北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心
2014年12月8日

12月15日上课时间调整到12月19日

2014 年 12 月				December		
星期一	星期二	星期三	星期四	星期五	星期六	星期日
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15 18:30-20:20 此次课程不上	16	17	18	19 改到周五 18:30-20:20	20	21
22	23	24	25	26	27	28
29	30	31				

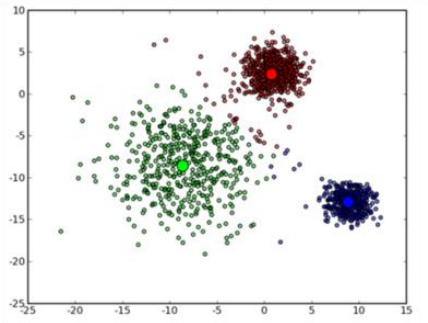
课程回顾：深度业务分析——原方法

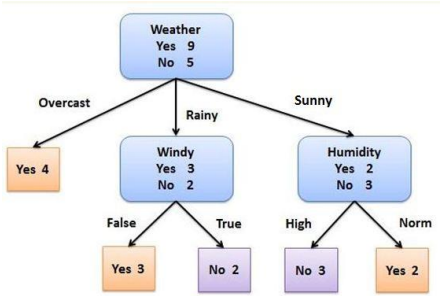
- 聚类（Clustering）
- 分类（Classification）
- 关联（Association）
- 模式（Pattern）
- ……

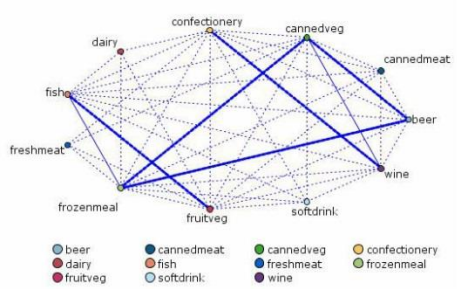
类别

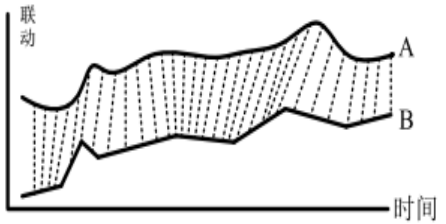
联系

轨迹









本次课程小结

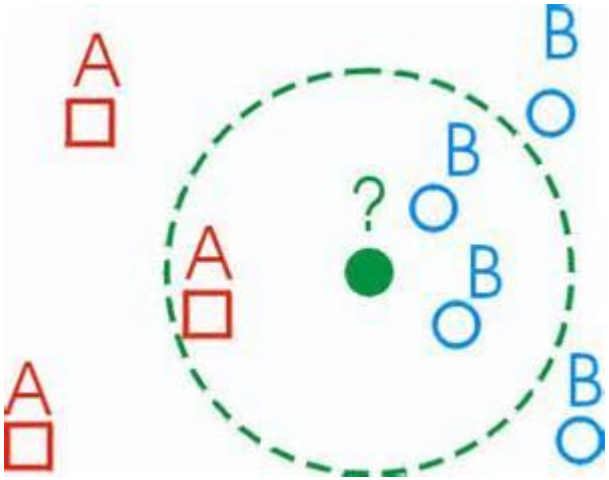
- **分类、预测分析的基本概念**
  - 懒惰型分类 v.s. 急切性分类
  - 分类分析 v.s. 聚类分析
  - 分类分析 v.s. 预测分析
  - 分类方法评估的角度
  - 训练集、测试集、交叉验证、Bootstrap

- **分类分析的典型方法**
  - K近邻方法（K Nearest Neighbors, KNN）
  - 决策树方法（Decision Tree）

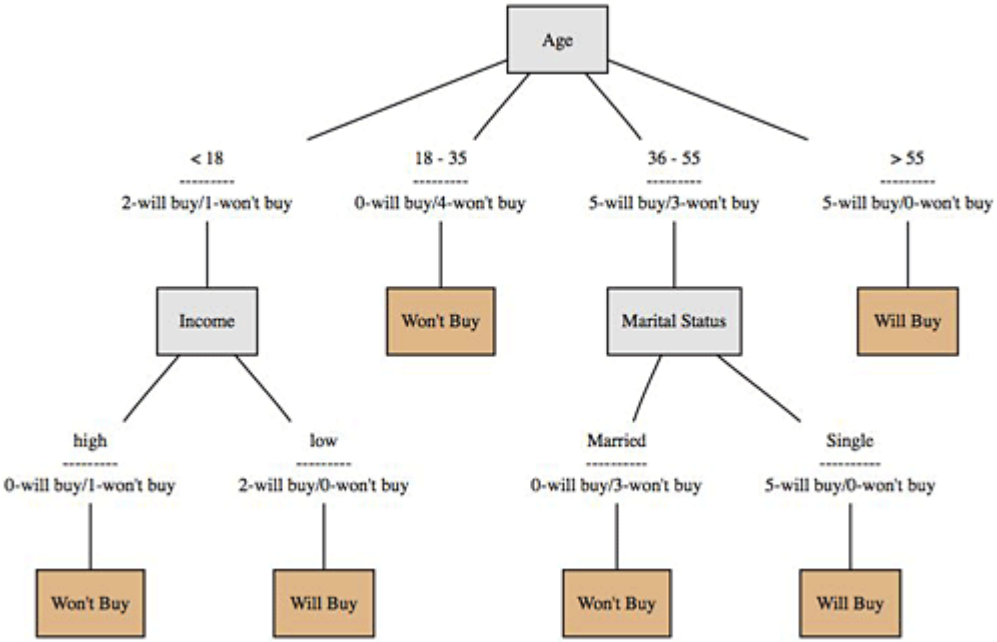
分类、预测分析方法

Classification & Prediction Analysis

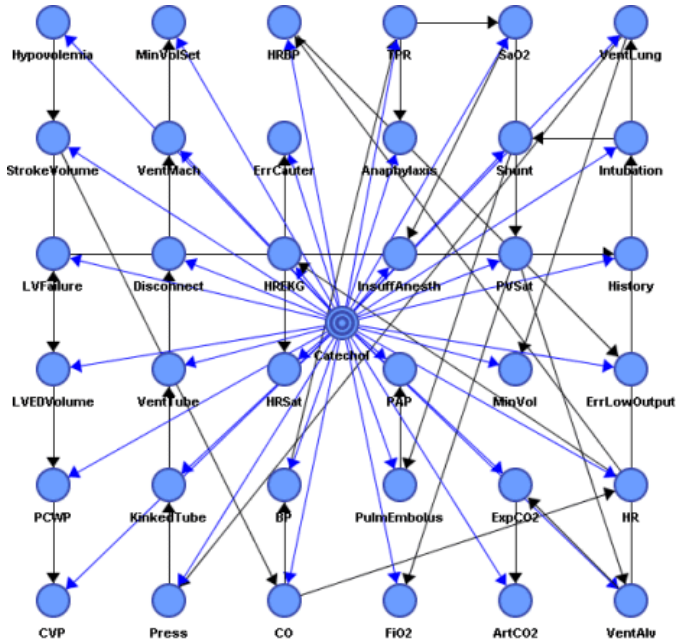
K nearest Neighbors



Decision Tree



Naïve Bayes


Hypovolemia
MinVolSet
HRBP
TPR
SaO2
VentLung
StrokeVolume
VentMach
ErrCauter
Anaphylaxis
Shunt
Intubation
LVFailure
Disconnect
HREKG
InsuffAnesth
PVSat
History
Catechol
LVEDVolume
VentTube
HRSat
PAP
MinVol
ErrLowOutput
PCWP
KinkedTube
BP
PulmEmbolus
ExpCO2
HR
CVP
Press
CO
FiO2
ArtCO2
VentAlv

贝叶斯分类

通常，分类学习模型通过**分类规则、决策树或数学公式**的形式提供

- **理论基础是贝叶斯定理，是一个统计分类器，能够预测一个数据对象属于每个类别的概率，从而判断该数据对象最可能的类别**
- **贝叶斯定理**

$$P(H_1 \mid A) = \frac{P(A \mid H_1) * P(H_1)}{P(A)}$$

- **贝叶斯分类器**
  - f: X → C，finite set of values
  - 样本x∈X有n个属性值：

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

- **给定一个样本，将其划分给类别集合C中隶属概率值最大的一个类别：**

$$C_{MAP} = \mathbf{argmax}_{c_j \in C} \mathbf{P}(c_j \mid \mathbf{x}) = \mathbf{argmax}_{c_j \in C} \mathbf{P}(c_j \mid x_1, x_2, \ldots x_n)$$

$$P(c_j \mid X) = \frac{P(X \mid c_j) P(c_j)}{P(X)}$$

$$C_{MAP} = \mathbf{argmax}_{c_j \in C} \mathbf{P}(x_1, x_2, \ldots x_n \mid c_j) \mathbf{P}(c_j)$$

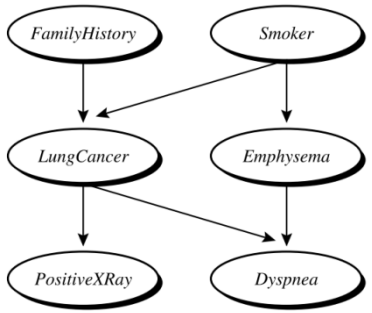
贝叶斯分类

$$C_{MAP} = \text{argmax}_{c_j \in C} P(x_1, x_2 \ldots, x_n \mid c_j) P(c_j)$$

- 给定训练集，估计上面等式右侧两个概率值
- $P(c_j)$：很容易估计
- $P(x_1,x_2,\ldots,x_n|c_j)$：不太好估计
- 朴素贝叶斯分类：独立性假设
  - 样本的属性值与类别标签是相互独立的

如果实际情况不满足独立性假设：

贝叶斯信念网络



	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

$$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

$$C_{NB} = \text{argmax}_{c_j \in V} P(c_j) \prod_i P(x_i \mid c_j)$$

朴素贝叶斯分类：例子

$$C_{NB} = \text{argmax}_{c_j \in V} P(c_j) \prod_i P(x_i \mid c_j)$$

Class-labeled training tuples from the *AllElectronics* customer database.

