

华东理工大学
EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

CRM介绍

华东理工大学计算机系 翟洁

数据仓库

定义：
数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合。

- (1) 面向主题——面向某一决策问题而制定的。
- (2) 数据不可修改。
- (3) 集成的数据——消除不一致和错误的地方。
- (4) 数据随时间不断变化。

客户基本信息库

客户编号
客户姓名
客户地址

客户订货信息库

订货单编号
订货日期

事实表

客户编号
产品编号
订货单编号
时间标识
地区标识
销售商

产品信息库

产品编号
品牌
产品名称
产品型号

地理库

地区标识
国家
地区
城市

时间维表

时间标识
时间
日期
月
年

数据挖掘的主要策略

聚合

一种用来使记录子集聚集在一起的技术。可用于客户群细分或发现高潜在的销售机会。

CRM系统概述

核心理念：

——以客户为中心提供个性化的服务以降低客户的流失率，通过实现客户效用的最大化获得最大利润。p118

最高境界：

——为客户创造一生的最大价值。

例如：

某公司有一百万客户资料。

- 20万是重要用户（忠诚度高），就构成了每年2亿的忠诚客户基础。
- 80万是一般的客户，就构成了每年8亿的忠诚客户基础。

数据挖掘

数据挖掘(Data Mining)，是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式。

例如

- 基于规则的数据挖掘

——一天的不同时刻和不同产品网上购买量之间的关系。

- 协同过滤

——推荐“亲密群体”中其他消费者购买的产品。

数据挖掘的主要策略

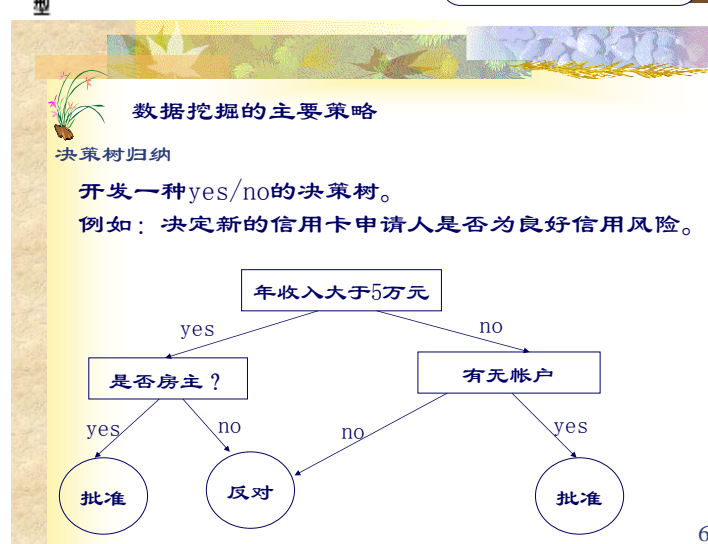
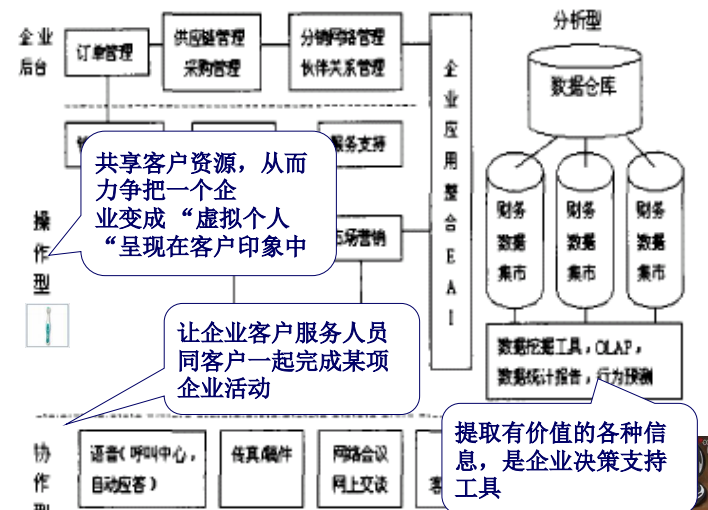
顺序关联——找出根据时间将事件连接起来的关联。

