技术
- Boosting技术

- 均包括建模和投票两个阶段

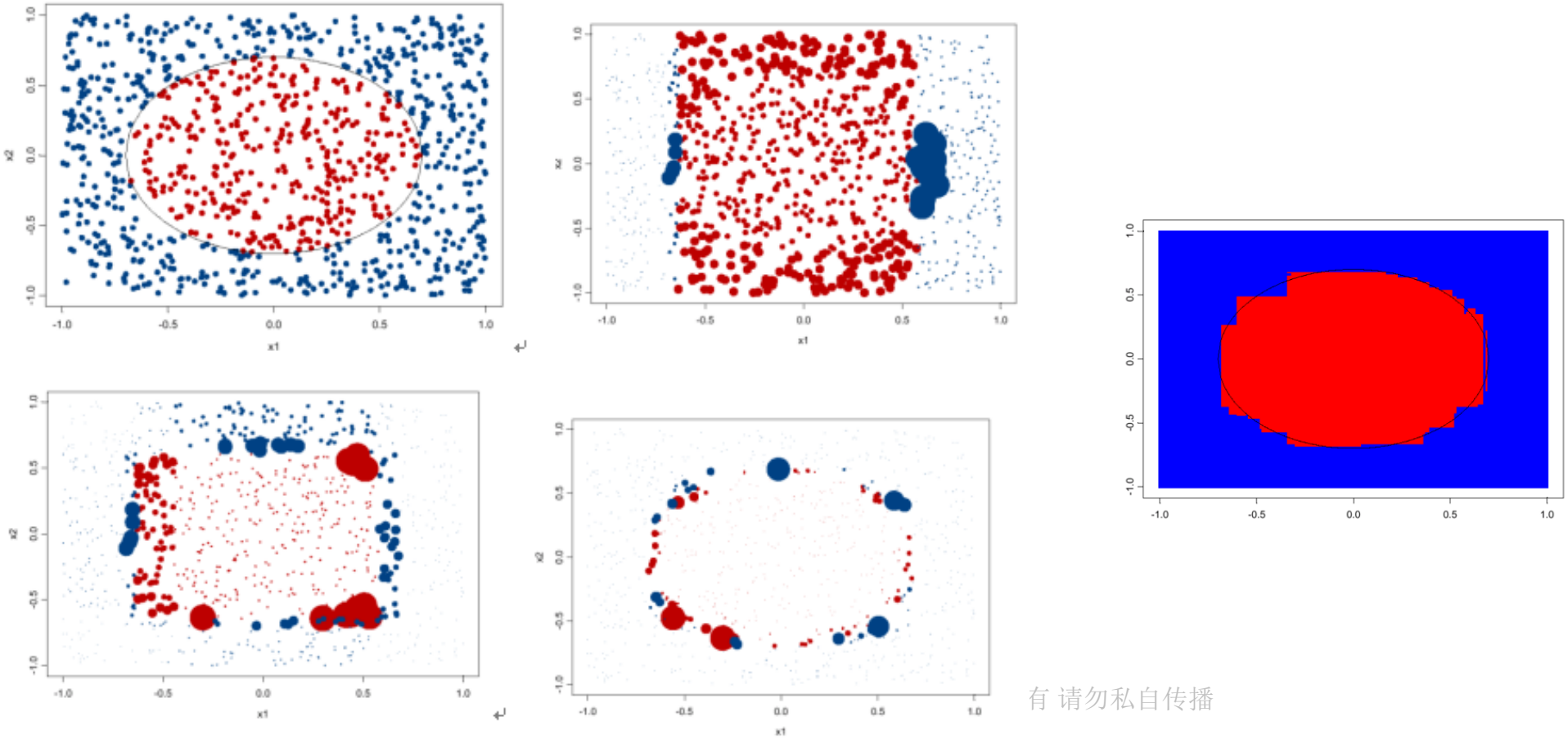
Bagging技术

- 建模过程（输入：训练样本集T，训练次数k；输出：多个决策树模型C1,C2,…Ck)

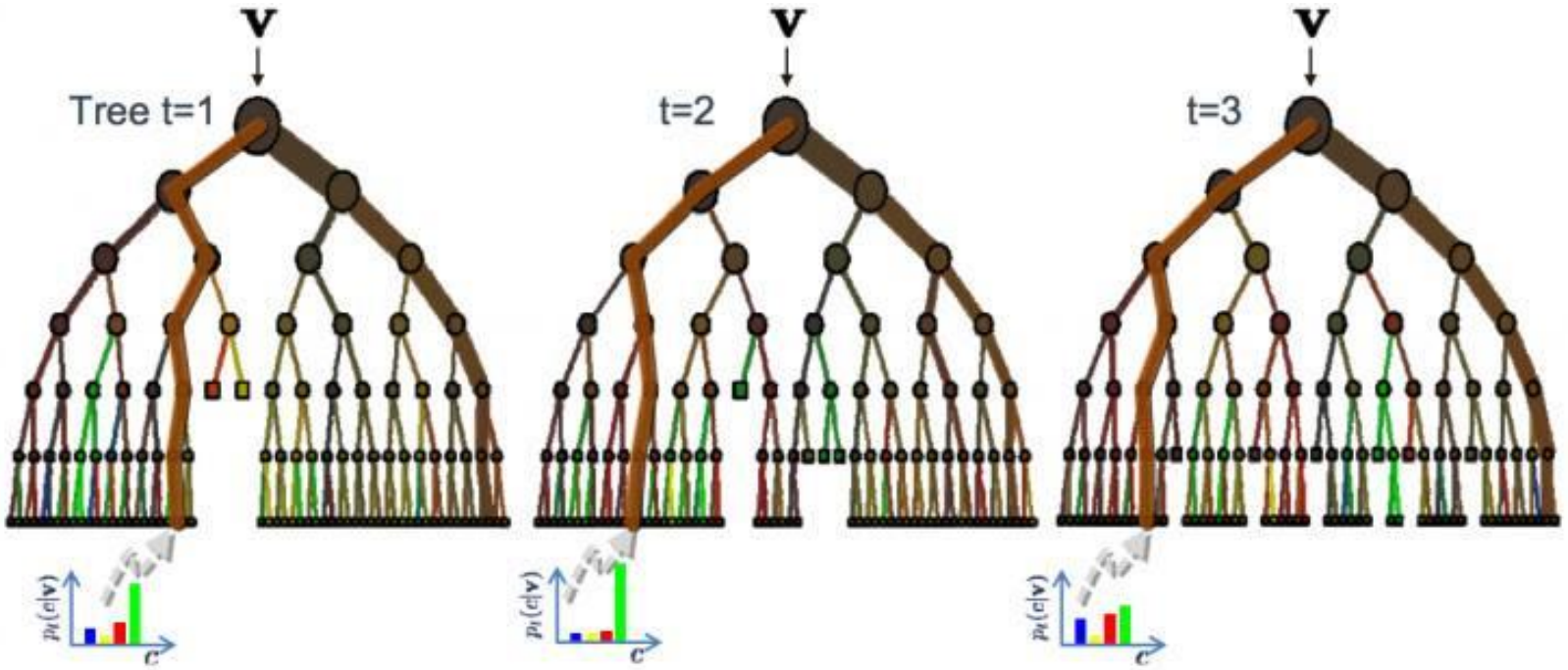
- For i=1,2,…,k do
- 从T中随机有放回抽取样本，形成有相同样本容量的样本集合Ti
- 以Ti为训练集构造模型Ci
- End for

Boosting技术

- 建立k个模型； k个模型投票

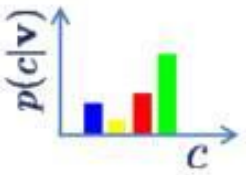


随机森林（Radom Forest）

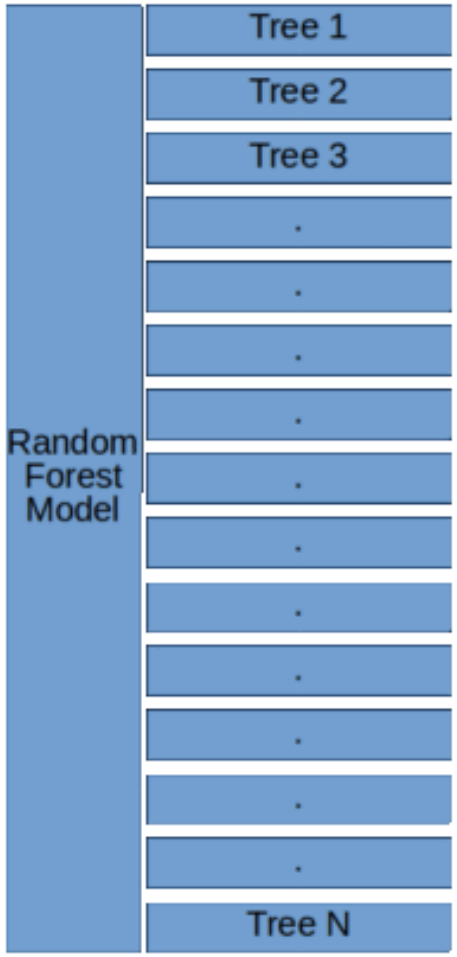


The ensemble model

Forest output probability $p(c|\mathbf{v}) = \frac{1}{T}\sum_{t}^{T} p_t(c|\mathbf{v})$

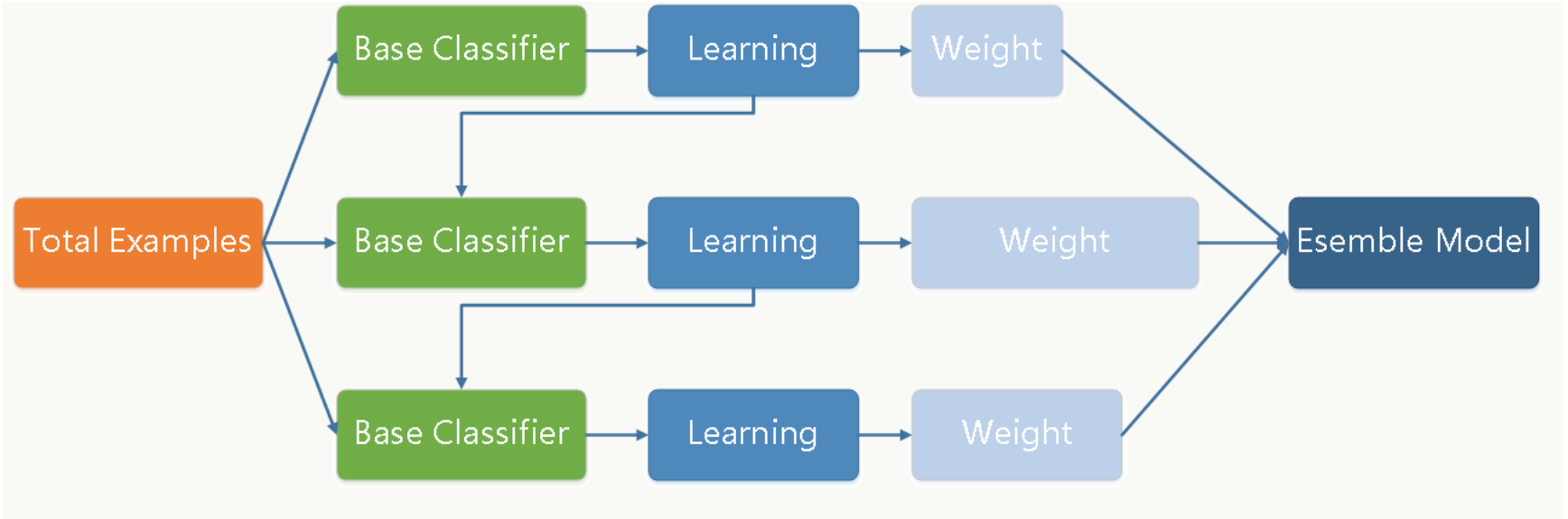


随机森林的决策结果



GBDT和XGBoost

- 随机森林是目前商业领域运用最为广泛的一种算法。
- 特别是模型算法竞赛领域。
- GBDT (Gradient Boosting Decision Tree)，XGBoost (eXtreme GBoost)



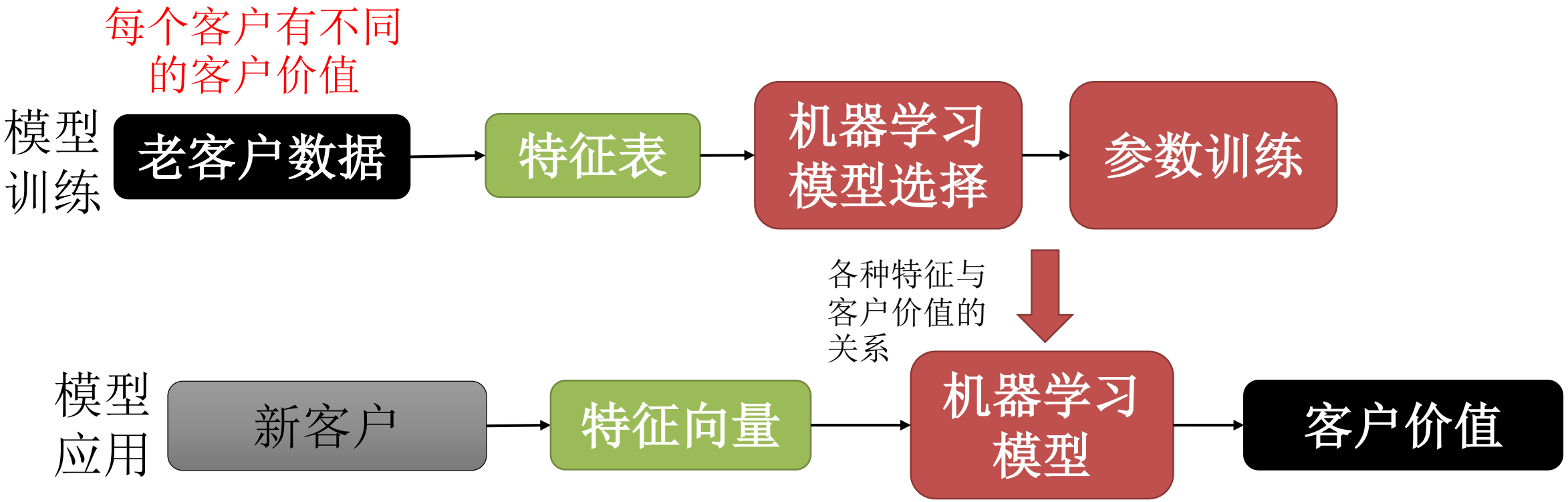
3. 回归分析

- 回归分析：
  - 用一些变量的加权求和来预测目标变量
  - **科学的积分方法**

  - 例：流失的可能性
  - 积分方法：开通了某服务，积x分；通话量，积x分
  - 回归方法：流失的可能性=f(系数1*是否开通X服务+系数2*通话量)