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

C_1=yes, C_2=no

分类任务：

$$X = (age = youth, income = medium, student = yes, credit\_rating = fair)$$

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

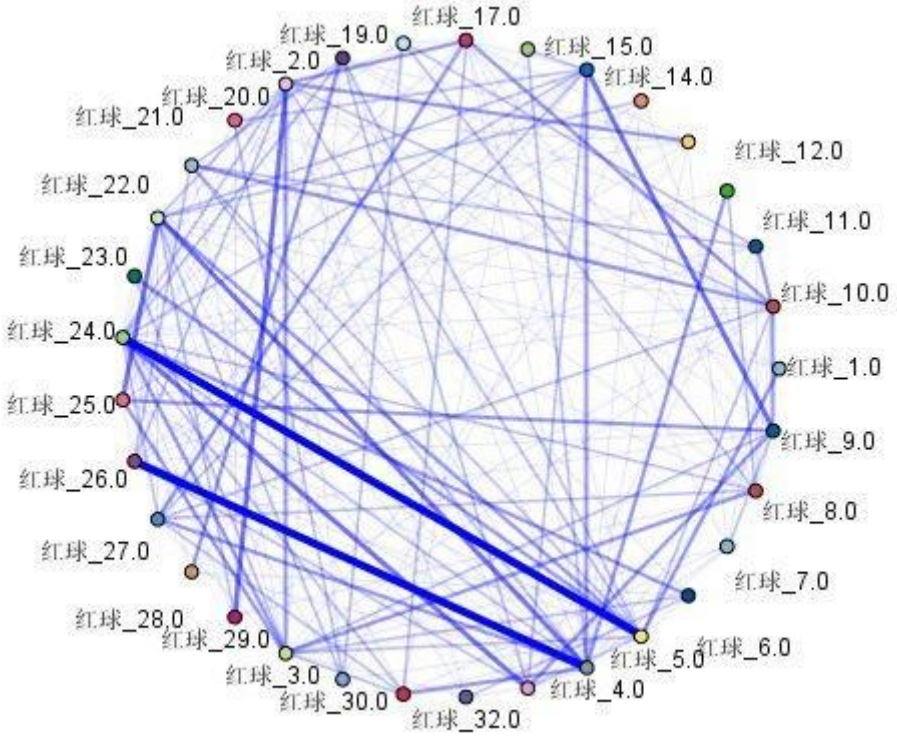
$$P(age = youth \mid buys\_computer = yes) = 2/9 = 0.222$$
$$P(age = youth \mid buys\_computer = no) = 3/5 = 0.600$$
$$P(income = medium \mid buys\_computer = yes) = 4/9 = 0.444$$
$$P(income = medium \mid buys\_computer = no) = 2/5 = 0.400$$
$$P(student = yes \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(student = yes \mid buys\_computer = no) = 1/5 = 0.200$$
$$P(credit\_rating = fair \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(credit\_rating = fair \mid buys\_computer = no) = 2/5 = 0.400$$

$$P(X \mid buys\_computer = yes) = P(age = youth \mid buys\_computer = yes) \times P(income = medium \mid buys\_computer = yes) \times P(student = yes \mid buys\_computer = yes) \times P(credit\_rating = fair \mid buys\_computer = yes)$$
$$= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.$$

$$P(X \mid buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

$$P(X \mid buys\_computer = yes) P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(X \mid buys\_computer = no) P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

Associative Classification



关联分类

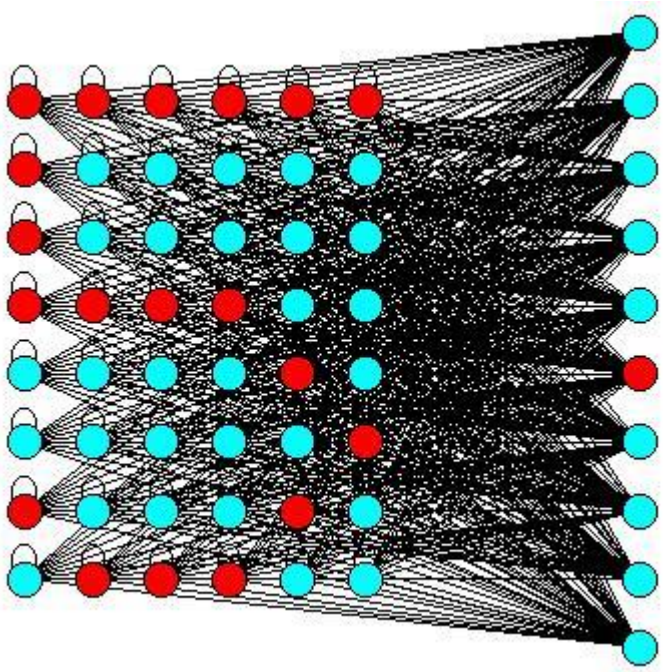
关联规则：X => Y

- **将关联规则挖掘生成的关联规则作为分类规则，进行分类**
- **当关联规则的后项为类别标签时，该规则可以作为分类关联规则（Class Association Rules - CAR）：X => C**
- **两种思路**
  - 1. 首先挖掘出所有的关联规则（满足阈值限制），然后从中挑选出后项为类别属性的规则
  - 2. 直接挖掘后项为单个类别属性的分类关联规则
- **分类关联规则的优先度（Priority）**
  - 1. 置信度大的规则优先度大；
  - 2. 置信度相同的，支持度大的规则优先度大；
  - 3. 置信度和支持度均相同的，先生成的规则优先度大。

常见的关联分类方法

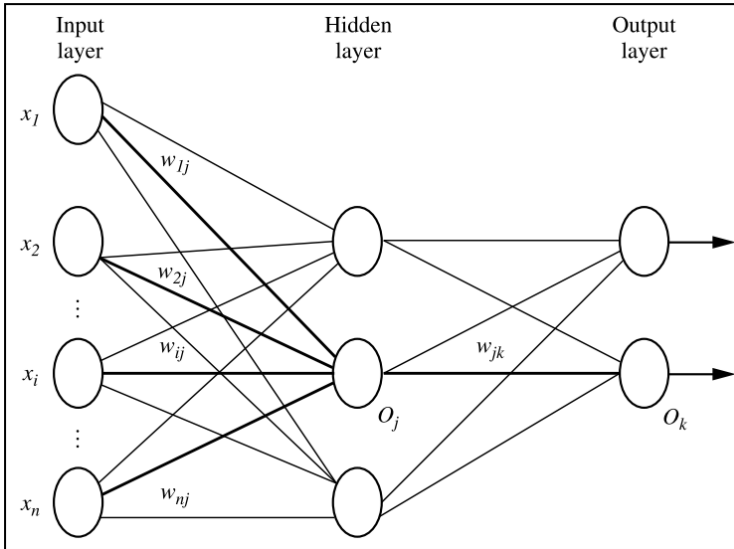
- **CBA**
  - Classification By Association: Bing Liu, Hsu & Ma @ KDD’98
- **CMAR**
  - Classification based on Multiple Association Rules: Li, Han, Pei @ ICDM’01)
- **CPAR**
  - Classification based on Predictive Association Rules: Xiaoxin Yin & Jiawei Han @ SDM’03)
- **GARC**
  - Gain-based Association Rule Classification: Guoqing Chen et al. @ DSS 2006

Neural Network

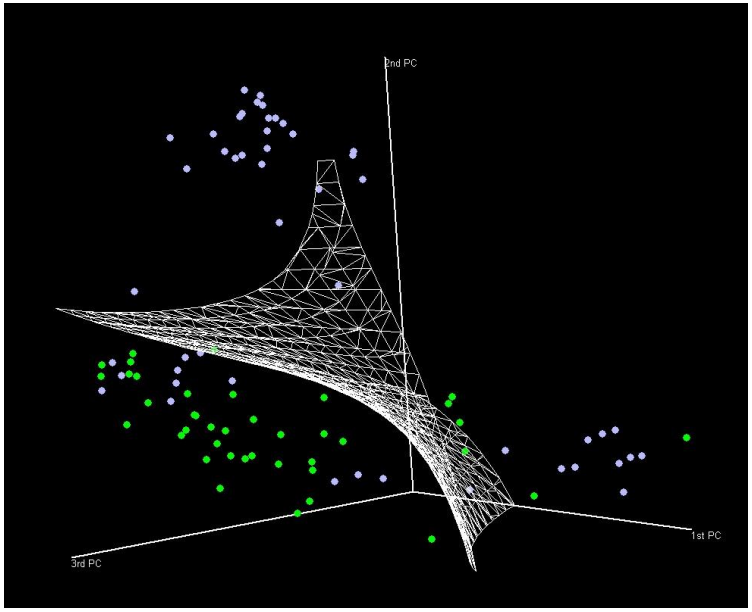


神经元网络学习算法

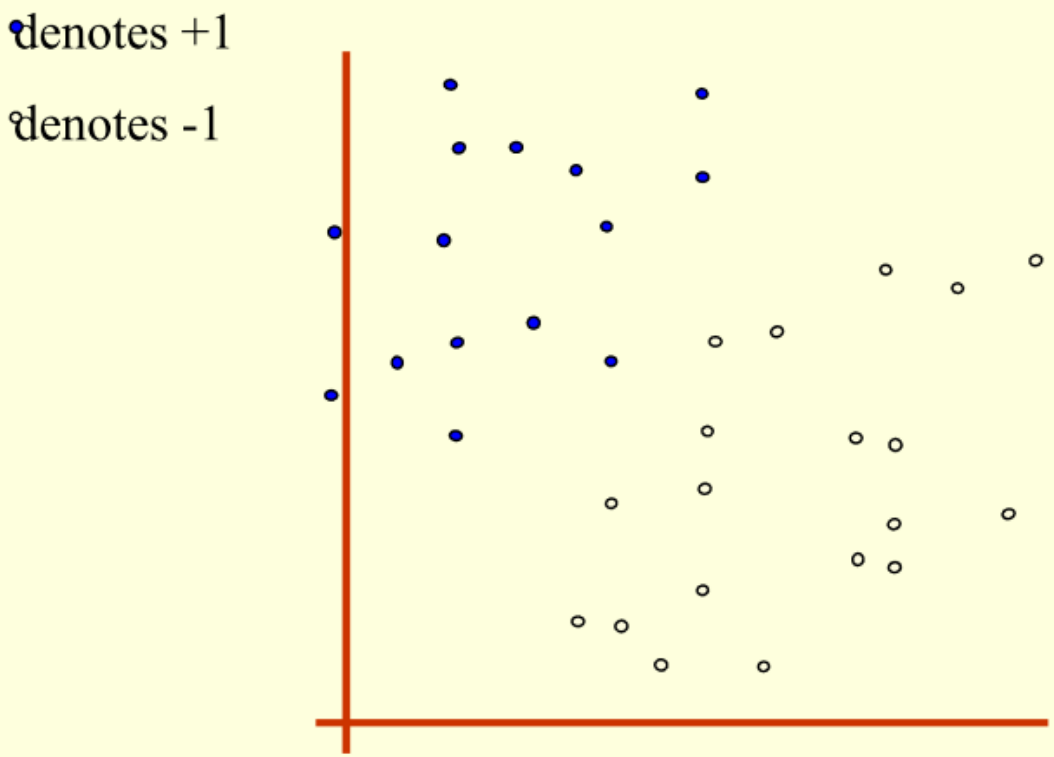
- **神经元网络起源于生理学和神经生物学中有关神经细胞计算本质的研究工作，是在对人脑组织结构和运行机制的认识理解基础之上模拟其结构和智能行为的一种工程系统**
- **神经元网络是一组连接的输入、输出单元，其中每个连接都与一个权相关联；在学习阶段，通过调整神经元网络的权，使得能够预测输入样本的正确类标签来学习**
- **劣势**
  - 神经元网络需要的训练时间较长，需要大量依靠经验确定的参数（如权重、网络拓扑结构），像黑盒子人们很难理解模型的学习过程，可解释性较差
- **优势**
  - 对噪音数据的高承受能力，对未经训练数据的分类能力很好
- **深度学习（Deep Learning）**