支票帐户+三个月后储蓄帐户---6个月后信用卡24%

关联

——关联算法能够发现一套项目与另一套项目的相互关联的所有规则。

55%----75%



决策树的算法

如何制定分类标准?

rid	age	income	student	Credit-rating	Buys-computer
1	<30	High	No	fair	No
2	<30	High	No	excellent	No
3	30-40	High	No	fair	yes
4	>40	Medium	No	fair	yes
5	>40	Low	yes	fair	yes
6	>40	Low	yes	excellent	No
7	30-40	Low	yes	excellent	yes
8	<30	Medium	No	fair	No
9	<30	Low	yes	fair	yes
10	>40	Medium	yes	fair	yes
11	<30	Medium	yes	excellent	yes
12	30-40	Medium	No	excellent	yes
13	30-40	High	yes	fair	yes
14	>40	Medium	No	excellent	No

rid	age	income	student	Credit-rating	Buys-computer
1	<30	High	No	fair	No
2	<30	High	No	excellent	No
3	30-40	High	No	fair	yes
4	>40	Medium	No	fair	yes
5	>40	Low	yes	fair	yes
6	>40	Low	yes	excellent	No
7	30-40	Low	yes	excellent	yes
8	<30	Medium	No	fair	No
9	<30	Low	yes	fair	yes
10	>40	Medium	yes	fair	yes
11	<30	Medium	yes	excellent	yes
12	30-40	Medium	No	excellent	yes
13	30-40	High	yes	fair	yes
14	>40	Medium	No	excellent	No

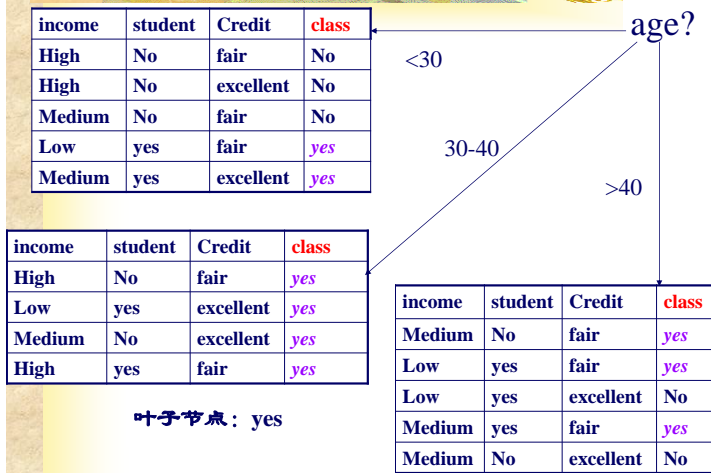
age

age	income	student	Credit-rating	Buys-computer
<30				
<30				
30-40	High	No	fair	yes
>40	Low	yes	excellent	yes
>40	Medium	No	excellent	yes
>40	High	yes	fair	yes

是否购买计算机: yes类别4个样本; no类别0个样本
计算过程
 $I(S_{12}, S_{22})$
 $= I(4, 0)$
 $= -(4/4) \log_2 (4/4) - (0/4) \log_2 (0/4)$
 $= 0$

16

决策树



rid	age	income	student	Credit-rating	Buys-computer
1	<30	High	No	fair	No
2	<30	High	No	excellent	No

是否购买计算机: yes类别9个样本; no类别5个样本
计算过程
 $I(S_1, S_2)$
 $= I(9, 5)$
 $= -(9/14) \log_2 (9/14) - (5/14) \log_2 (5/14)$
 $= 0.94$

11	<30	Medium	yes	excellent	yes
12	30-40	Medium	No	excellent	yes
13	30-40	High	yes	fair	yes
14	>40	Medium	No	excellent	No

age

age	income	student	Credit-rating	Buys-computer
<30				
<30				
30-40	High	No	fair	yes
>40	Low	yes	fair	yes
>40	Low	yes	excellent	No
>40	Medium	yes	fair	yes
>40	Medium	No	excellent	No