- 一般线性回归分析：目标变量是连续变量
- 逻辑回归分析（Logistic Regression，LR）：目标变量 0~1

客户价值预测（数量预测）



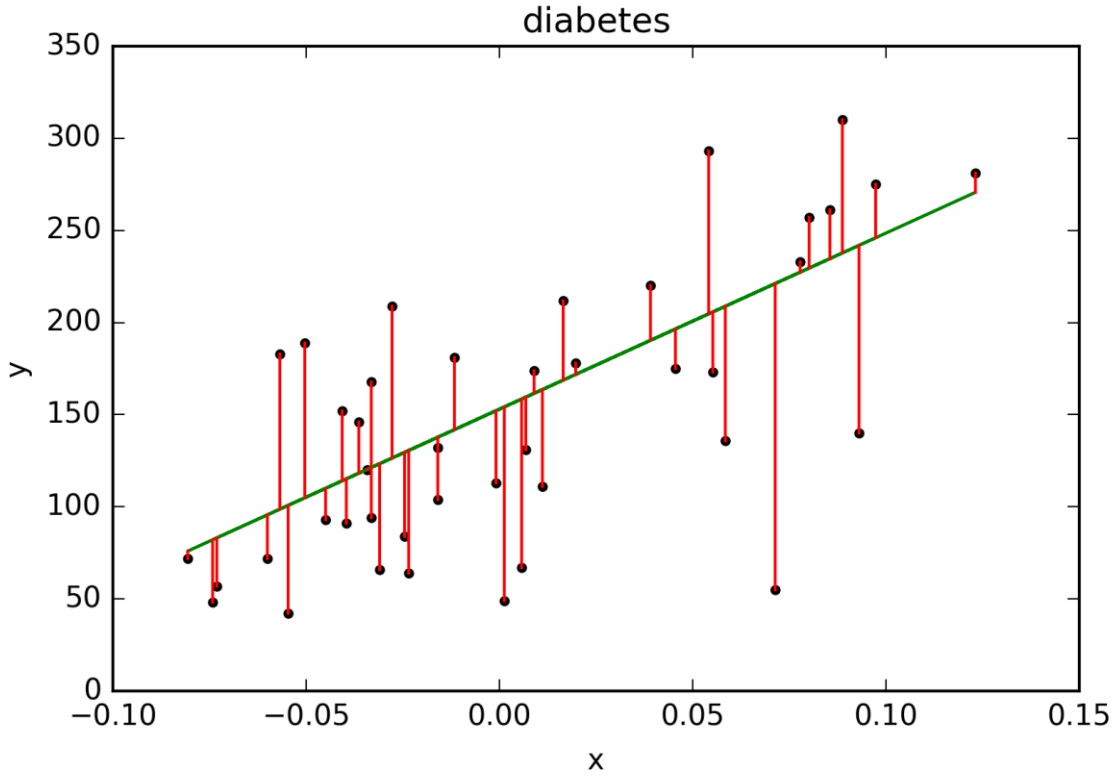
一般线性回归

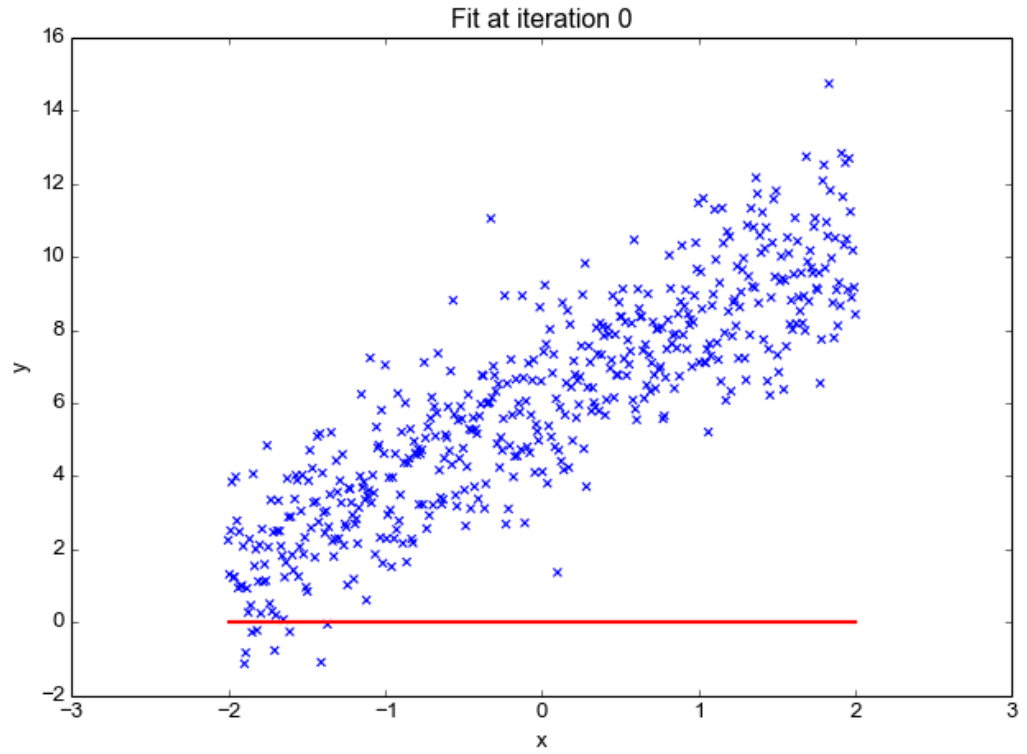
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

客户购买量　基准　性别　月收入　历史购买量　其他

最小二乘法

- 显式解 及 梯度下降法





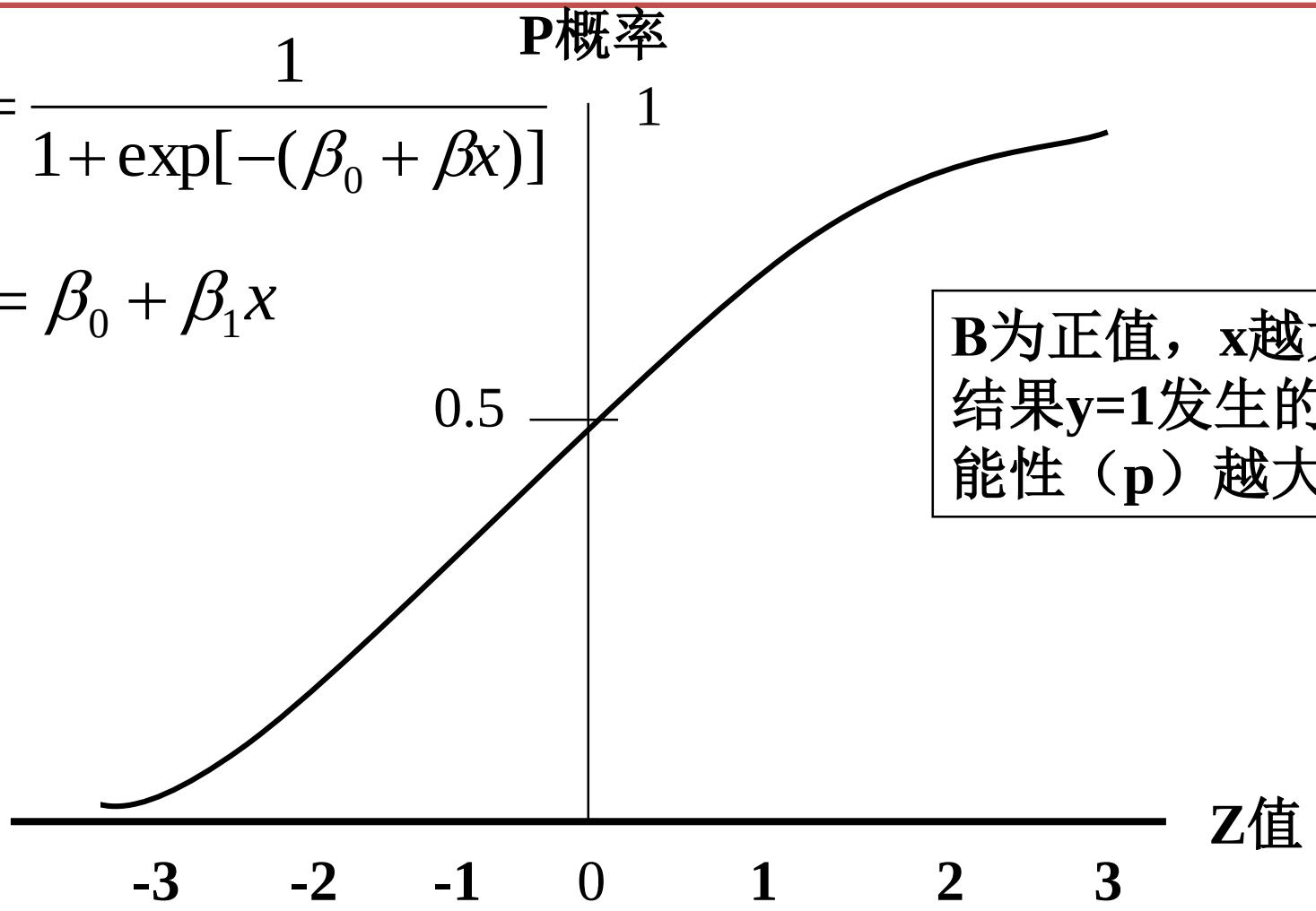
回归分析结果举例：P值与R方

变量名	估计值	P-value	变量名	估计值	P-value
常数项	-0.162	<0.001	当期保养总花费	0.258	<0.001
车型-A	0.001	0.067	当期保养总次数	0.129	<0.001
车型-B	0.260	<0.001	当期新增里程数	0.198	<0.001
车型-C	0.113	0.159	累积购车数量	0.055	<0.001
车型-D	0.176	0.035	车价	0.143	<0.001
车型-其它	-0.094	0.273			
调整R方：36.37%					