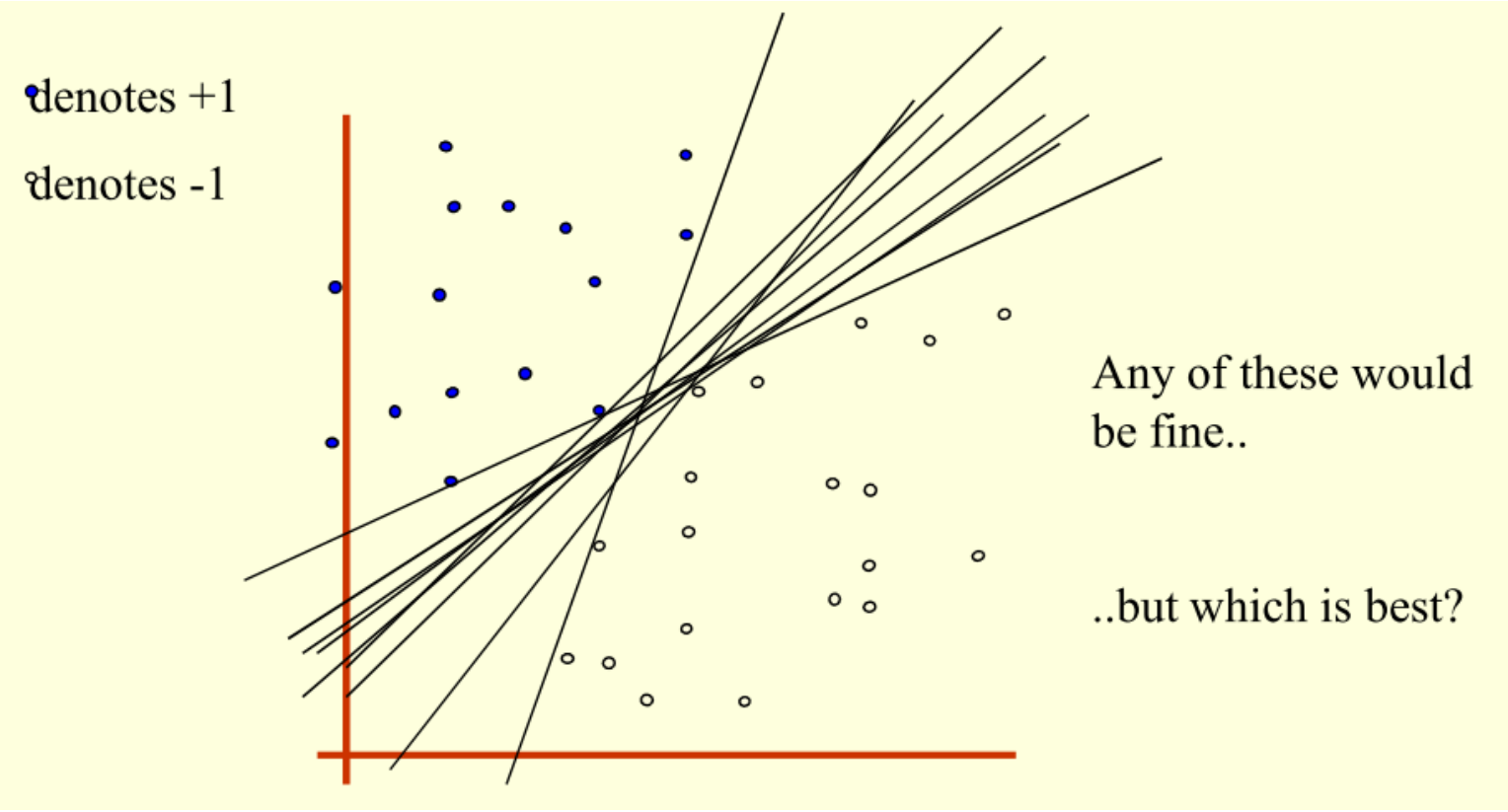
Support Vector Machines



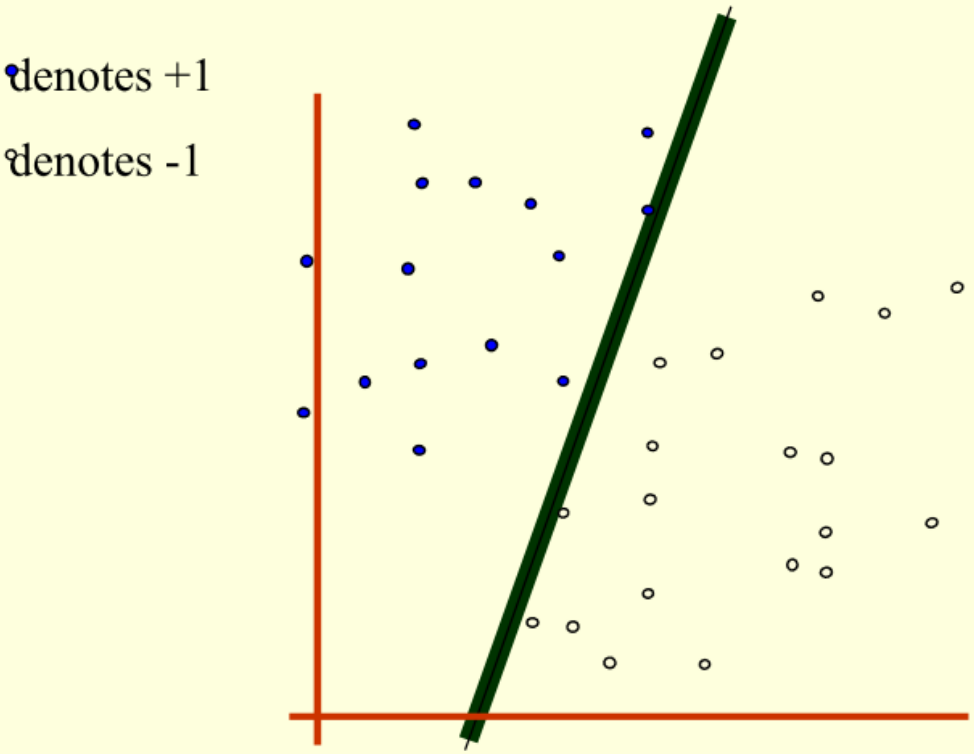
线性分类器



线性分类器

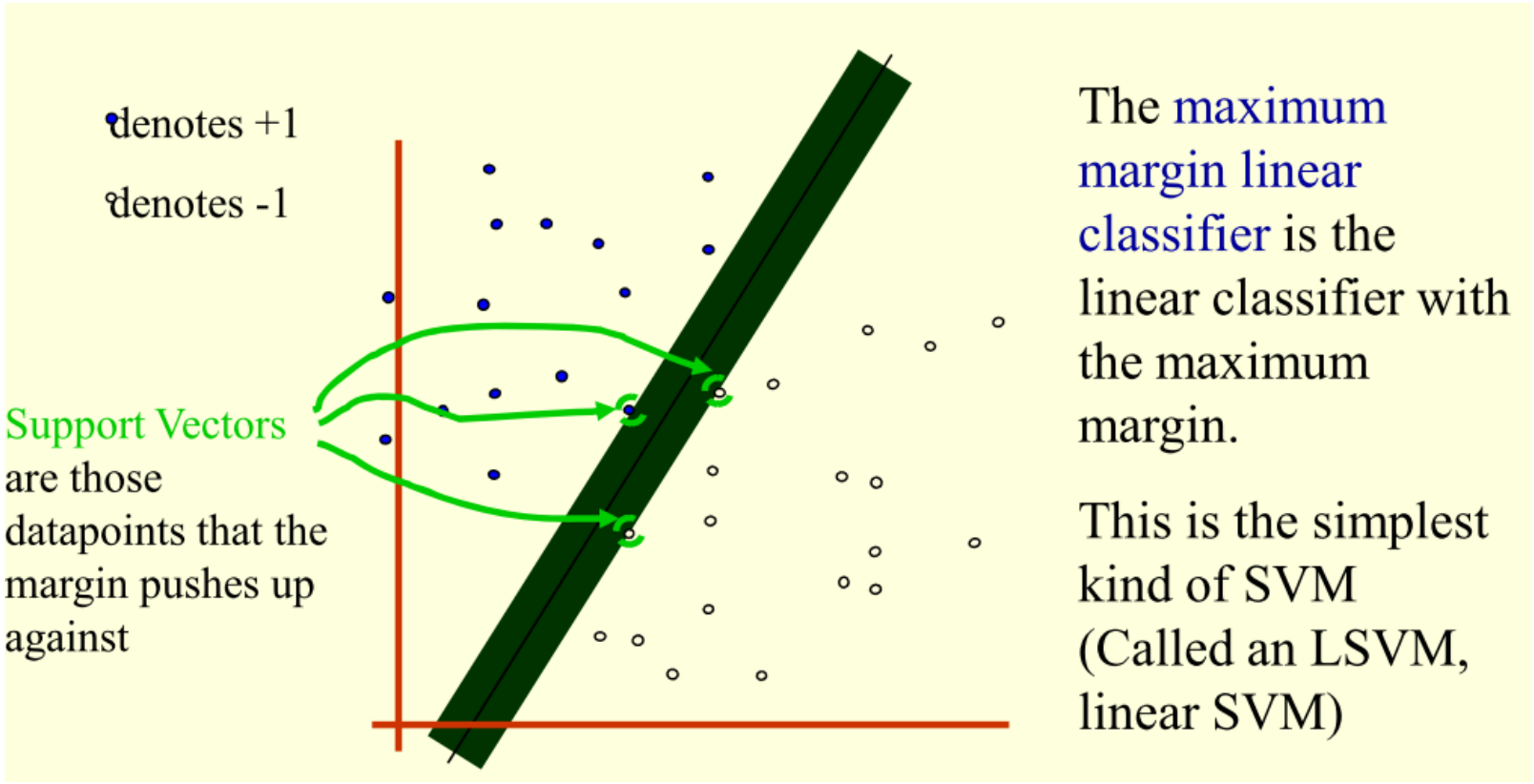

denotes +1
denotes -1
Any of these would be fine..
..but which is best?

线性分类器

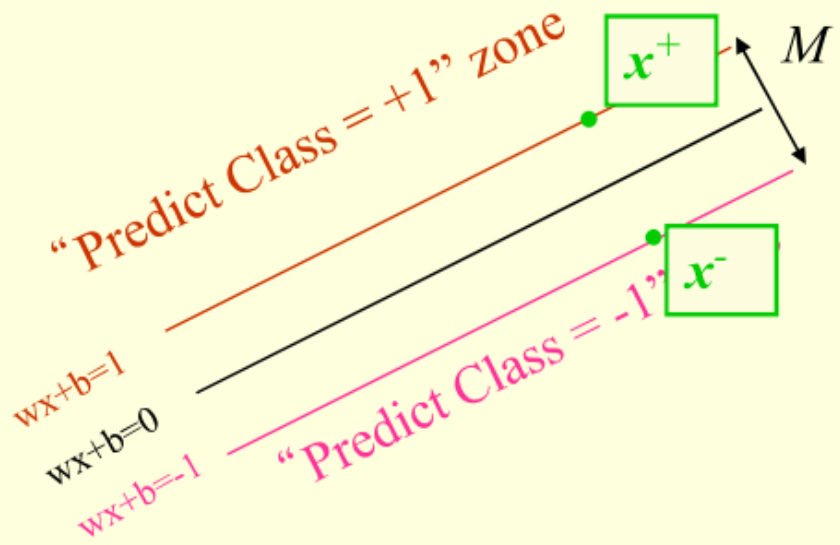


Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

线性分类器


denotes +1
denotes -1
Support Vectors are those datapoints that the margin pushes up against
The maximum margin linear classifier is the linear classifier with the maximum margin.
This is the simplest kind of SVM (Called an LSVM, linear SVM)