是否购买计算机: yes类别3个样本; no类别2个样本
计算过程
 $I(S_{13}, S_{23})$
 $= I(3, 2)$
 $= -(3/5) \log_2 (3/5) - (2/5) \log_2 (2/5)$
 $= 0.971$

17

ID3算法的核心:

在决策树各级节点上选择属性时, 用信息增益作为属性的选择标准, 以使得在每一个非叶节点进行测试时能获得关于被测试记录最大的类别信息。

$$\text{Gain}(A) = I(s_1, s_2, \dots, s_m) - E(A)$$

I: 所有信息, E: 按照某类分析所需的信息熵

12

age	income	student	Credit-rating	Buys-computer
<30				
<30				
30-40	High	No	fair	No
>40	Medium	No	fair	No
>40	Low	yes	fair	yes
>40	Medium	yes	excellent	yes

是否购买计算机: yes类别2个样本; no类别3个样本
计算过程
 $I(S_{11}, S_{21})$
 $= I(2, 3)$
 $= -(2/5) \log_2 (2/5) - (3/5) \log_2 (3/5)$
 $= 0.971$

15

age

age	income	student	Credit-rating	Buys-computer
<30				
<30				
30-40	High	No	fair	No
>40	Low	yes	fair	yes
>40	Medium	yes	excellent	yes

$E(\text{age})$
 $= (5/14)I(S_{11}, S_{21}) + (4/14)I(S_{12}, S_{22}) + (5/14)I(S_{13}, S_{23})$
 $= 0.694$
 $\text{Gain}(\text{age}) = I(s_1, s_2) - E(\text{age})$
 $= I(9, 5) - E(\text{age})$
 $= 0.94 - 0.694$
 $= 0.245$

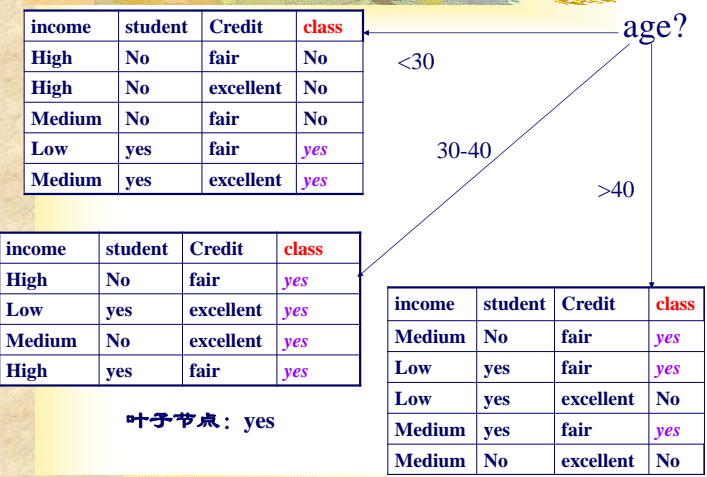
Gain(income)=0.029

Gain(student)=0.151

Gain(credit-rating)=0.048

所以age的信息增益最大, age作为测试属性用于当前分支节点

18



•Apriori算法：使用候选项集找频繁项集

•由频繁项集产生关联规则

设A=足球服, B=足球鞋, C=足球。某网上商城的销售情况如下：

客户号	客户	商品
C1	李鸣	足球服A
C1	李鸣	足球鞋B
C1	李鸣	足球C
C2	金珊	足球C
C3	冯君	足球服A
C3	冯君	足球鞋B
C4	丁贝	足球鞋B
C5	陈骋	足球服A

•支持度
 设W 中有s %的事务同时支持物品集A 和B, s %称为关联规则A→B 的支持度。

关联算法简介

•Apriori算法：使用候选项集找频繁项集

•由频繁项集产生关联规则

客户号	客户	商品
C1	李鸣	足球服A
C1	李鸣	足球鞋B
C1	李鸣	足球C
C2	金珊	足球C
C3	冯君	足球服A
C3	冯君	足球鞋B
C4	丁贝	足球鞋B
C5	陈骋	足球服A