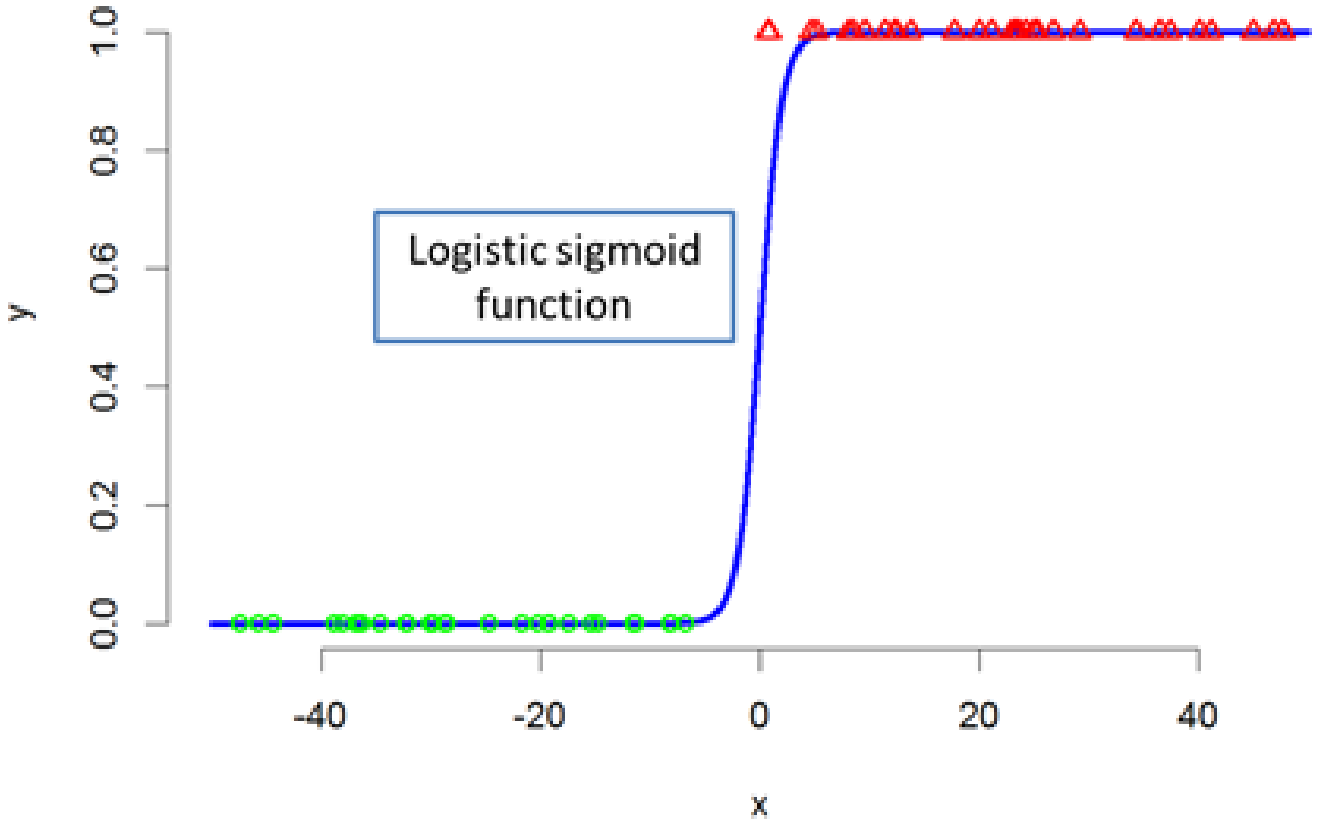
Logistic回归 逻辑回归 逻辑斯迪/蒂克回归

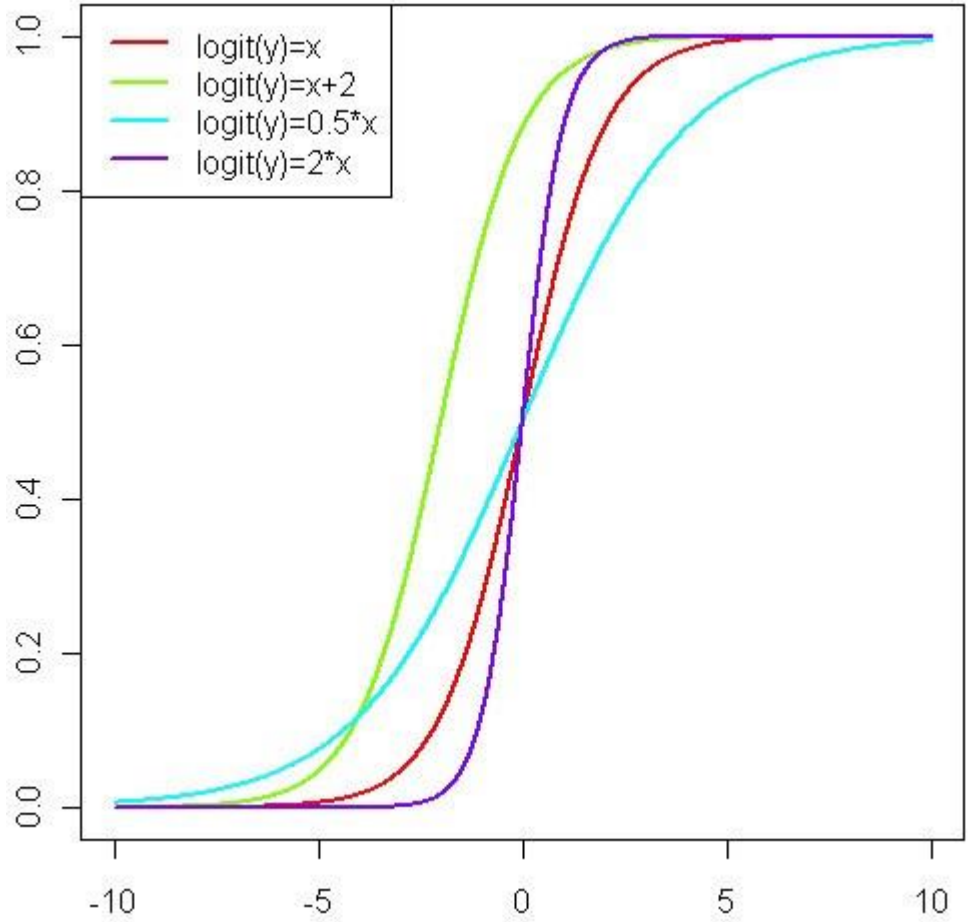
$$p(y=1)=\frac{1}{1+\exp[-(\beta_0+\beta x)]}$$

$$z=\beta_0+\beta_1 x$$



B为正值，x越大，结果y=1发生的可能性（p）越大。





预测值的计算

- 计算公式：
- 得分是一个0-1之间的分数
- 得分=1/(1+exp(-最后得分))

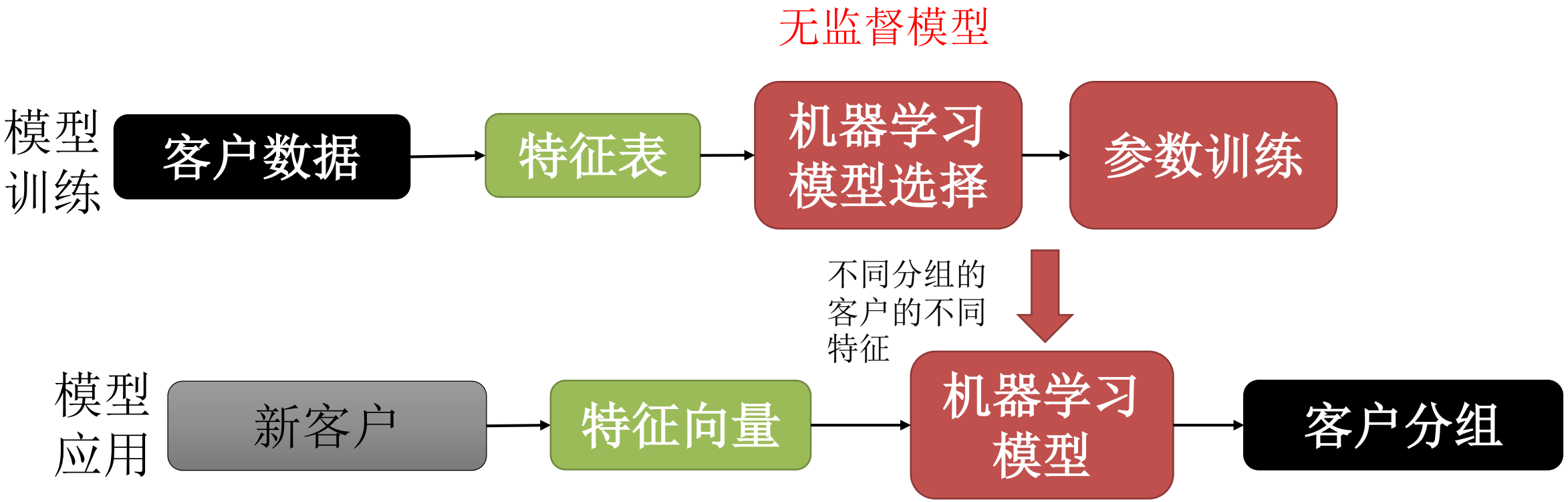
$$p(y=1)=\frac{1}{1+\exp[-(\beta_0+\beta x)]}$$

- 得分>0.5，判定为1，概率得分为原始得分
- 得分<0.5，判定为0，概率得分为1-原始得分

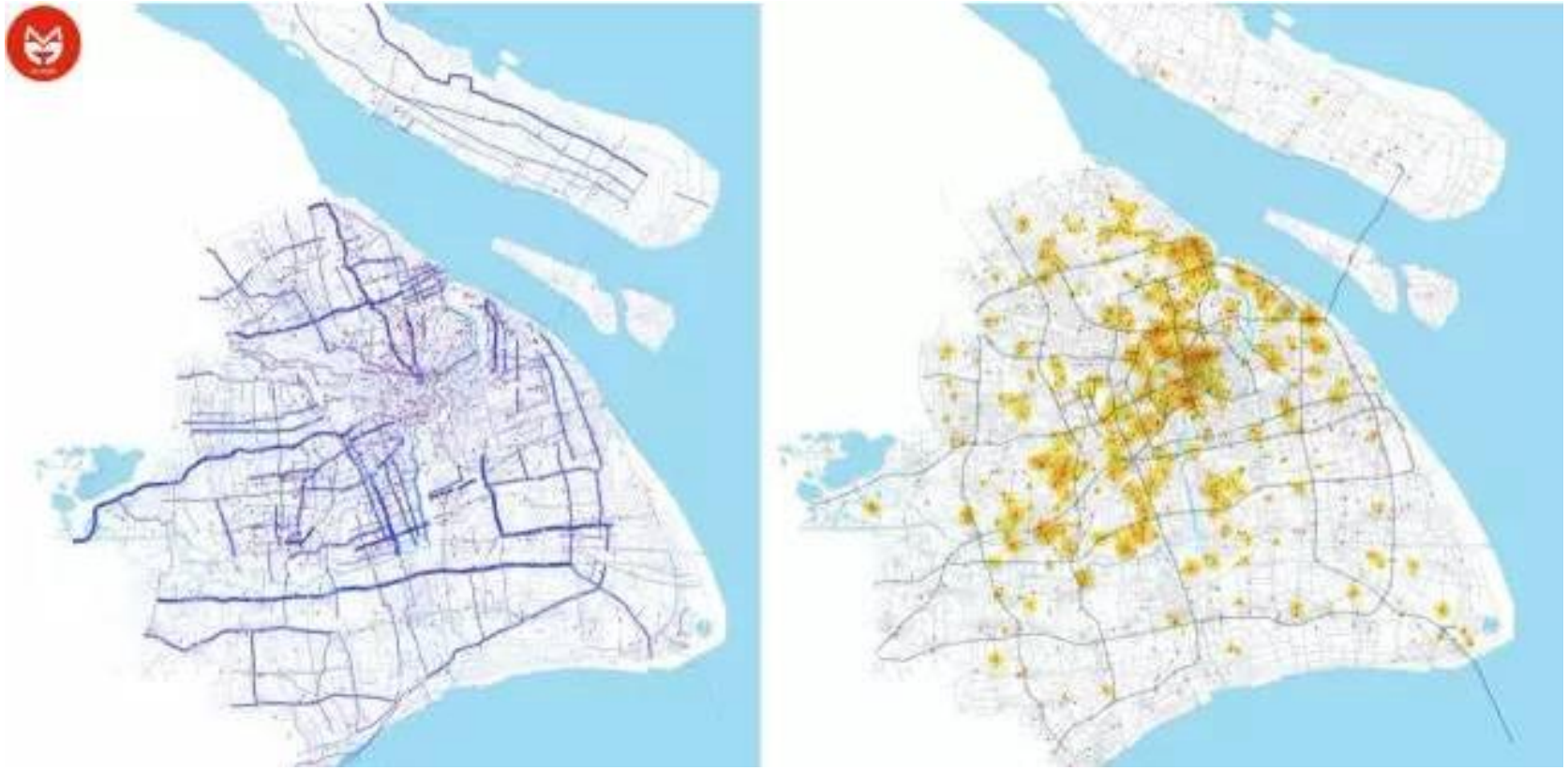
4. 聚类分析

- 聚类分析是对数据进行描述建模型的方法，目的探索数据中是否存在“自然的子类”。
- 聚类分析：对记录进行聚类，使得相同类别的记录之间尽可能相似，不同类别的记录之间尽可能不同。

客户分组研究（没有标签）

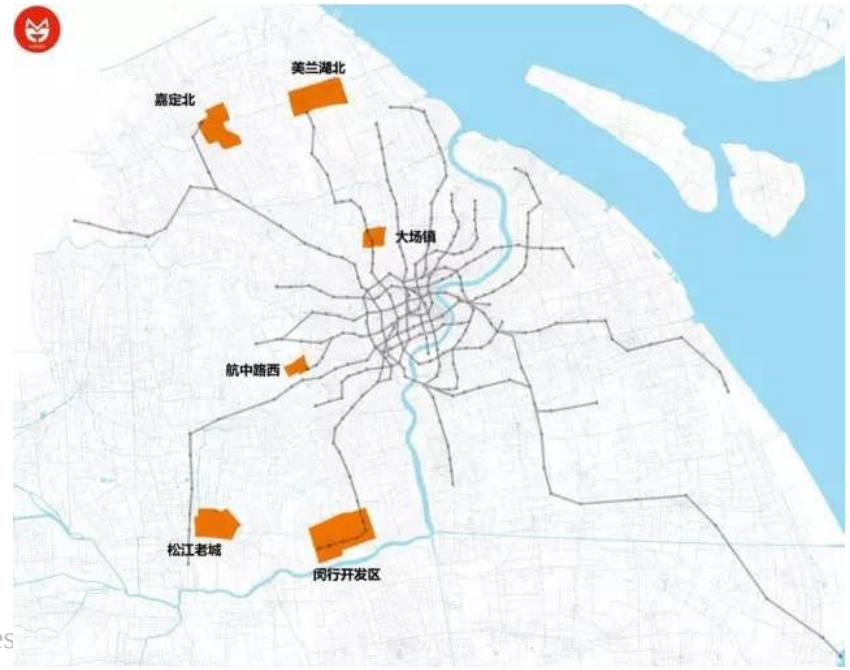


案例：上海市严重交通事故分析



案例：上海市严重交通事故分析

- 上海容易发生交通事故的主要有以下三类地区：
  - 人多车多、出行量大、路况复杂的中心城区；
  - 接驳轨交通勤、助动车盛行的近郊；
  - 高速公路沿线和货运码头附近。



案例：中国社会各阶级的分析

- 《毛泽东选集》
  - 中国社会各阶级的分析，第一卷 3-11
    - 地主阶级和买办阶级——敌人
    - 中产阶级——动摇不定
    - 小资产阶级——最接近的朋友
    - 半无产阶级——最接近的朋友
    - 无产阶级——革命的领导力量
    - 游民无产者——其它
  - 怎样分析农村阶级，第一卷127-129