线性分类器：训练最大边界分类器



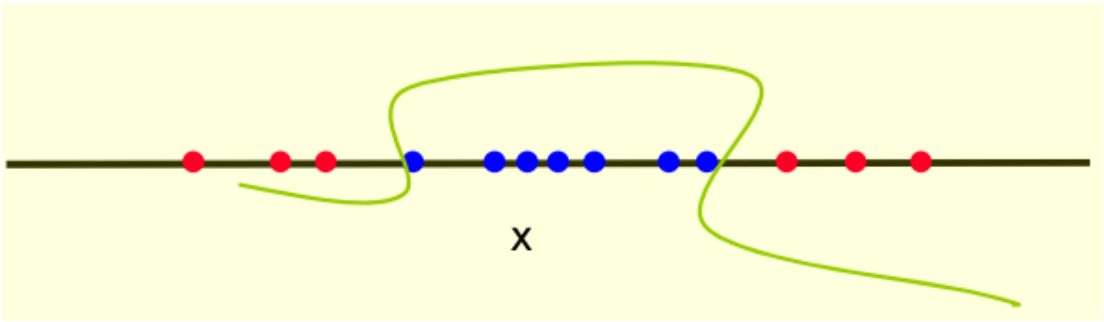
Learn *w* and *b* *s.t.*

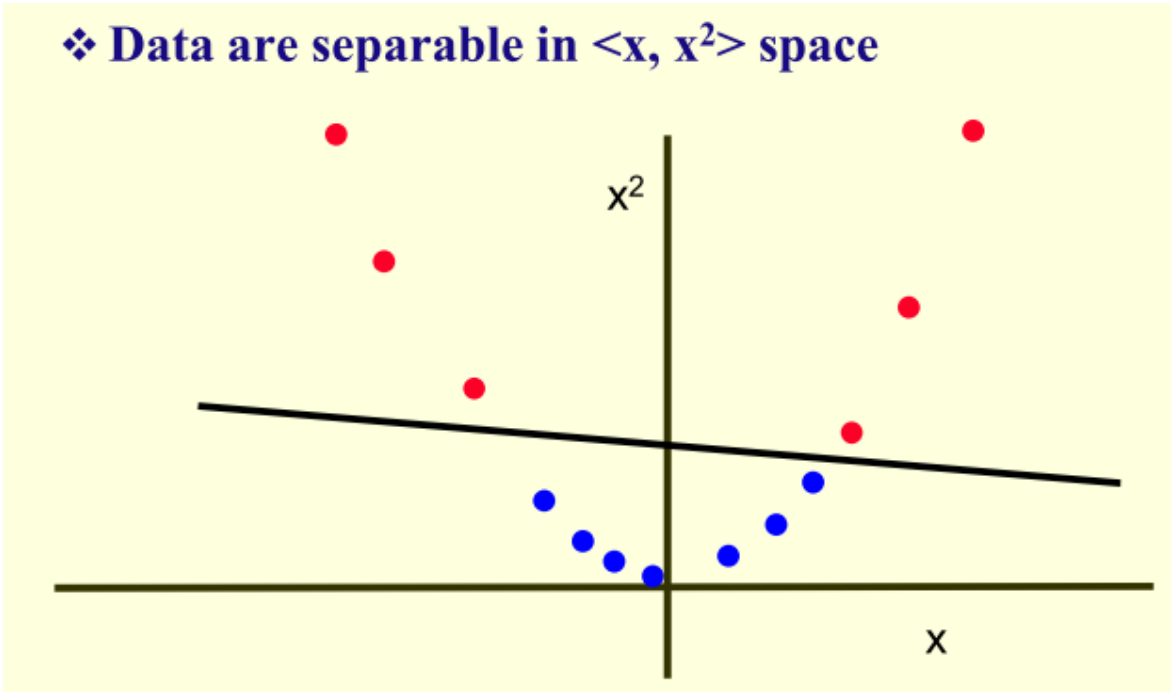
- **satisfying min | w . x + b | = 1**
- **maximizing the margin width =** $\frac{2}{\|\mathbf{w}\|}$

支持向量机分类机 (Support Vector Machine, SVM)

- 支持向量机能够非常成功地处理回归问题(时间序列分析)和模式识别(分类问题、判别分析)等诸多问题，并可推广于预测和综合评价等领域
- 分类问题中，需要分类的对象可以看作是n维实空间中的点。为了确定每个点的类别，我们希望能够把这些点通过一个**n-1维的超平面**分开，通常这个超平面被称为线性分类器
- 我们希望找到分类最佳的超平面，也就是使得属于两个不同类的数据点间隔最大的那个面，该面亦称为**最大间隔超平面**。支持向量机的目的就是寻找最大间隔超平面
- 支持向量机通过引入**核函数**将数据样本对应的向量映射到一个**更高维**的空间里，在这个空间里找到一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面，支持向量机通过建立**方向合适**的间隔超平面使两个与之平行的超平面间的距离最大化。
- 支持向量机的基本假设为，平行超平面间的距离或差距越大，分类器的总误差越小。

特征空间：向高维空间转换





分类、预测分析（Classification & Prediction）

- 基本概念
- 分类分析的经典方法
- **预测分析的常用方法**
- 分类、预测方法的评估
- 分类方法的应用案例
- 总结

什么是预测分析？

- **预测与分类的相似之处**
  - 首先，构建一个模型
  - 其次，使用模型预测未知的值

- **预测最主要使用的方法**
  - 线性和多元回归
  - 非线性回归

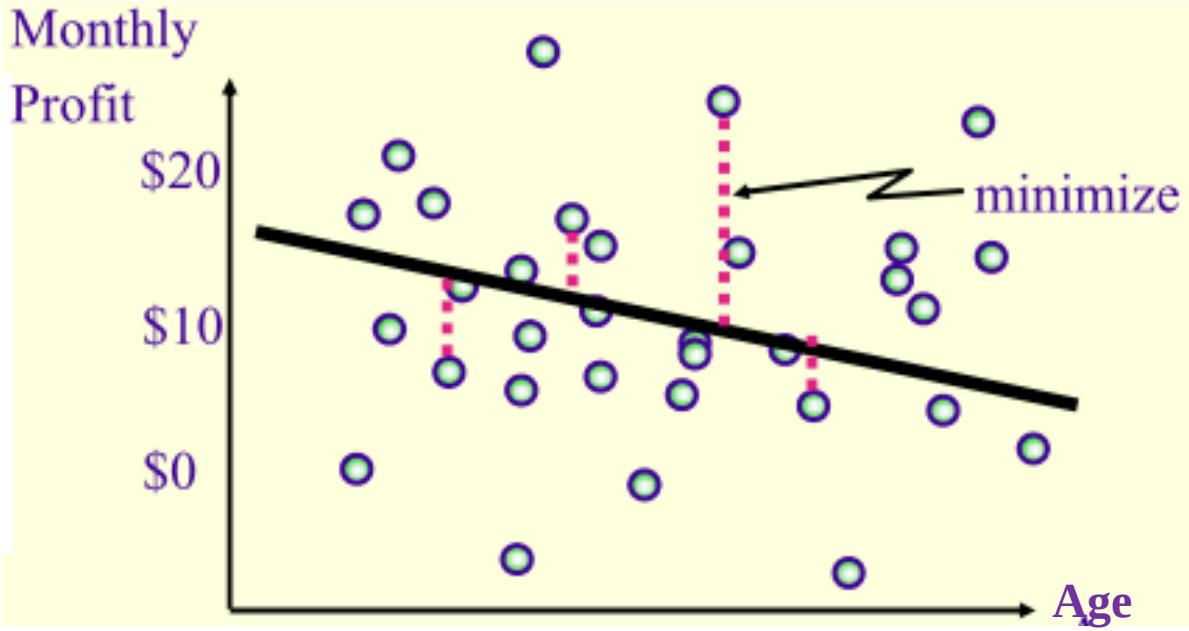
- **预测与分类的区别**
  - 分类分析主要处理类别标签的判定和预测
  - 预测分析模型一般处理连续数值的预测问题

回归分析

- **一元线性回归：$y = w_0 + w_1 x$**
  - 使用样本数据来估计能够确定直线的两个参数w_0和w_1
  - 使用最小平方法求解，使得实际数据与该直线的估计之间误差最小

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

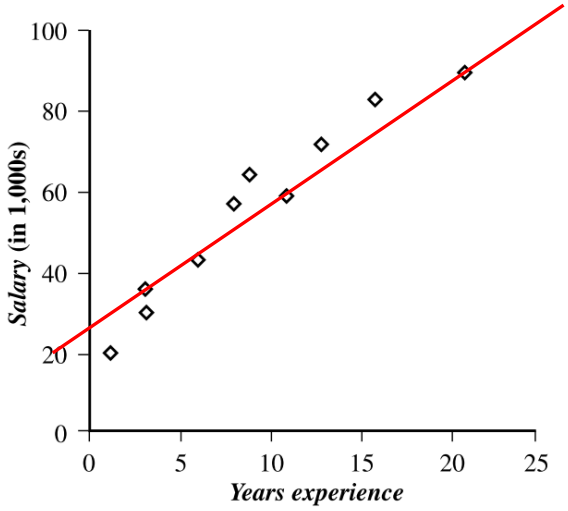


一元线性回归：例子

Salary data.

x years experience	y salary (in $1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

$$y = 23.6 + 3.5x$$



$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5$$

$$w_0 = 55.4-(3.5)(9.1) = 23.6$$

回归分析

- **多元线性回归：$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$**
  - 很多非线性模型可以转换成上述线性模型：

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3.$$

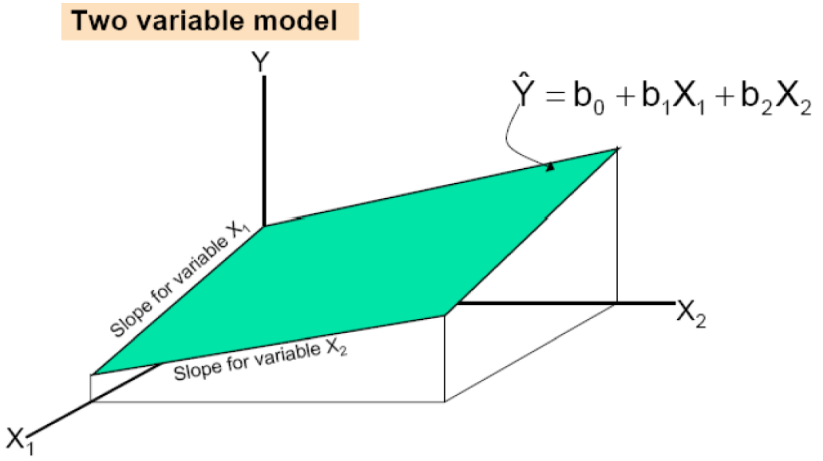
  - 解决多元线性回归问题通常使用一些统计软件，例如：SAS, SPSS, R等

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{Y} = (Y_1, \cdots, Y_n)^T_{n\times 1}$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{k1} \\ 1 & X_{12} & \ldots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \ldots & X_{kn} \end{pmatrix}_{n\times p}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$