A的支持度？ A→B 的支持度 ？

A的支持度=60% A→B 的支持度 = 40%

关联：哪些商品组合受欢迎

搭配套餐1




限时38元GV... 原价：138.00

限时99元 GV... 原价：168.00

套餐价格：128.00
节省：178.00

查看套餐

设A=足球服, B=足球鞋, C=足球。某网上商城的销售情况如下：

客户号	客户	商品	数量
C1	李鸣	足球服A	10
C1	李鸣	足球鞋B	8
C1	李鸣	足球C	60
C2	金珊	足球C	20
C3	冯君	足球服A	50
C3	冯君	足球鞋B	60
C4	丁贝	足球鞋B	20
C5	陈骋	足球服A	30

•不关心商品的销量，只关心商品间的关联度

Apriori算法：找频繁项集

- 1、在第一轮循环中，所有1项目（只有1项）集是候选项目集，从中筛选出
- 支持度>最小支持度的项目放入频繁项目集
- 2、在第二轮循环中，只有频繁1项目组成的2项目才是候选项目，从中筛选出
- 支持度>最小支持度的2项目放入频繁项目集。
- 3、继续循环，直至n轮循环，所有的频繁项目选出，频繁项目集形成。

客户号	客户	商品
C1	李鸣	足球服A
C1	李鸣	足球鞋B
C1	李鸣	足球C
C2	金珊	足球C
C3	冯君	足球服A
C3	冯君	足球鞋B
C4	丁贝	足球鞋B
C5	陈聘	足球服A

设最小支持度为30%

(1) 1项目候选项: A、B、C

支持度: 60% 60% 40%

所以频繁项目集: 进入下一轮单项:

{A、B、C} {A、B、C}

28

•Apriori算法: 使用候选项集找频繁项集

•由频繁项集产生关联规则

31

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

可信度: 它是针对规则而言的。指在出现了物品集A的事务T中, 物品集B也同时出现的概率有多大。

可信度= $p(\text{condition and result}) / p(\text{condition})$ 。

规则2: if A THEN C, 可信度?

可信度= $(A \text{和} C) / A = 20\% / 60\% = 33\%$

AC的支持度: 20%

34

客户号	客户	商品
C1	李鸣	足球服A
C1	李鸣	足球鞋B
C1	李鸣	足球C
C2	金珊	足球C
C3	冯君	足球服A
C3	冯君	足球鞋B
C4	丁贝	足球鞋B
C5	陈聘	足球服A

设最小支持度为30%

(1) 2项目候选项: AB、BC、AC

支持度: 40% 20% 20%

所以频繁项目集: 进入下一轮单项:

{A、B、C、AB} {A、B}

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

32

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

可信度: 它是针对规则而言的。指在出现了物品集A的事务T中, 物品集B也同时出现的概率有多大。

可信度= $p(\text{condition and result}) / p(\text{condition})$ 。

假设可信度>50%, 规则有意义!

规则1: if A THEN B, 可信度=67%

规则2: if A THEN C, 可信度=33%

规则1有意义

35

客户号	客户	商品
C1	李鸣	足球服A
C1	李鸣	足球鞋B
C1	李鸣	足球C
C2	金珊	足球C
C3	冯君	足球服A
C3	冯君	足球鞋B
C4	丁贝	足球鞋B
C5	陈聘	足球服A

设最小支持度为30%

(1) 3项目候选项: 无

算法终止

所以频繁项目集:

{A、B、C、AB}

30

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

可信度: 它是针对规则而言的。指在出现了物品集A的事务T中, 物品集B也同时出现的概率有多大。

可信度= $p(\text{condition and result}) / p(\text{condition})$ 。

规则1: if A THEN B, 可信度?

可信度= $(A \text{和} B) / A = 40\% / 60\% = 67\%$

33

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

兴趣度: 物品集A的出现对物品集B的出现有多大的影响。

兴趣度= $p(\text{condition and result}) / p(\text{condition}) * p(\text{result})$ 。

——当兴趣度大于1的时候, 这条规则就是比较好的;

——当兴趣度小于1的时候, 这条规则就是没有很大意义的。

规则1: if A THEN B, 兴趣度?