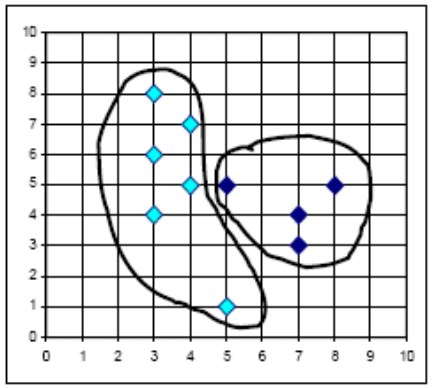
聚类算法的种类

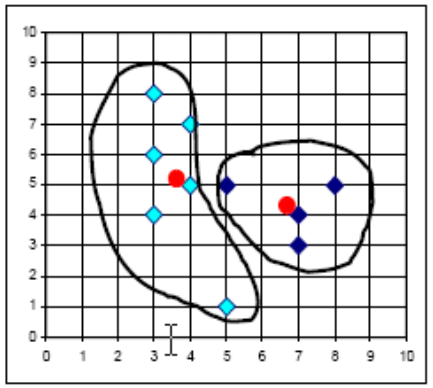
- 从聚类结果角度划分：
  - 覆盖聚类与非覆盖聚类：每个数据点都至少属于一个类，为覆盖聚类，否则为非覆盖聚类
  - 层次聚类和非层次聚类：存在两个类，其中一个类是另一个类的子集，为层次聚类，否则为非层次聚类
  - 确定聚类和模糊聚类：任意两个类的交集为空，一个数据点最多只属于一个类，为确定聚类（或硬聚类）。否则，如果至少一个数据点属于一个以上的类，为模糊聚类

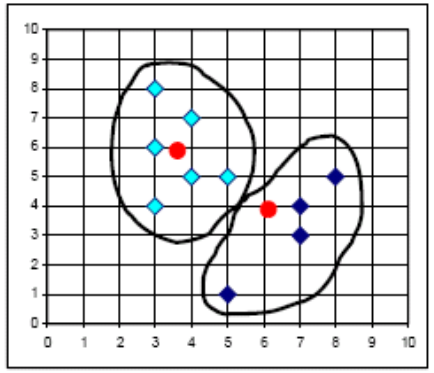
- 从聚类变量类型角度划分
  - 数值型聚类算法、分类型聚类算法、混合型聚类算法
- 从聚类的原理角度划分
  - 划分聚类（Partitional clustering）
  - 层次聚类（Hierarchical clustering）
  - 基于密度的聚类（Density-based clustering）
  - 网格聚类（Rid clustering）

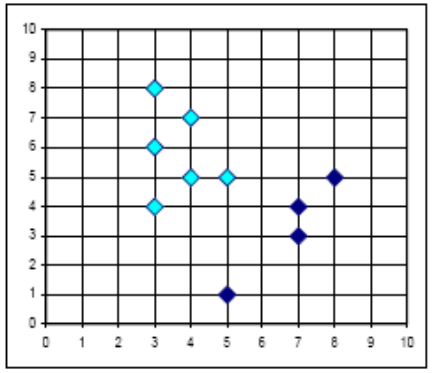
聚类算法：K-均值聚类

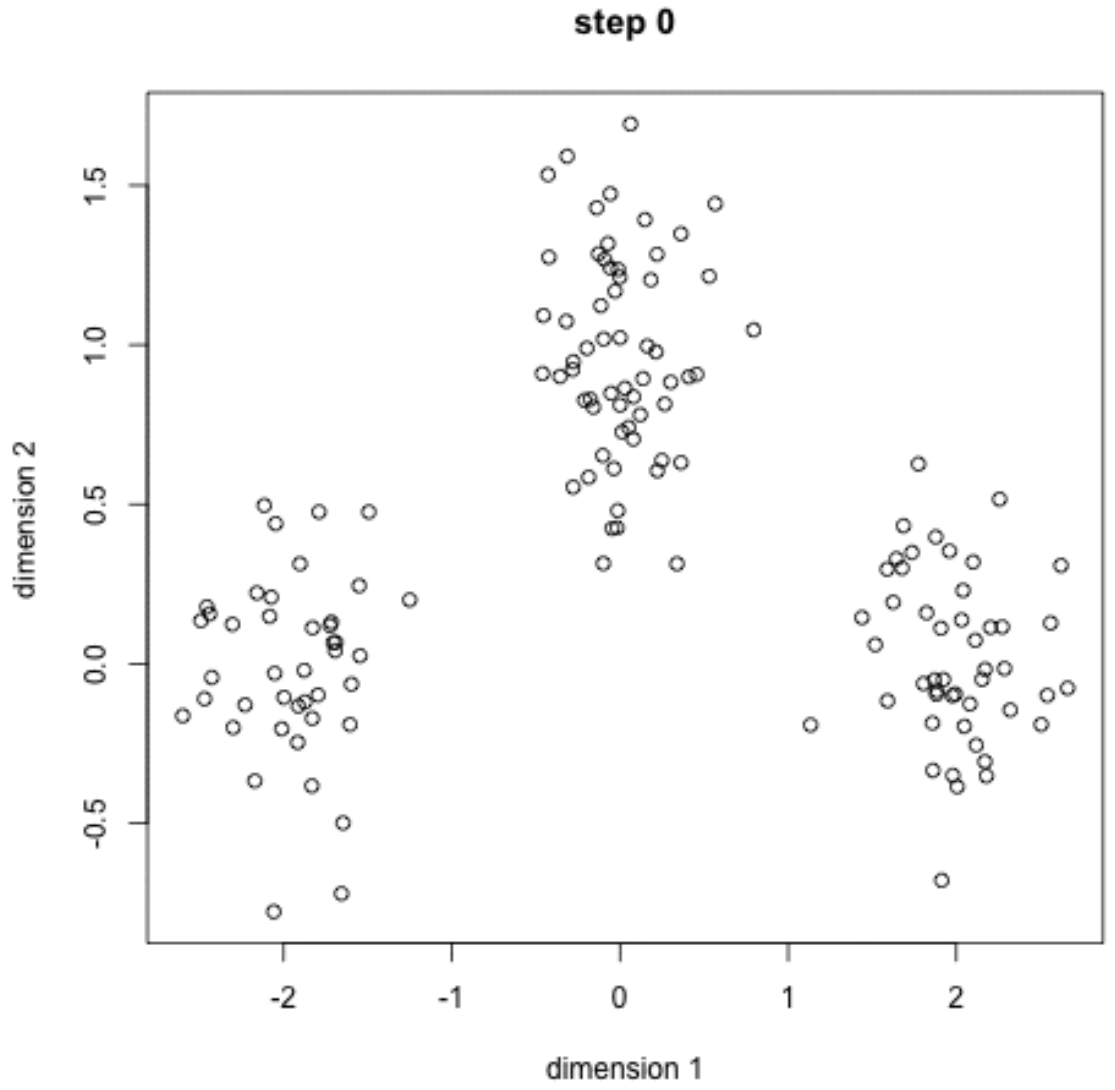
- k-means算法，也被称为k-平均或**k-均值**，是一种得到最广泛使用的聚类算法。
- K-means是将各个聚类子集内的所有数据样本的均值作为该聚类的代表点，算法的主要思想是通过迭代过程把数据集划分为不同的类别，使得评价聚类性能的准则函数达到最优，从而使生成的每个聚类内紧凑，类间独立。
- K-means算法不适合处理离散型属性，但是对于连续型具有较好的聚类效果。







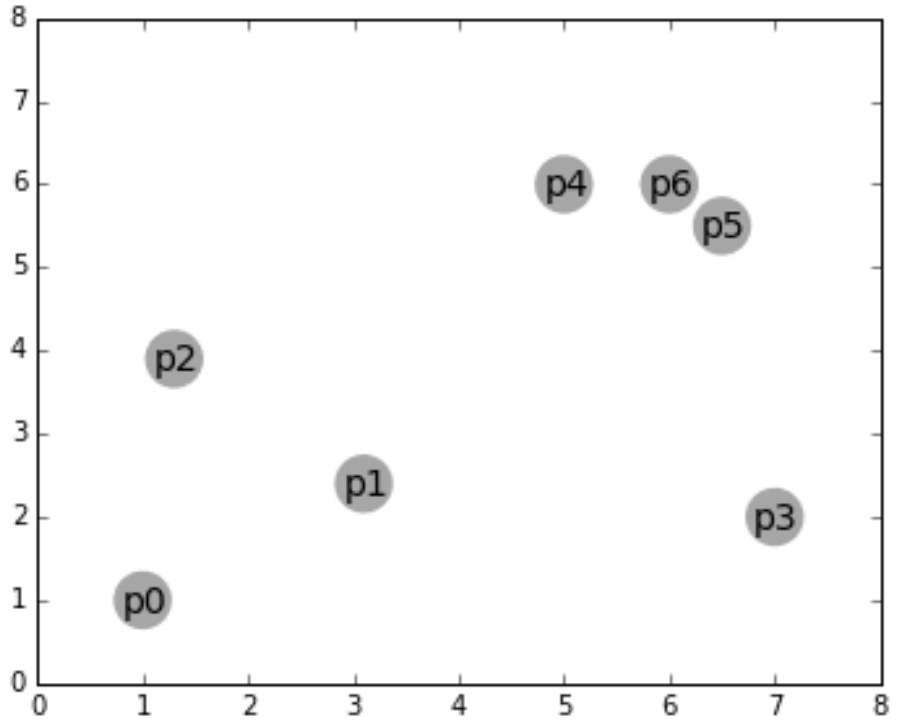


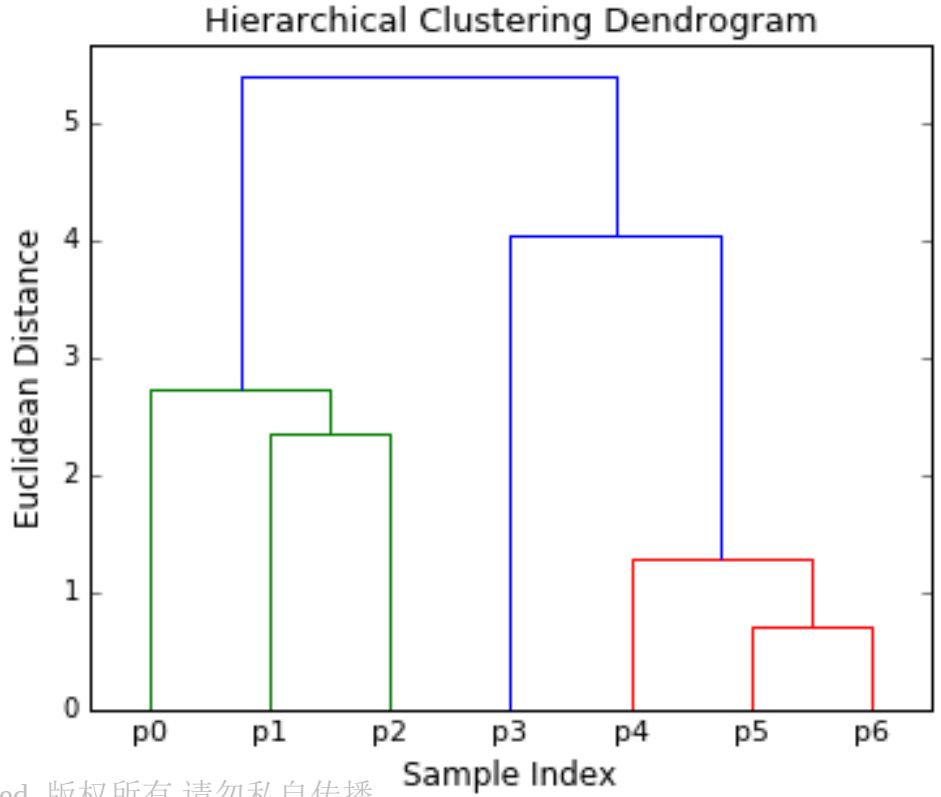


K-均值聚类示例

聚类算法：层次聚类

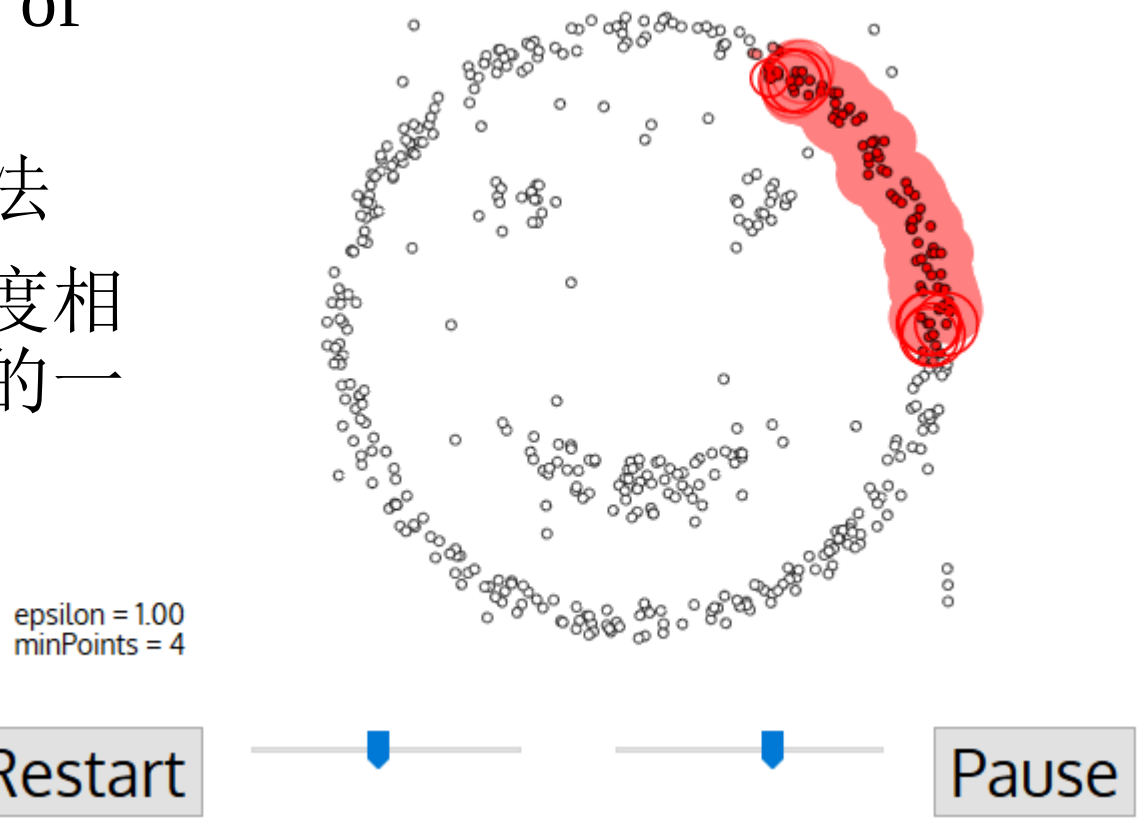
- 逐步合并距离最小的元素/组
- 生成嵌套的类结构，树图





聚类算法：DBScan聚类算法

- Density-Based Spatial Clustering of Applications with Noise
- 考虑噪声的基于密度的聚类方法
- 由密度可达关系导出的最大密度相连的样本集合，即为最终聚类的一个类别，或者说一个簇。