回归分析

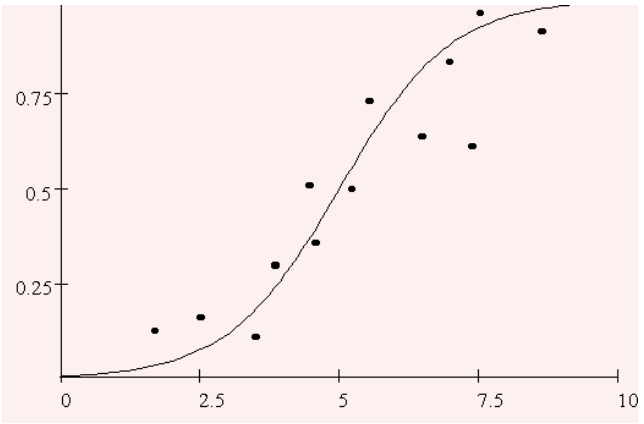
- **Logistic回归**
  - 用来预测某个事件发生的概率，该事件的发生基于一个或多个独立的输入变量
  - 一般用来处理因变量是二项分布的回归问题，即因变量是二分类

- **Logistic回归模型：**
  - log-likelihood maximum (对数似然最大化)

$$\pi(x) = P(Y = 1 \mid X_1, X_2, \cdots, X_n)$$

$$\ln \frac{\pi(x)}{1-\pi(x)} = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

$$\pi(x) = \frac{e^{(\mu+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}{1+e^{(\mu+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$



Logistic回归：例子

- Y：life insurance promotion
- Credit: credit card insurance
- P: probability of choosing life insurance

#	Income	Credit	sex	age	Y	P
1	40K	0	1	45	0	0.007
2	30K	0	0	40	1	0.987
3	40K	0	1	42	0	0.024
4	30K	1	1	43	1	1
5	50K	0	0	38	1	0.999
6	20K	0	0	55	0	0.049
7	30K	1	1	35	1	1
8	20K	0	1	27	0	0.584
9	30K	0	1	43	0	0.005
10	30K	0	0	41	1	0.981
11	40K	0	0	43	1	0.985
12	20K	0	1	29	1	0.38
13	50K	0	0	39	1	0.999
14	40K	0	1	55	0	0
15	20K	1		19	1	1

$$P(Y=1 \mid X)=\frac{e^{(17.691+0.0001\text{Income}+19.871\text{CreditCardIns}-8.314\text{Sex}-0.415\text{Age})}}{1+e^{(17.691+0.0001\text{Income}+19.871\text{CreditCardIns}-8.314\text{Sex}-0.415\text{Age})}}$$

分类、预测分析（Classification & Prediction）

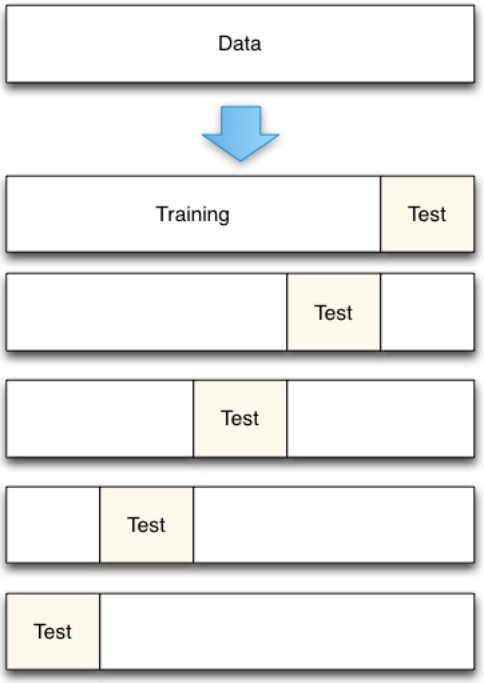
- 基本概念
- 分类分析的经典方法
- 预测分析的常用方法
- **分类、预测方法的评估**
- 应用案例
- 总结

分类方法的结果评估

- **分类的准确性（Accuracy）**
  - 最基础、最重要，模型正确预测新的或先前未见过的数据的类标签的能力
- **分类的速度（Speed）**
  - 计算成本
- **分类器的鲁棒性（Robustness）**
  - 给定噪音数据或有遗漏值的数据，模型正确预测的能力
- **分类器的可扩展性（Scalability）**
  - 给定大规模数据，有效构建分类模型的能力
- **分类器的可解释性（Interpretability）**
  - 分类器表示的知识被用户理解的程度

分类实践中需要关注的问题

- 训练集 和 测试集的选择
  - Training set (2/3), Test set (1/3)
  - 随机选择
  - Problem 1: What if all examples with a certain class were missed out of the training set?
- 交叉验证（Cross-validation）
  - Every example is used in training and testing in turn
  - Folds: partition of the data
  - 常用：ten-fold cross-validation
  - Leave-one-out（留一法）
  - Problem 2: What if the number of all data examples was too small?



分类实践中需要关注的问题（续）

- 自引导、重采样（Bootstrap）
  - Sampling with replacement（有放回的随机采样）
  - Dataset with *n* instances is sampled *n* times → training set
  - Instances which are not picked → test set
  - Every time, Probability of being picked: $\frac{1}{n}$
  - Probability of not being picked: $1-\frac{1}{n}$
  - Probability of an instance not picked by *n* times：

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

分类准确率的评估测度

- **混淆矩阵：Confusion Matrix**

		Predicted class	
		C_1	C_2
Actual class	C_1	true positives	false negatives
	C_2	false positives	true negatives

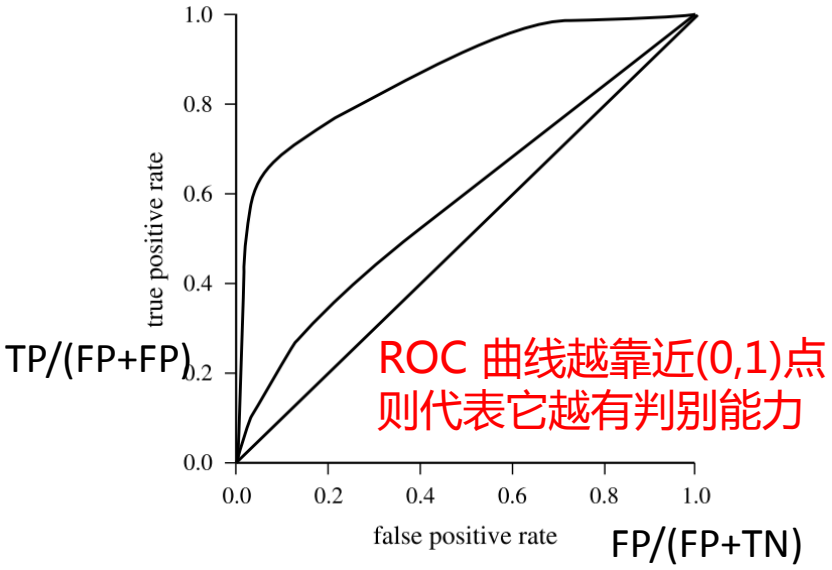
$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$error\ rate = 1 - accuracy = \frac{FP + FN}{TP + FP + FN + TN}$$

$$recall(sensitivity) = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP} \qquad specificity = \frac{TN}{FP + TN}$$

ROC curves
(Receiver Operating Characteristic)



预测方法的结果评估

- **与分类方法的相似之处**
  - 需要使用训练集、测试集
  - 交叉验证
- **与分类方法的不同之处**
  - 准确度、错误率等质量评估指标不再适用
  - 平均平方误差(mean squared error)

$$\text{Mean squared error}: \frac{\sum_{i=1}^{d}(y_i - y_i')^2}{d}$$

  - 平均绝对误差(mean absolute error)

$$\text{Mean absolute error}: \frac{\sum_{i=1}^{d}|y_i - y_i'|}{d}$$

  - 相对平方误差(relative squared error)

$$\text{Relative squared error}: \frac{\sum_{i=1}^{d}(y_i - y_i')^2}{\sum_{i=1}^{d}(y_i - \bar{y})^2}$$

  - 相对绝对误差(relative absolute error)

$$\text{Relative absolute error}: \frac{\sum_{i=1}^{d}|y_i - y_i'|}{\sum_{i=1}^{d}|y_i - \bar{y}|}$$

分类、预测分析（Classification & Prediction）

- 基本概念
- 分类分析的经典方法
- 预测分析的常用方法
- 分类、预测方法的评估
- 应用案例
- 总结

利用数据挖掘预防客户流失的案例

预防客户流失是解除管制后的电信业最关键的挑战之一

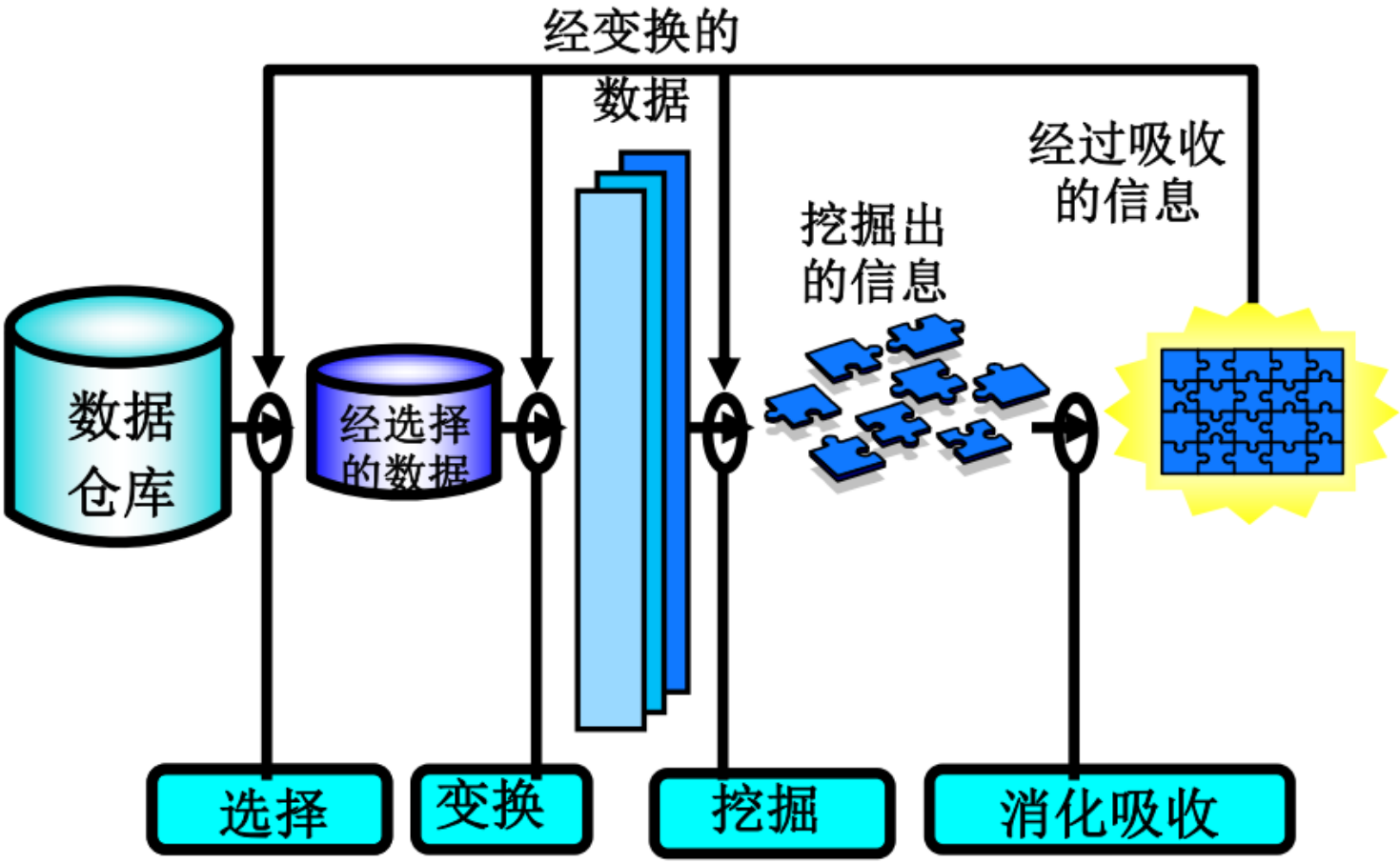
现 状

- 发展新客户的成本比较高 (例如:对新客户的折扣，市场促销活动等)
- 新客户在相当时间之后才会带来盈利
- 防止客户流失比发展新客户成本低
- 客户越来越不忠诚 (customers are less and less loyal)