兴趣度= $(A \text{和} B) / (A * B) = 40\% / (60\% * 60\%) = 1.11 > 1$ 有意义

36

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

规则2: if A THEN C, 兴趣度?

$$\text{兴趣度} = (A \text{和} C) / (A * C) \\ = 20\% / (60\% * 40\%) = 0.83 < 1$$

AC的支持度: 20%

37

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

可信度 = $p(\text{condition and result}) / p(\text{condition})$ 。

兴趣度: 物品集A的出现对物品集B的出现有多大的影响。

兴趣度 = $p(\text{condition and result}) / p(\text{condition}) * p(\text{result})$ 。


A、B、C的支持度为20%

规则: if A and B THEN C, 兴趣度?

$$\text{兴趣度} = (A \text{和} B, C) / ((A, B) * C) = 20\% / (40\% * 40\%) = 1.25 > 1$$

有意义

40

	含有蘑菇的披萨数: 100+400+300+100=900 支持度: 900/2000=45%
	含有香肠的披萨数: 150+400+200+100=850 支持度: 850/2000=42.5%
	含有奶酪的披萨数: 200+300+200+100=800 支持度: 800/2000=40%

一家披萨店卖了2000个披萨饼, 其中: 100个仅含蘑菇, 150个是意大利香肠, 200个是含干奶酪; 400个是蘑菇加意大利香肠, 300个是蘑菇加干奶酪, 200个是意大利香肠加干奶酪, 100个是蘑菇、意大利香肠加干奶酪; 550个没有配料。

43

练习

设A=足球服, B=足球鞋, C=足球。


分析结果:

项目	支持度	项目	支持度
A	60%	C	40%
B	60%	A和B	40%
ABC	20%		

规则: if A and B THEN C, 求可信度和兴趣度, 并确定此规则是否有意义(按照兴趣度 $C \geq 1$ 判断)?

38

关联: 一个商品中, 哪些元素的组合受欢迎?

	含有蘑菇+香肠的披萨数: 400+100=500 支持度: 500/2000=25%
	含有蘑菇+奶酪的披萨数: 300+100=400 支持度: 400/2000=20%
	含有香肠+奶酪的披萨数: 200+100=300 支持度: 300/2000=15%
	含有蘑菇+香肠+奶酪的披萨数: 100 支持度: 100/2000=5%

一家披萨店卖了2000个披萨饼, 其中: 100个仅含蘑菇, 150个是意大利香肠, 200个是含干奶酪; 400个是蘑菇加意大利香肠, 300个是蘑菇加干奶酪, 200个是意大利香肠加干奶酪, 100个是蘑菇、意大利香肠加干奶酪; 550个没有配料。

44

设A=足球服, B=足球鞋, C=足球。

分析结果:

频繁项目	支持度	频繁项目	支持度
A	60%	C	40%
B	60%	A和B	40%

可信度 = $p(\text{condition and result}) / p(\text{condition})$ 。

兴趣度: 物品集A的出现对物品集B的出现有多大的影响。

兴趣度 = $p(\text{condition and result}) / p(\text{condition}) * p(\text{result})$ 。

A、B、C的支持度为20%

规则: If A and B then C?

规则: if A and B THEN C, 可信度?

$$\text{可信度} = (A, B \text{和} C) / (A, B) = 20\% / 40\% = 50\%$$

39



一家披萨店卖了2000个披萨饼, 其中: 100个仅含蘑菇, 150个是意大利香肠, 200个是含干奶酪; 400个是蘑菇加意大利香肠, 300个是蘑菇加干奶酪, 200个是意大利香肠加干奶酪, 100个是蘑菇、意大利香肠加干奶酪; 550个没有配料。



支持度: 45%



支持度: 42.5%



支持度: 25%



可信度?

$$= \text{蘑菇+香肠的支持度} / \text{蘑菇的支持度} \\ = 25\% / 45\% = 0.588$$

兴趣度?

$$= \text{蘑菇+香肠的支持度} / (\text{蘑菇的支持度} * \text{香肠的支持度}) \\ = 25\% / (42.5\% * 45\%) = 1.31 > 1$$

45

关联规则的应用举例?