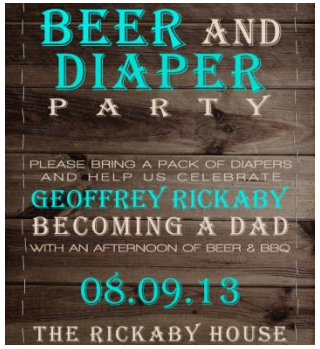


5. 关联规则

- “沃尔玛拥有世界上最大的数据仓库系统，为了能够准确了解顾客在其门店的购买习惯，沃尔玛对其顾客的购物行为进行购物篮分析，想知道顾客经常一起购买的商品有哪些。沃尔玛数据仓库里集中了其各门店的详细原始交易数据。在这些原始交易数据的基础上，沃尔玛利用数据挖掘方法对这些数据进行分析和挖掘。一个意外的发现是：‘跟尿布一起购买最多的商品竟是啤酒！’”

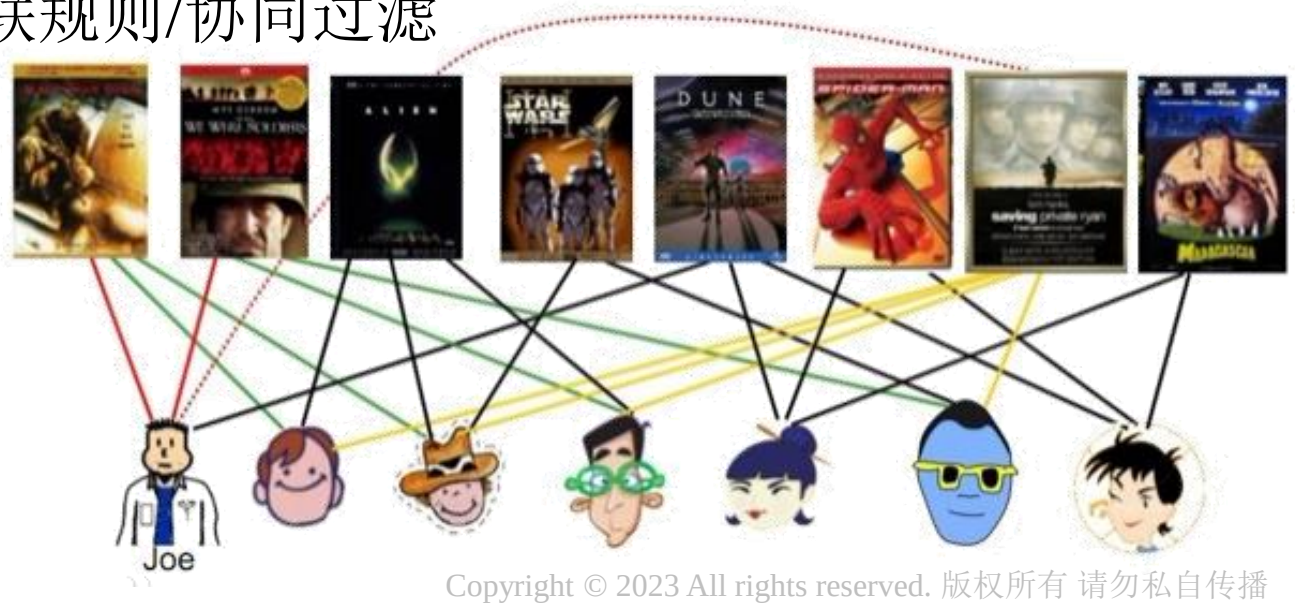
案例：啤酒与尿布（Beer & Diaper）

- 为什么是“啤酒与尿布”？



- 不可全信的 都市传奇（Urban Legend）
  - 沃尔玛/一家超市/7-11
    - *Sometimes the data can throw up surprises: mining of databases held by 7-Eleven stores in the US revealed a link between purchases of beer and nappies. When they were moved together, sales of both increased, says Williams.*
  - 夸大的效果
    - *The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and increased sales on peanuts and pretzels by more that 27%.*

- 互联网内容推荐
  - 为什么刷抖音/快手停不下来
  - 今日头条/新闻推荐
- 互联网商品推荐
  - 关联规则/协同过滤



关注 分享 举报

企业批量购书

基于大数据的商务智能分析

从大数据的角度进行商务智能分析！

[美] 伯特·布瑞吉斯 著，费凡，段世惠，肖春虹，王玲芳 等 译

机械工业出版社 CHINA MACHINE PRESS

金秋风暴 | 金秋 · 无潮不欢

京 东 价 ¥ 58.50 [7.5折] [定价 ~~¥ 78.00~~] 降价通知　累计评价 93

优 惠 券 满300减20

促　　销 加价购 满12元另加26.90元，或满15元另加16.90元，或满18元另加9.90元，即可在购物车换购热销商品 详情>>

增值业务 助力环保，传递知识，旧书换新

配 送 至 云南昆明市西山区碧鸡街道 有货

京东物流 预约送货 | 部分收货 | 送货上门

由 京东 发货，并提供售后服务. 11:10前下单，预计明天(09月25日)送达

重　　量 0.4kg

服务支持 放心购 上门换新 | 破损包退换 | 闪电退款

可配送海外49元免基础运费

增值保障 意外换新 ¥2.50 | 2年爱心收 ¥1.00

白条分期 不分期 | ¥19.79 x 3期 | ¥10.04 x 6期 | ¥5.11 x 12期 | ¥2.67 x 24期

1 加入购物车

温馨提示 · 支持7天无理由退货

木棉图书旗舰店	¥40.40
博库网旗舰店	¥40.56
万里路图书专营店	¥53.80

6个卖家在售

人气单品 | 七日畅销榜 | 新书热卖榜


R语言：从数据思维到数据实战
¥70.30

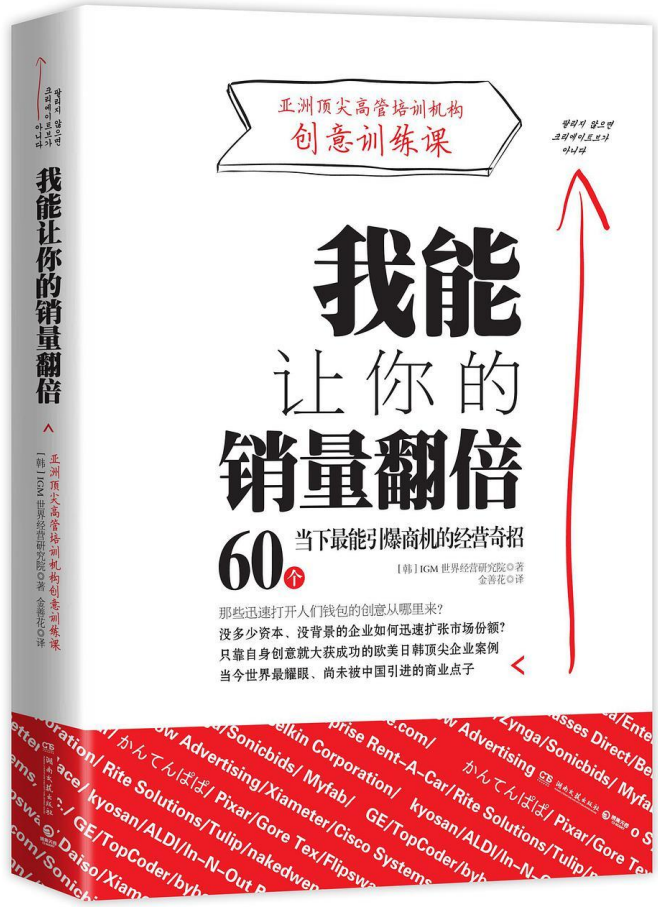

商务智能与分析：决策支持系统（原书第10
¥119.10

商务智能（第四版）/清华科技大讲堂
¥49.00


大数据地理信息系统：原理、技术与应用
¥51.60

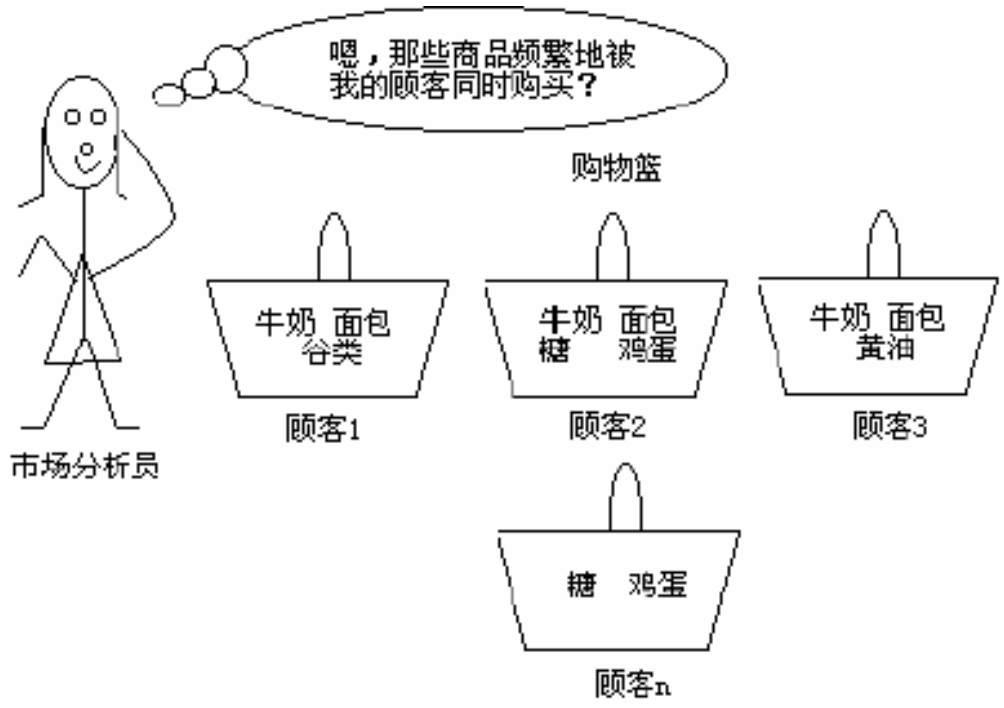
大数据时代（大数据系统研究的先河之作）
¥33.30


商业数据分析
¥83.70



简单关联规则核心概念

- 目的：寻找事物之间的联系规律，发现它们之间的关联关系
- 关联关系包括：简单关联关系、序列关联关系
- 关联分析的主要技术是关联规则（Association Rule）
- 最早用于研究超市顾客购买商品之间的规律，称为购物篮分析
- 无监督学习方法



面包→牛奶（S=85%, C=90%）

前项

后项

支持度

置信度

前项与后项

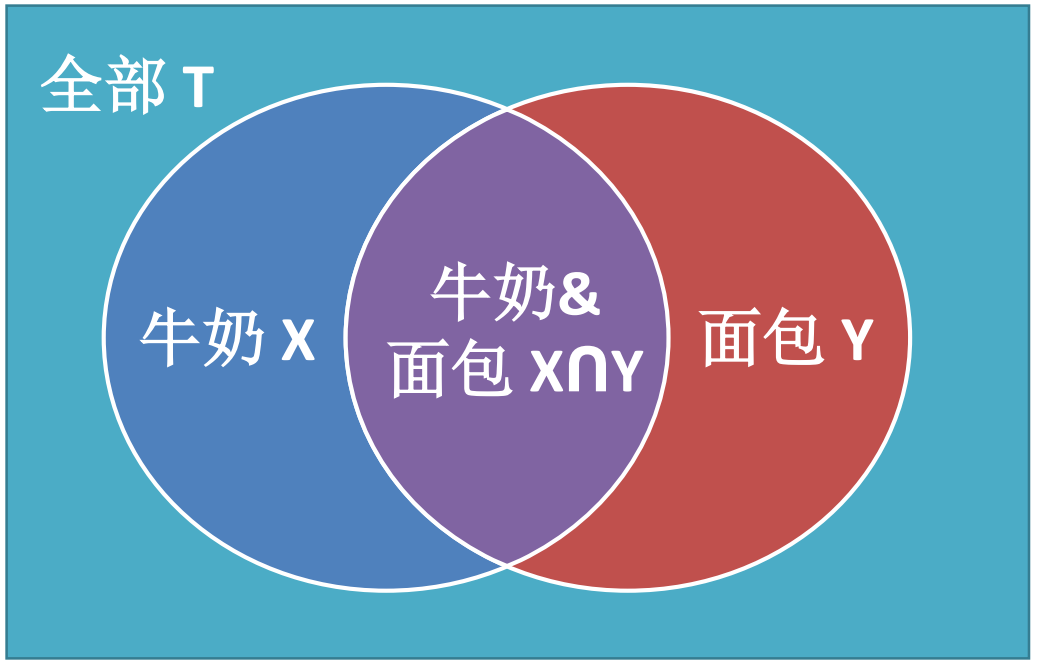
- 简单关联规则的一般表示形式：
- X→Y（规则支持度，规则置信度）

- X为规则的前项，可为项目或项集或包含逻辑与（∩）或（∪）非（ ¬ ）的逻辑表达式
- Y为规则的后项，一般为一个项目，表示某种结论或事实

- 举例：
  - 面包→牛奶
  - 性别(女) ∩ 收入(>5000)→品牌(A)