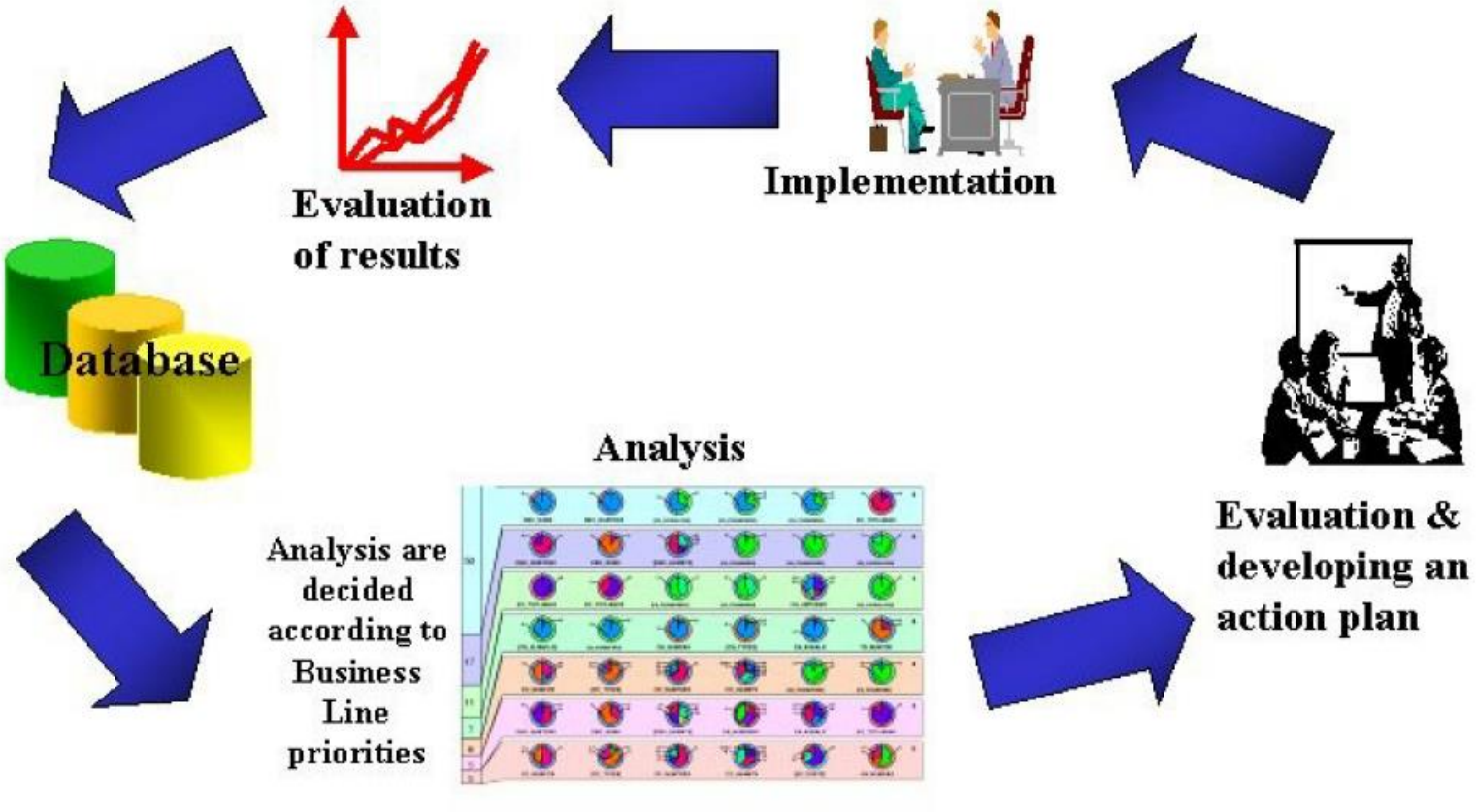
挑 战

- 发展新客户
  - 发展“正确的” （忠诚的并有利可图的)客户
- 预防流失
  - 防止有价值的客户流失

数据挖掘是预防客户流失的关键...



…并且被整合到电信公司市场营销过程之中…



预测模型的挖掘过程

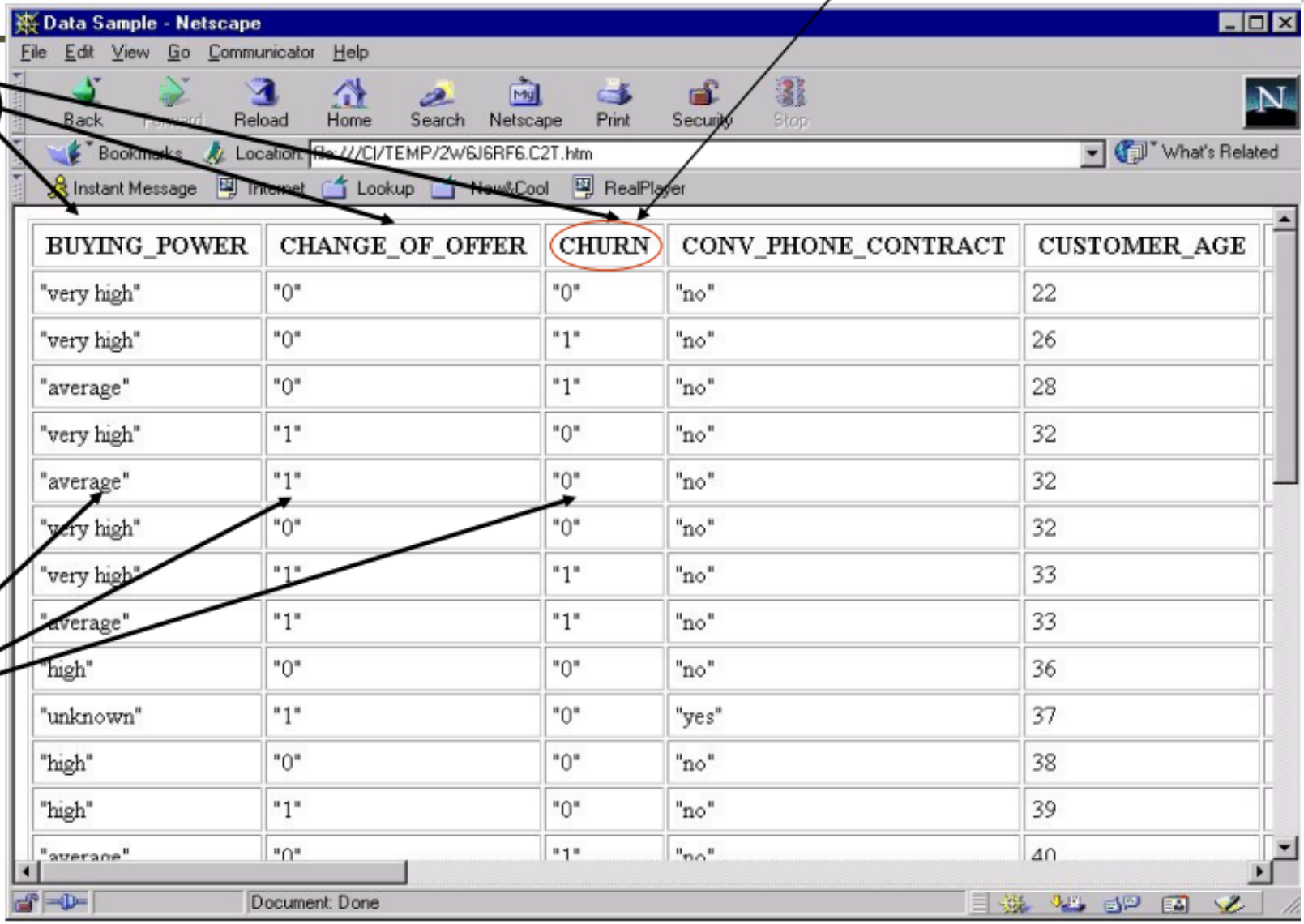
- **抽取和准备数据**
  - 浏览并分析数据
  - 将数据拆分为：训练集、测试集
- **建立模型**
  - 选择并运行算法：
    - 决策树、RBF预测(Radial Basis Function)、线性回归
  - 在测试数据集上检查并验证结果
  - 如果需要重复这个过程
- **选择模型**
  - 用ROC曲线(或增益图)比较各个模型
  - 选择最佳模型
- **根据结果采取行动**

浏览数据

Class label attr.