选购热点: 磨砂 帆布 反绒 一脚蹬 厚底 高档皮质 复古风

商务鞋 软底 耐磨

鞋头款: 圆头 尖头 方头 扁头

46

客户响应预测

- 优惠服务——
 - 下一个有可能响应优惠服务的客户, 或许与以前已经响应的客户类似。
- 抱怨服务——
 - 客户抱怨的文本, 归入一系列固定的分类代码

49

2、每个字段只建立一个距离函数

分别为性别、薪金建立一个距离函数

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

52



促销 春季大促 **138.65** 元

配送 至 上海 商家承担运费

月销量 **19550** 件

货号: k01 品牌: 鞋头款式: 圆头

闭合方式: 系带 鞋面材质: 牛筋 鞋面材质: 真皮二层皮

帮面材料: 牛皮 皮革风格: 反绒皮 内里材质: 皮质

47

步骤

- 1、建立训练集
- 2、每个字段只建立一个距离函数
- 3、组合距离函数

50

2、每个字段只建立一个距离函数

常见的距离函数:

- 差的绝对数值: $|A-B|$
- 差的平方: $(A-B)^2$
- 归一化绝对值: $|A-B| / (\text{最大差值})$

不希望年龄或薪金主导组合函数, 用哪个函数?

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

53

最近邻方法—— 距离和相似性的衡量

48

1、建立训练集

营销数据库中客户情况

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

51

2、每个字段只建立一个距离函数

常见的距离函数:

- 差的绝对数值: $|A-B|$
- 差的平方: $(A-B)^2$
- 归一化绝对值: $|A-B| / (\text{最大差值})$

归一化绝对值优点: 所有属性值均在0-1之间

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

54

2、每个字段只建立一个距离函数

归一化绝对值：|A-B| / (最大差值)
年龄归一化：所有属性值均在0-1之间

	27	51	52	33
27				
51	0.96			
52				
33				

$|51-27| / (52-27)$
 $= 24/25$
 $= 0.96$

55

3、建立组合距离函数

为年龄、薪金建立组合距离函数——
欧几里得几何距离：
 $d(A,B)=\text{sqrt}(d_{\text{年龄}}^2 + d_{\text{薪金}}^2)$

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

58

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000

新记录归为哪类？

5	30	\$28000
---	----	---------

$d_{1.5} = \text{sqrt}((3/25)^2 + (9/86)^2) = 0.16$
 $d_{2.5} = \text{sqrt}((21/25)^2 + (36/86)^2) = 0.94$
 $d_{1.5} < d_{2.5}$
所以5和1、4为一类

61

2、每个字段只建立一个距离函数

归一化绝对值：|A-B| / (最大差值)
年龄归一化：所有属性值均在0-1之间

	27	51	52	33
27				
51	0.96			
52		?		
33				

$|52-51| / (52-27)$
 $= 1/25$
 $= 0.04$

56

3、建立组合距离函数

欧几里得几何距离：
 $d(A,B)=\text{sqrt}(d_{\text{年龄}}^2 + d_{\text{薪金}}^2)$

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

求2号与1号的距离
 $d_{\text{年龄}} = |51-27| / (52-27) = 0.96$
 $d_{\text{薪金}} = |64000-19000| / (105000-19000) = 0.52$
 $d = \text{sqrt}(0.96^2 + 0.52^2) = 1.09$

59

小结

- 掌握：
 - CRM的概念、类型（操作、分析、协作）；
 - 数据仓库的定义
 - 数据挖掘的算法介绍
- 重点：
 - 决策树算法、关联算法、最近邻方法

62

2、每个字段只建立一个距离函数

归一化绝对值：|A-B| / (最大差值)
年龄归一化：所有属性值均在0-1之间

	27	51	52	33
27				
51	0.96			
52	1.00	0.04		
33	0.24	0.72	0.76	

57

号	d
1	1, 4, 2, 3

1、4：薪金比较低的年轻人
2、3：薪金比较高的较年长者
相似客户可以采用相同的优惠策略

记录号	年龄	薪金
1	27	\$19000
2	51	\$64000
3	52	\$105000
4	33	\$55000

60



谢谢大家！