有效性的测度指标

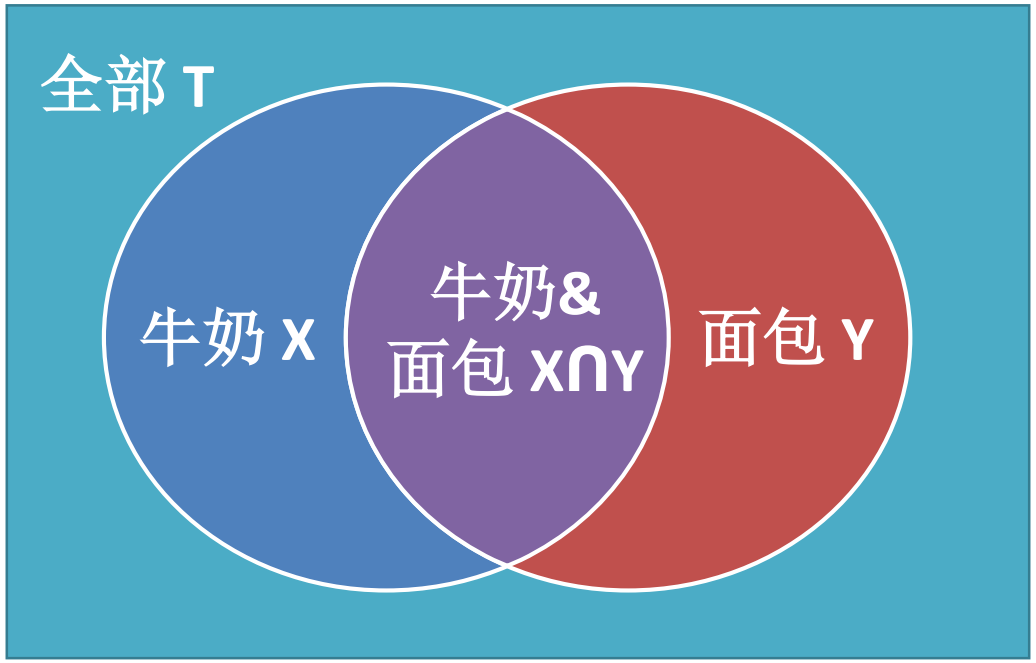
- 规则**置信度(Confidence)**：对准确度的测量，描述了包含项目X的事务中同时包含项目Y项的概率，反映X出现条件下Y出现的可能性
- 置信度高说明X出现则Y出现的可能性高
- 面包→牛奶（S=85%,C=90%），表示购买面包则同时购买牛奶的可能性为90%



$$C_{X \to Y} = \frac{|T(X \cap Y)|}{|T(X)|}$$

条件概率

- 规则支持度（Support）：测度了规则的普遍性，是项目X和项目Y项同时出现的概率
- **面包→牛奶（S=85%, C=90%）**表示顾客中同时购买面包和牛奶的概率为85%



$$S_{X \to Y} = \frac{|T(X \cap Y)|}{|T|}$$

模型扩展2：多层关联规则挖掘

- 对于很多的应用来说，由于数据分布的分散性，所以很难在数据最细节的层次上发现一些强关联规则。当我们引入概念层次后，就可以在较高的层次上进行挖掘[HF95, SA95]。虽然较高层次上得出的规则可能是更普通的信息，但是对于一个用户来说是普通的信息，对于另一个用户却未必如此。所以数据挖掘应该提供这样一种在多个层次上进行挖掘的功能。
- 多层关联规则的分类：根据规则中涉及到的层次，多层关联规则可以分为同层关联规则和层间关联规则。
- 多层关联规则的挖掘基本上可以沿用“支持度-可信度”的框架。不过，在支持度设置的问题上有一些要考虑的东西。

模型扩展3：多维关联规则挖掘

- 对于多维数据库而言，除维内的关联规则外，还有一类多维的关联规则。例如：
  - 年龄（“20...30”） 职业（“学生”）→ 购买(“笔记本电脑”)
  - 涉及到三个维上的数据：年龄、职业、购买。

- 根据是否允许同一个维重复出现，可以又细分为维间的关联规则（不允许维重复出现）和混合维关联规则（允许维在规则的左右同时出现）。
  - 年龄（“20...30”） 购买(“笔记本电脑”) → 购买(“打印机”)

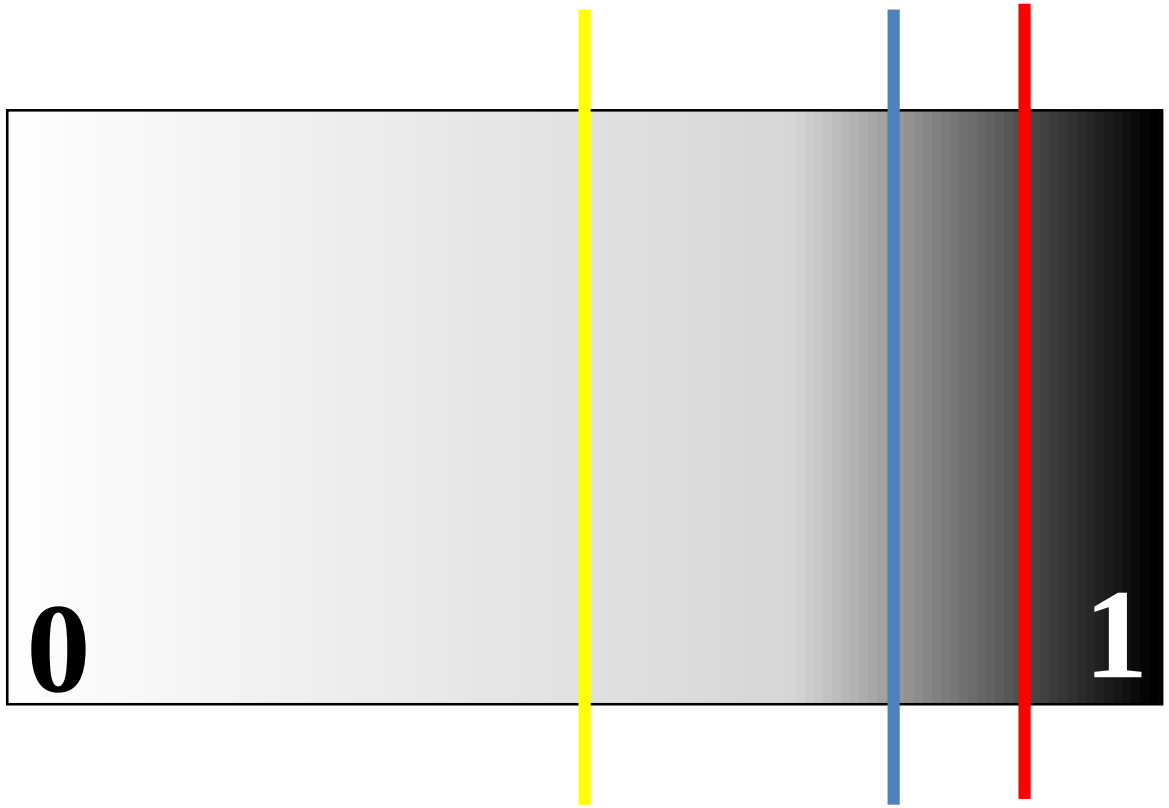
4. 机器学习模型的评估及弱点
Evaluation of Machine Learning Models

1. 混淆矩阵（Confusion Matrix）

混淆矩阵 Confusion Matrix		预测情况		求和
		1 Positive	0 Negative	
真实情况	1	TP	FN	TP+FN
	0	FP	TN	FP+TN
求和		P	N	总样本数

分类器的本质

- 按照分数排队，决定阈值，阈值以上为Positive



举例：假设逻辑回归模型输出0.9及以上认为购买

混淆矩阵 Confusion Matrix		预测情况 模型得分>=0.9?		求和
		分数>=0.9 Positive	分数<0.9 Negative	
真实情况 是否购买	购买	80 (TP)	20 (FN)	100 (TP+FN)
	不购买	400 (FP)	500 (TN)	900 (FP+TN)
求和		480 (P)	520 (N)	1000 (总样本数)

准确率、查准率（Precision）

- 总体正确率(Accuracy)=(TP+TN)/总样本数=580/1000=0.58
- **预测为1的准确率(Precision)=TP/P=80/480=0.17**

混淆矩阵 Confusion Matrix		预测情况 模型得分>=0.9?		求和
		分数>=0.9 Positive	分数<0.9 Negative	
真实情况 是否购买	购买	**80 (TP)**	20 (FN)	100 (TP+FN)
	不购买	400 (FP)	**500 (TN)**	900 (FP+TN)
求和		**480 (P)**	520 (N)	**1000 (总样本数)**

模型更好还是更差？

- 正确率=(TP+TN)/总样本数=810/1000=0.81
- 准确率=TP/P=10/110=0.09

Accuracy 不重要
我们真正关心的是准确率
Precision 重要

混淆矩阵 Confusion Matrix		预测情况 模型得分>=0.9?		求和
		分数>=0.9 Positive	分数<0.9 Negative	
真实情况 是否购买	购买	**10 (TP)**	90 (FN)	100 (TP+FN)
	不购买	100 (FP)	**800 (TN)**	900 (FP+TN)
求和		**110 (P)**	890 (N)	**1000 (总样本数)**

召回率、查全率（Recall）

- 召回率(Recall)=TP/(TP+FN)=80/100=0.8
- 准确率(Precision)=TP/P=80/480=0.17

混淆矩阵 Confusion Matrix		预测情况 模型得分>=0.9?		求和
		分数>=0.9 Positive	分数<0.9 Negative	
真实情况 是否购买	购买	**80 (TP)**	20 (FN)	**100 (TP+FN)**
	不购买	400 (FP)	500 (TN)	900 (FP+TN)
求和		480 (P)	520 (N)	1000 (总样本数)

模型更好还是更差？

- 召回率(Recall)=TP/(TP+FN)=90/100=0.9
- 准确率(Precision)=TP/P=90/690=0.13

召回率 Recall
和准确率 Precision
不得两全 此消彼长

混淆矩阵 Confusion Matrix		预测情况 模型得分>=0.8?		求和
		分数>=0.8 Positive	分数<0.8 Negative	
真实情况 是否购买	购买	**90 (TP)**	10 (FN)	**100 (TP+FN)**
	不购买	600 (FP)	300 (TN)	900 (FP+TN)
求和		690 (P)	310 (N)	1000 (总样本数)

如何决定模型的阈值？

- 选择 得分>0.9 还是 得分>0.8？

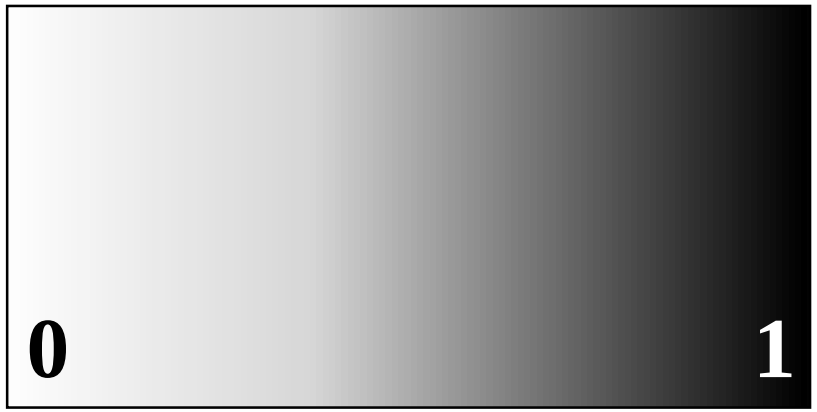
- 理论方法——F1得分 $F_1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$
- 例：
- 得分>0.9时，F1=2*0.17*0.8/(0.17+0.8)=0.28
- 得分>0.8时，F1=2*0.13*0.9/(0.13+0.9)=0.23