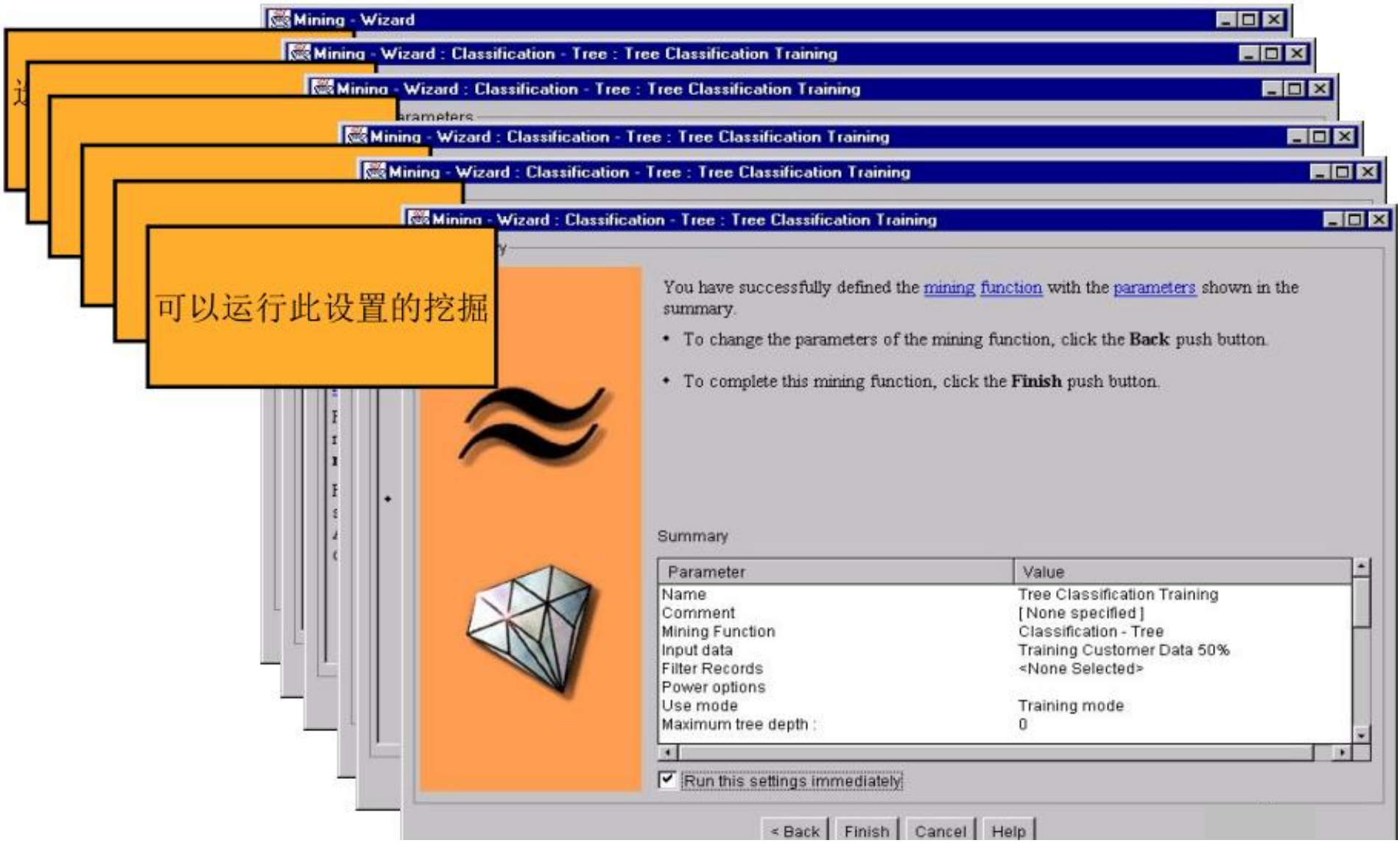
属性

值

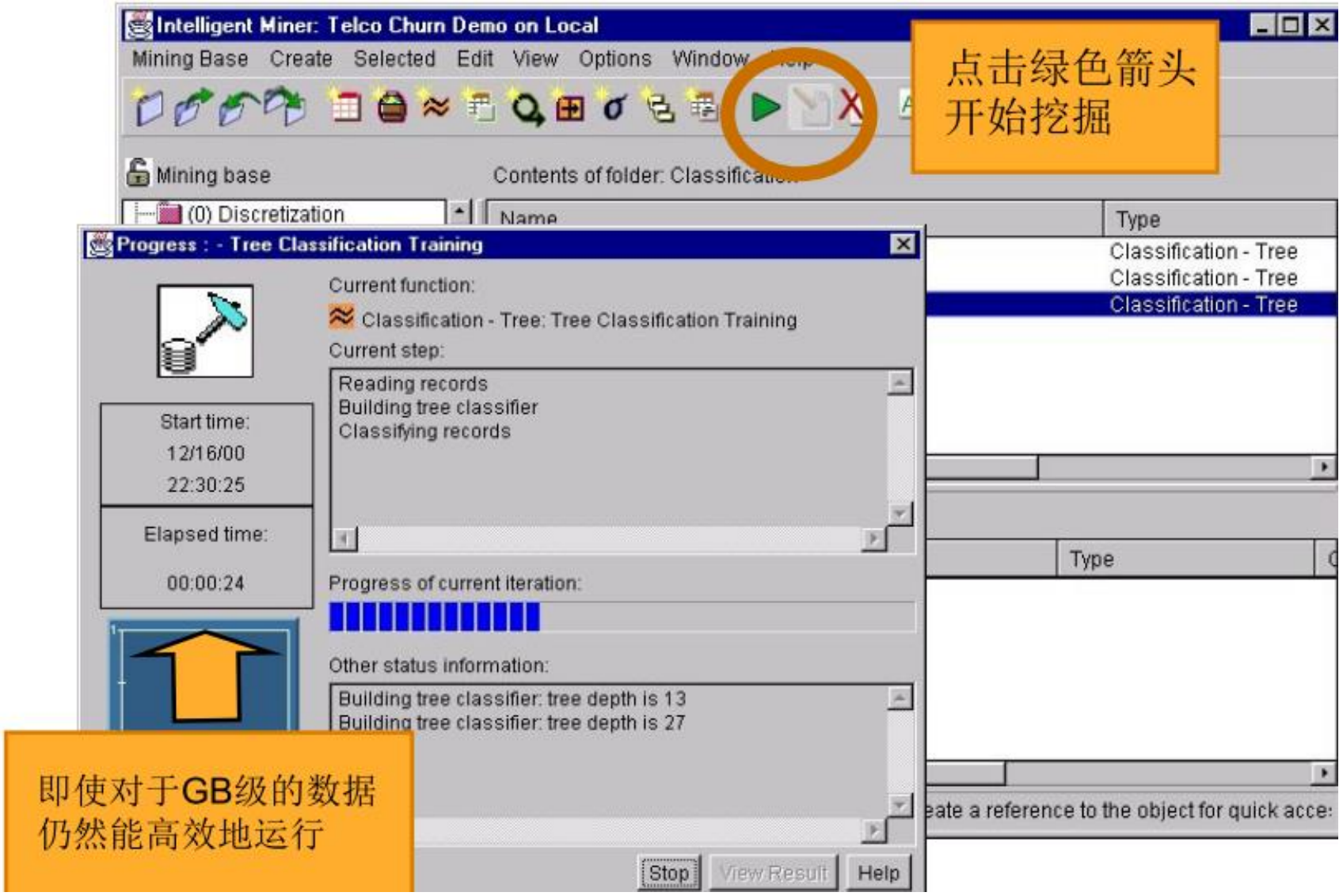


BUYING_POWER	CHANGE_OF_OFFER	CHURN	CONV_PHONE_CONTRACT	CUSTOMER_AGE
"very high"	"0"	"0"	"no"	22
"very high"	"0"	"1"	"no"	26
"average"	"0"	"1"	"no"	28
"very high"	"1"	"0"	"no"	32
"average"	"1"	"0"	"no"	32
"very high"	"0"	"0"	"no"	32
"very high"	"1"	"1"	"no"	33
"average"	"1"	"1"	"no"	33
"high"	"0"	"0"	"no"	36
"unknown"	"1"	"0"	"yes"	37
"high"	"0"	"0"	"no"	38
"high"	"1"	"0"	"no"	39
"average"	"0"	"1"	"no"	40

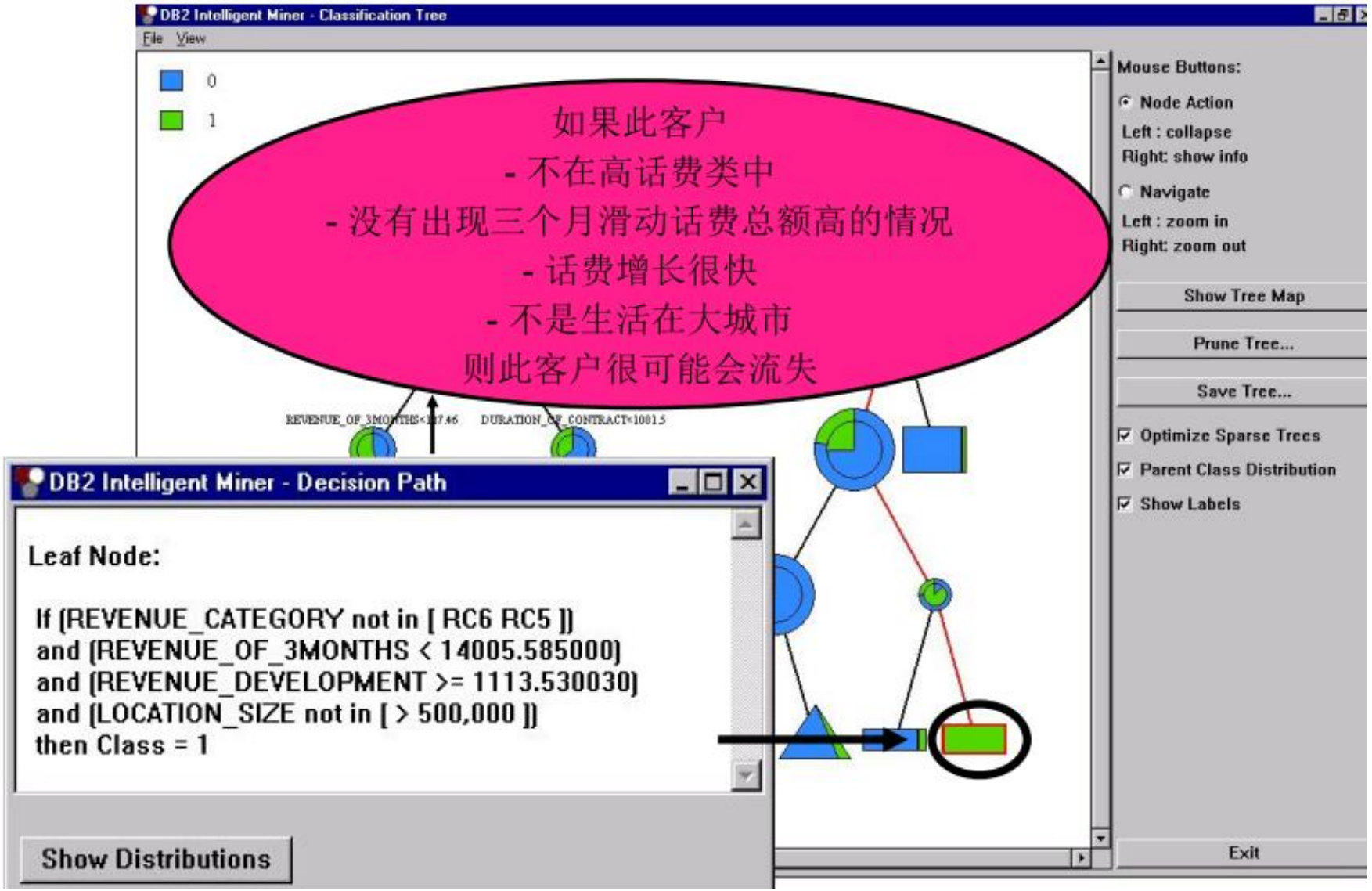
定义分类设置对象


Mining - Wizard
Mining - Wizard : Classification - Tree : Tree Classification Training
Mining - Wizard : Classification - Tree : Tree Classification Training
Mining - Wizard : Classification - Tree : Tree Classification Training
Mining - Wizard : Classification - Tree : Tree Classification Training
Mining - Wizard : Classification - Tree : Tree Classification Training
可以运行此设置的挖掘
You have successfully defined the mining function with the parameters shown in the summary.
To change the parameters of the mining function, click the Back push button.
To complete this mining function, click the Finish push button.
Summary
Parameter
Value
Name
Tree Classification Training
Comment
[None specified]
Mining Function
Classification - Tree
Input data
Training Customer Data 50%
Filter Records
<None Selected>
Power options
Use mode
Training mode
Maximum tree depth :
0
Run this settings immediately
< Back
Finish
Cancel
Help