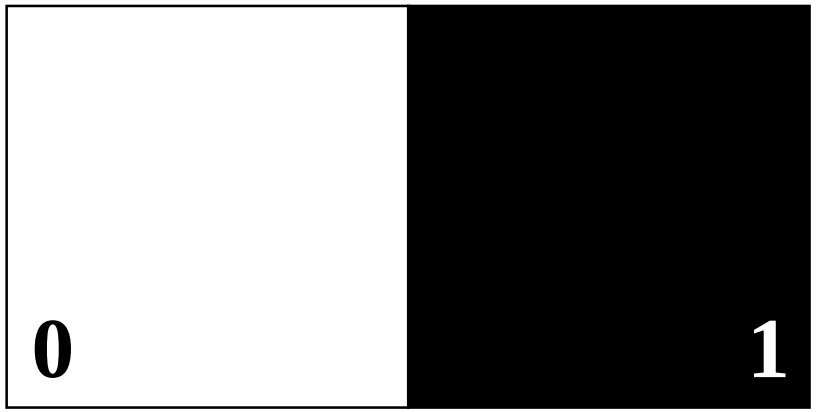
- **其实没必要**

- 根据具体的情况，决定工作范围。
- 范围大，成果多，准确率低
- 范围小，成果少，准确率高

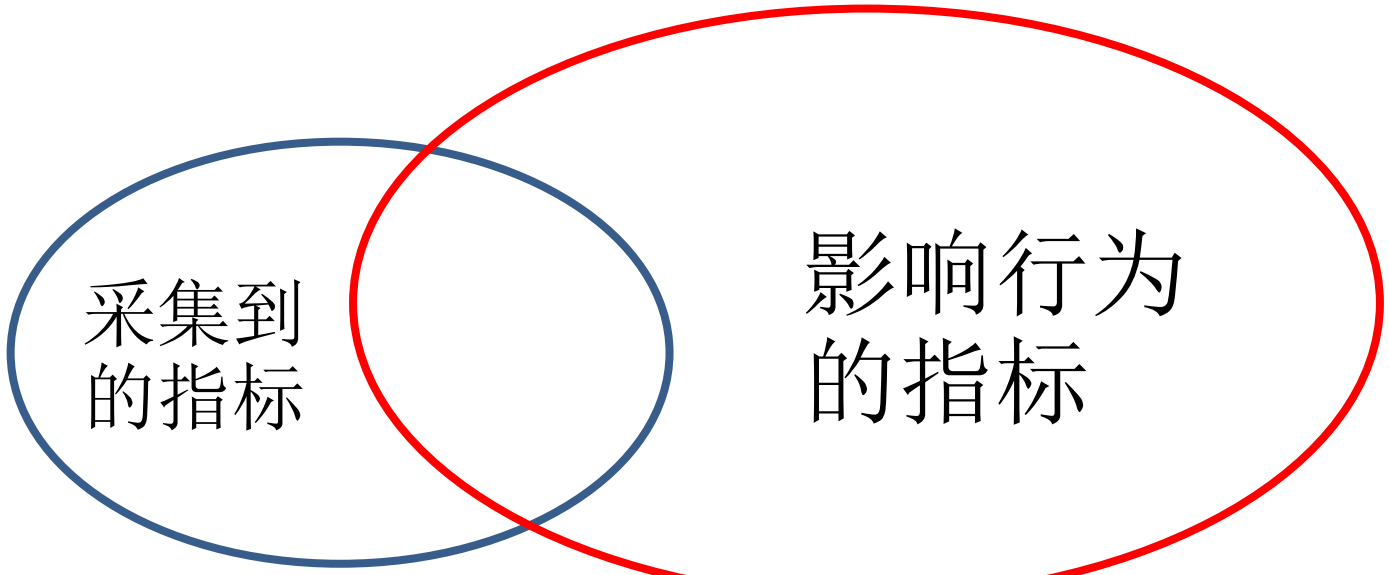
逻辑回归/其它机器学习模型能不能做到100%准确？

- 有没有一个足够牛的模型，可以黑白分明？

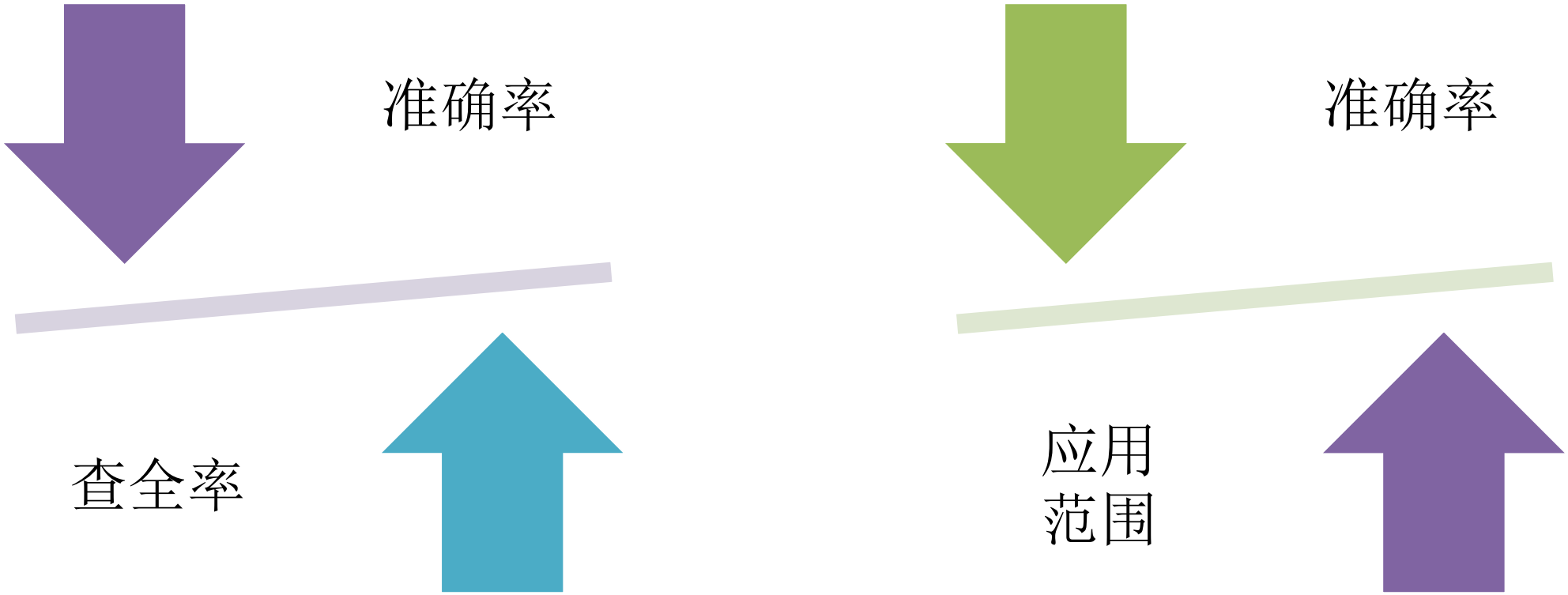




- 在模型训练的层面，比较容易！！！
  - 用采集到的指标反复训练模型，容易达到模型与训练数据的100%契合
- 在模型实用的层面，极其困难！！！
  - 真实的指标浩如烟海，不可能采集完全，也不可能训练出这样的模型



模型的权衡



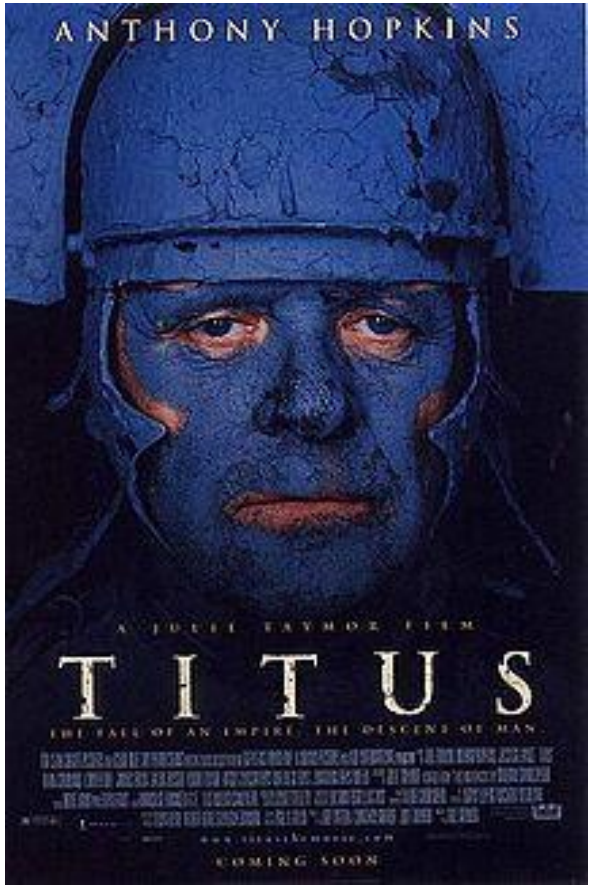
5. 机器学习与商业模式的结合
Machine Learning and Business Models

1. 案例：差别定价实验

amazon.com and you're done

~~$26.24~~ $22.74

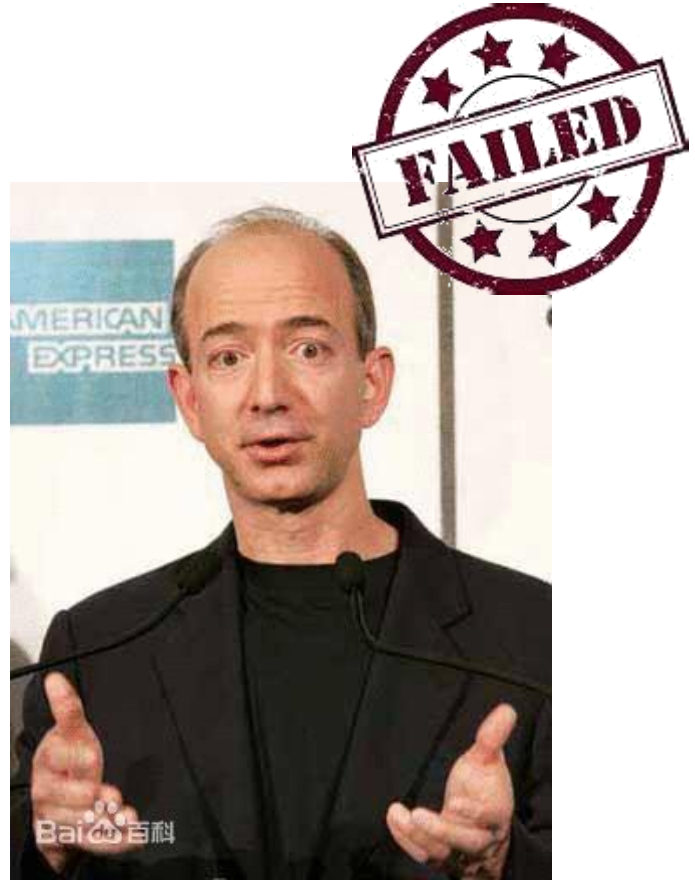
- Streitfeld, David. "On the web, price tags blur: What you pay could depend on who you are." *The Washington Post* September 27 (2000).
- 亚马逊选择了68种DVD碟片进行动态定价试验，试验当中，亚马逊根据潜在客户的人口统计资料、在亚马逊的购物历史、上网行为以及上网使用的软件系统确定对这68种碟片的报价水平。
- 名为《泰特斯》（Titus）的碟片对新顾客的报价为22.74美元，而对那些对该碟片表现出兴趣的老顾客的报价则为26.24美元。
- 通过这一定价策略，部分顾客付出了比其他顾客更高的价格，亚马逊因此提高了销售的毛利率。



- 差别定价策略实施不到一个月，就有细心的消费者发现了这一秘密：
  - “一个亚马逊的用户很偶然地发现：他想买的DVD售价$26.24，但是当他把电脑的上网记录清理干净，然后再登录亚马逊的时候，同一款DVD的价格变成了$22.74。”
- 通过在名为DVDTalk (www.dvdtalk.com)的音乐爱好者社区的交流，成百上千的DVD消费者知道了此事
  - 那些付出高价的顾客当然怨声载道，纷纷在网上以激烈的言辞对亚马逊的做法进行口诛笔伐。
  - 有人公开表示以后绝不会在亚马逊购买任何东西。

- 亚马逊前不久才公布了它对消费者在网站上的购物习惯和行为进行了跟踪和记录。
- 这次事件曝光后，消费者和媒体开始怀疑亚马逊是否利用其收集的消费者资料作为其价格调整的依据，这样的猜测让亚马逊的价格事件与敏感的网络隐私问题联系在了一起。

- 亚马逊的首席执行官贝佐斯只好亲自出马做危机公关，他指出亚马逊的价格调整是随机进行的，与消费者是谁没有关系，价格试验的目的仅仅是为测试消费者对不同折扣的反应，亚马逊“**无论是过去、现在或未来，都不会利用消费者的人口资料进行动态定价。**”
- 贝佐斯为这次的事件给消费者造成的困扰向消费者公开表示了道歉。不仅如此，亚马逊还试图用实际行动挽回人心，亚马逊答应给所有在价格测试期间购买这68部DVD的消费者以最大的折扣，据不完全统计，至少有6896名没有以最低折扣价购得DVD的顾客，已经获得了亚马逊退还的差价。



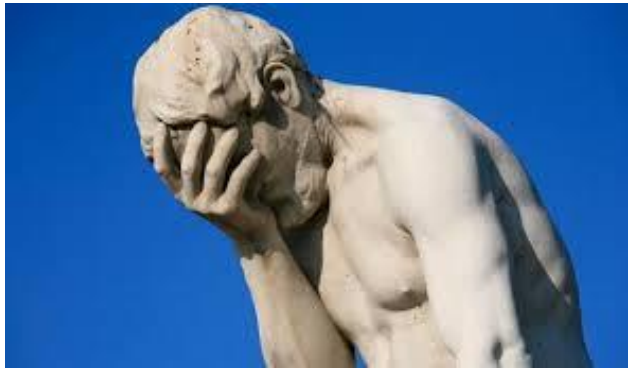
2. 案例：Farecast预测机票票价

- Etzioni, O., R. Tuchinda, C. A. Knoblock and A. Yates (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
- 维克托•迈尔-舍恩伯格《大数据时代：生活、工作与思维的大变革》

Oren Etzioni
University of Washington
Computer Science & Engineering
Director of Turing Center

UNIVERSITY OF WASHINGTON

- “2003年，奥伦·埃齐奥尼（Oren Etzioni）准备乘坐从西雅图到洛杉矶的飞机去参加弟弟的婚礼。他知道飞机票越早预订越便宜，于是他在这个大喜日子来临之前的几个月，就在网上预订了一张去洛杉矶的机票。”
- “在飞机上，埃齐奥尼好奇地问邻座的乘客花了多少钱购买机票。当得知虽然那个人的机票比他买得更晚，但是票价却比他便宜得多时，他感到非常气愤。于是，他又询问了另外几个乘客，结果发现大家买的票居然都比他的便宜。”



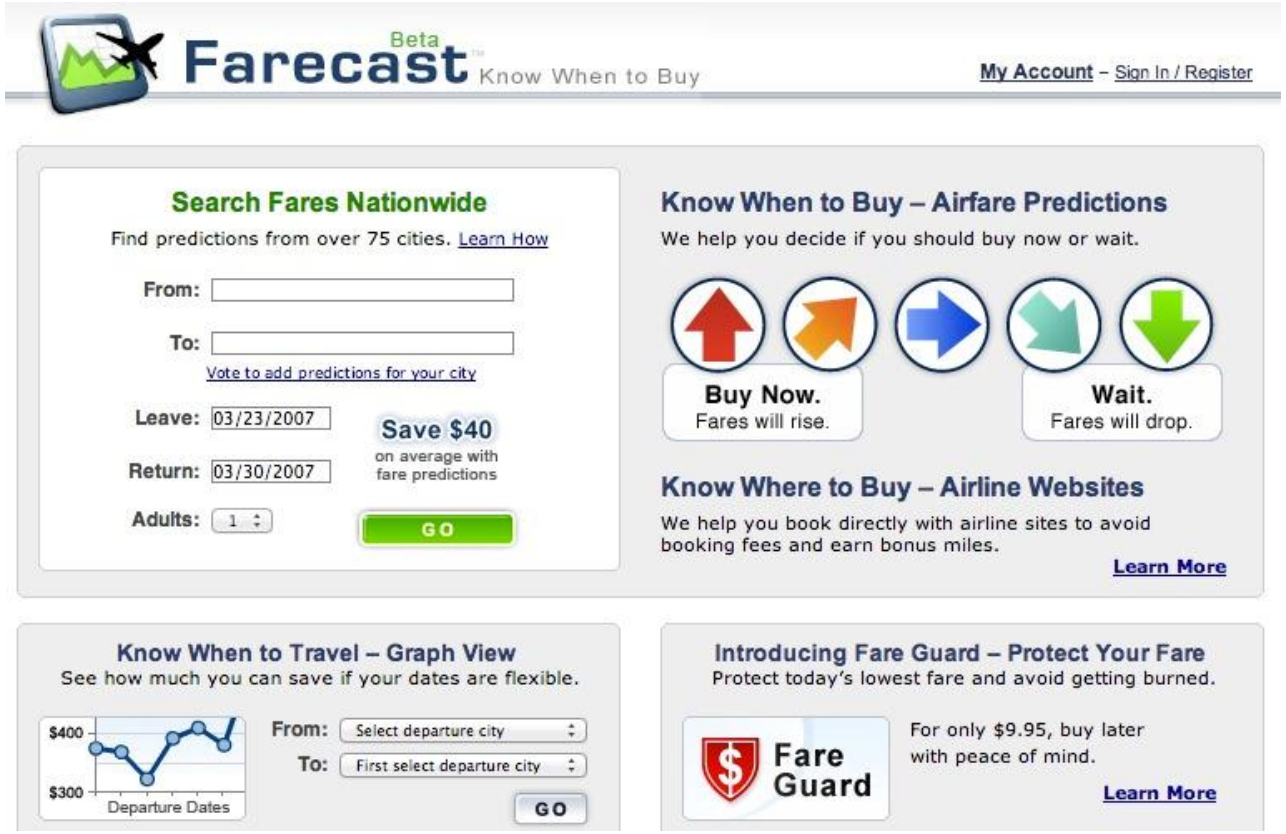
- Etzioni, O., R. Tuchinda, C. A. Knoblock and A. Yates (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. Proceedings of the ninth **ACM SIGKDD** international conference on Knowledge discovery and data mining, ACM.

This paper reports on a pilot study in the domain of airline ticket prices where we recorded over 12,000 price observations over a 41 day period. When trained on this data, Hamlet — our multi-strategy data mining algorithm — generated a predictive model that saved 607 simulated passengers $283,904 by advising them when to buy and when to postpone ticket purchases. Remarkably, a clairvoyant algorithm with complete knowledge of future prices could save at most $320,572 in our simulation, thus HAMLET's savings were 88.6% of optimal. The algorithm's savings of $283,904 represents an average savings of 27.1% per simulated passenger for whom savings are possible. Our pilot study suggests that mining of price data available over the web has the potential to save consumers substantial sums of money per annum, at least until corporations begin to fight back.

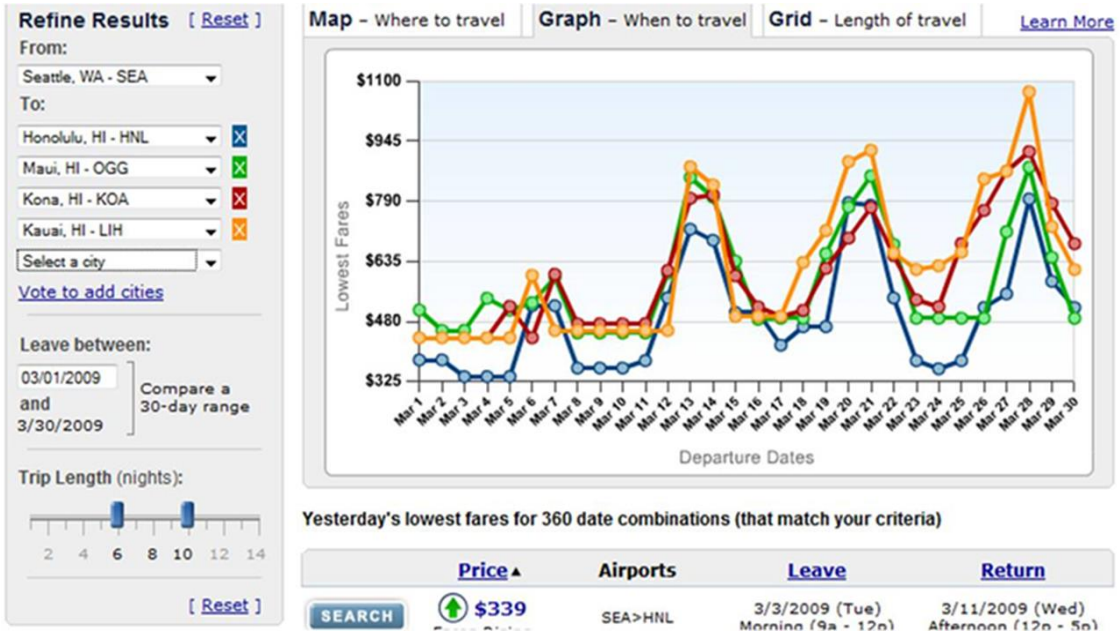
- “埃齐奥尼创立了一个预测系统，它帮助虚拟的乘客节省了很多钱。这个预测系统建立在41天之内的12000个价格样本基础之上，而这些数据都是从一个旅游网站上爬取过来的。”
- “这个预测系统并不能说明原因，只能推测会发生什么。也就是说，它不知道是哪些因素导致了机票价格的波动。机票降价是因为有很多没卖掉的座位、季节性原因，还是所谓的“周六晚上不出门”，它都不知道。这个系统只知道利用其他航班的数据来预测未来机票价格的走势。”

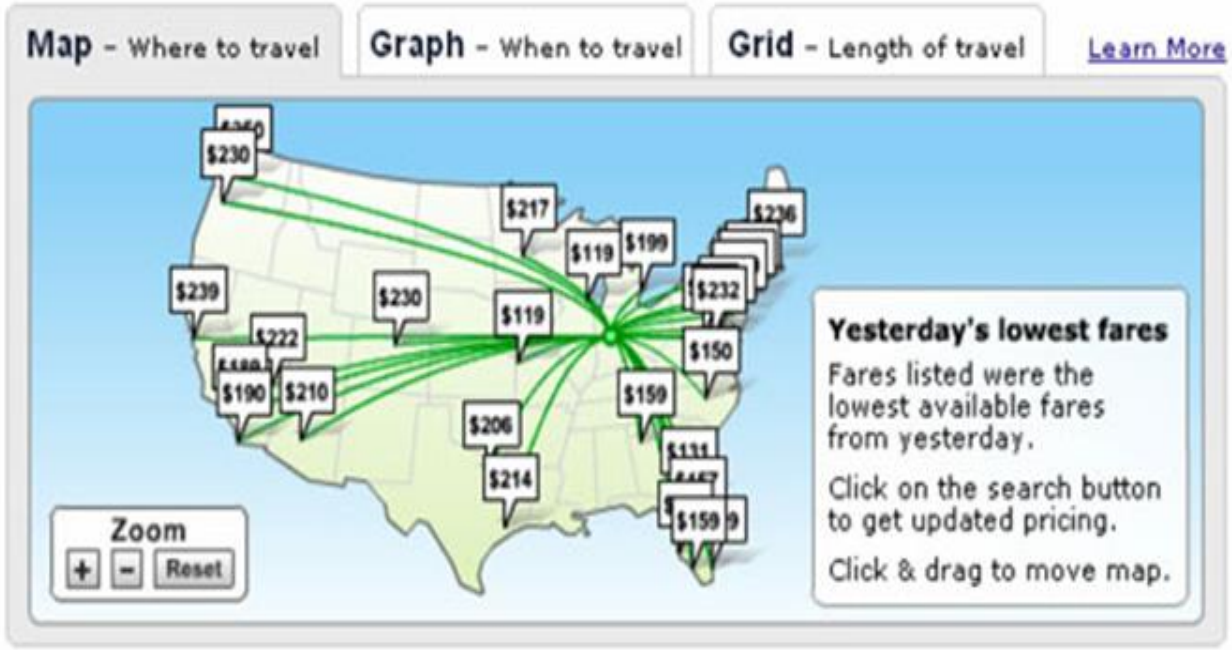
Farecast
Know When to Buy

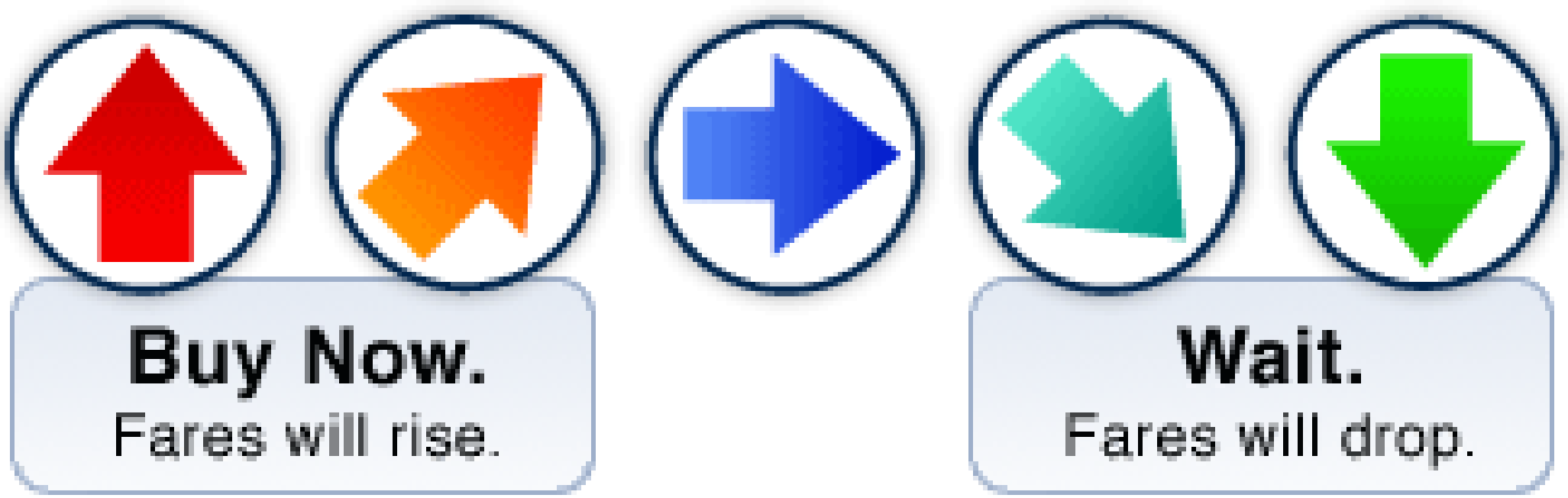
- “这个小项目逐渐发展成为一家得到了风险投资基金支持的科技创业公司，名为Farecast。通过预测机票价格的走势以及增降幅度，Farecast票价预测工具能帮助消费者抓住最佳购买时机，而在此之前还没有其他网站能让消费者获得这些信息。”



- 技术的接受度、实用性







大数据的“阿基里斯之踵”：准确率

谢谢！
Thank you for your attention.

liuyuewen@xjtu.edu.cn