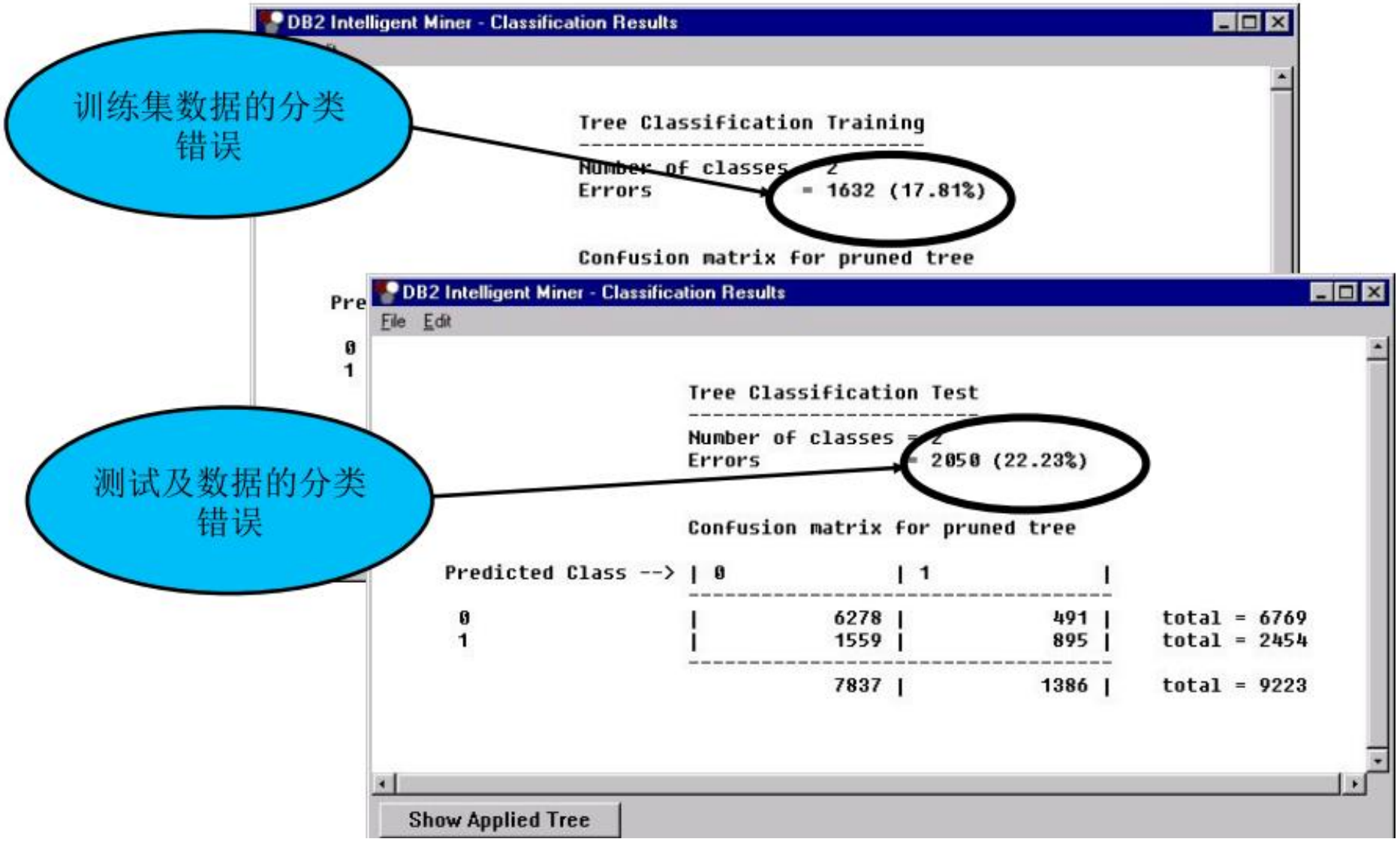
运行分类训练设置


Intelligent Miner: Telco Churn Demo on Local
Mining Base Create Selected Edit View Options Window
点击绿色箭头
开始挖掘
Mining base
Contents of folder: Classification
(0) Discretization
Name
Type
Classification - Tree
Classification - Tree
Classification - Tree
Progress : - Tree Classification Training
Current function:
Classification - Tree: Tree Classification Training
Current step:
Reading records
Building tree classifier
Classifying records
Start time:
12/16/00
22:30:25
Elapsed time:
00:00:24
Progress of current iteration:
Other status information:
Building tree classifier: tree depth is 13
Building tree classifier: tree depth is 27
Stop
View Result
Help
Type
eate a reference to the object for quick acce
即使对于GB级的数据
仍然能高效地运行

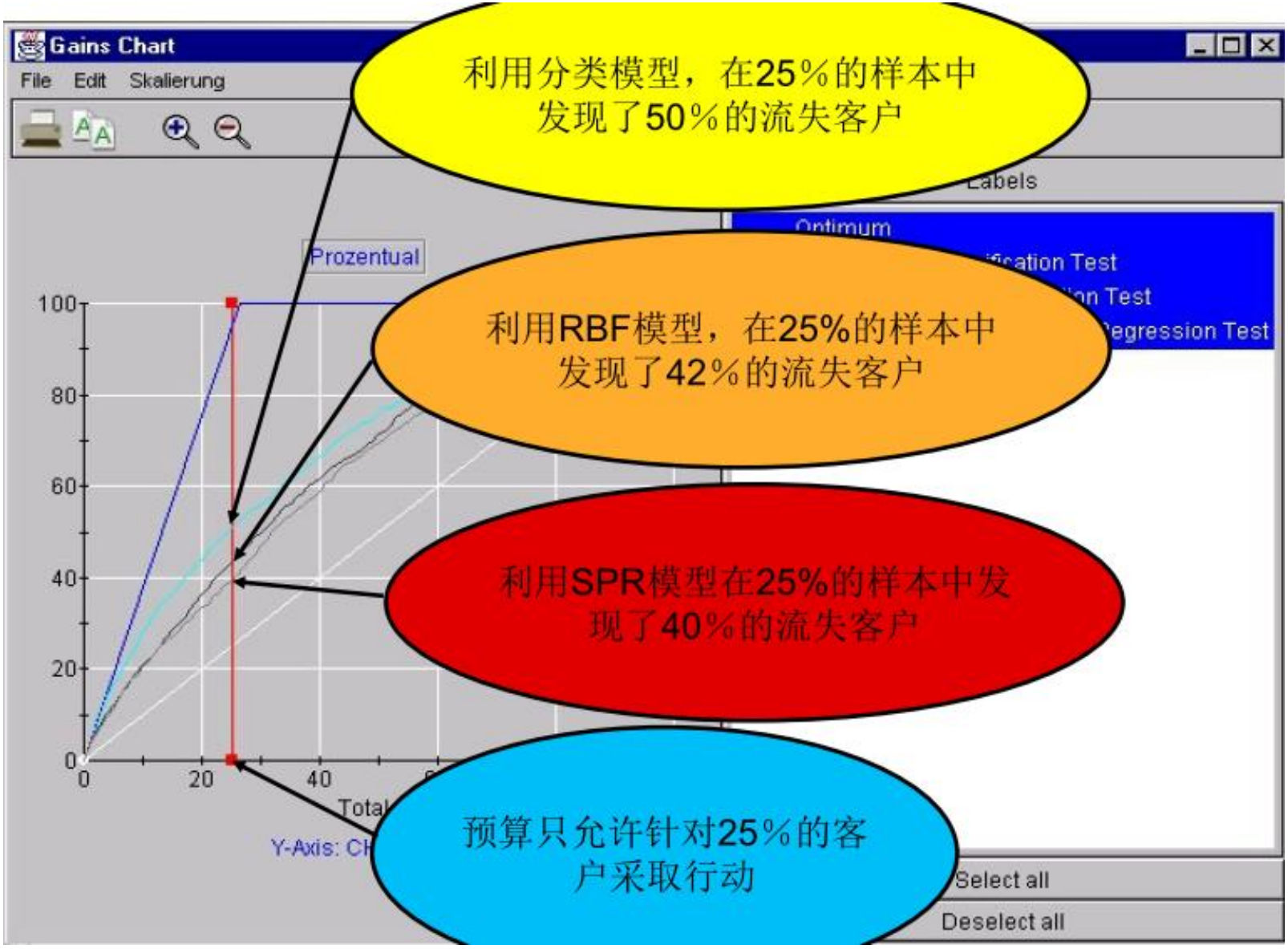
验证分类训练的结果


DB2 Intelligent Miner - Classification Tree
File View
0
1
如果此客户
- 不在高话费类中
- 没有出现三个月滑动话费总额高的情况
- 话费增长很快
- 不是生活在大城市
则此客户很可能会流失
REVENUE_OF_3MONTHS<187.46
DURATION_OF_CONTRACT<10815
Mouse Buttons:
Node Action
Left : collapse
Right: show info
Navigate
Left : zoom in
Right: zoom out
Show Tree Map
Prune Tree...
Save Tree...
Optimize Sparse Trees
Parent Class Distribution
Show Labels
Exit
DB2 Intelligent Miner - Decision Path
Leaf Node:
If (REVENUE_CATEGORY not in [RC6 RC5])
and (REVENUE_OF_3MONTHS < 14005.585000)
and (REVENUE_DEVELOPMENT >= 1113.530030)
and (LOCATION_SIZE not in [> 500,000])
then Class = 1
Show Distributions

确定分类模型


训练集数据的分类错误
DB2 Intelligent Miner - Classification Results
Tree Classification Training
Number of classes = 2
Errors = 1632 (17.81%)
Confusion matrix for pruned tree
DB2 Intelligent Miner - Classification Results
File Edit
测试及数据的分类错误
Tree Classification Test
Number of classes = 2
Errors = 2050 (22.23%)
Confusion matrix for pruned tree
Predicted Class --> | 0 | 1 |
0 | 6278 | 491 | total = 6769
1 | 1559 | 895 | total = 2454
7837 | 1386 | total = 9223
Show Applied Tree

挑选最佳模型并采取行动



快速投资回报

- **假使在一年之内**
  - 100,000客户， 每年26%的流失率
  - 25%的客户被选定为需要采取防止流失措施
  - 按照决策树分类模型进行选择，即这25%的客户包括了所有流失客户的 50%
  - 防止流失措施的成功率为50%
  - 平均每个流失客户损失￥500
  - 对每个客户采取防止流失措施平均花费￥100

快速投资回报

- **投资：**
  - 购买数据挖掘产品Intelligent Miner for Data(一次性费用)：￥70,000
  - 30天服务费用: ￥45,000
  - 可变费用100,000 * 25% * 100 = ￥2,500,000
- **回报：**
  - 100,000 * 26% * 50% * 50 % * 500 = ￥3,250,000
- **头一年的收益：**
  - ￥635,000：（3,250,000 - 2,500,00）

分类、预测分析（Classification & Prediction）

- 基本概念
- 分类分析的经典方法
- 预测分析的常用方法
- 分类、预测方法的评估
- 应用案例
- 总结

期末课程论文说明

- **主题要求**
  - 必须与“大数据管理”相关
  - 建议围绕所学专业背景下的“大数据管理问题”展开
- **内容要求**
  - 不少于4000字，版式：word中正文小四字体，1.5倍行距
  - 独立完成，不得大段拷贝或直接引用网上、书上及他人已发布内容，需要适当引用时请在引用位置注明参考文献来源（查重）
  - 论文内容框架（建议）：
    - 1. 学习本课程的心得体会、感受，对本课程教学的建议和意见（必有）
    - 2. 论文背景介绍
    - 3. 论文涉及的大数据问题及管理需求、策略和意义（可举实例说明）
    - 4. 本人对该大数据问题的看法、观点及讨论
    - 5. 总结
    - 6. 参考文献和资料

期末课程论文说明（续）

- **论文提交要求**
  - 需要以电子版提交，建议提交word版本
  - 作业提交邮箱：bigdata_homework@163.com
  - 作业提交截止时间：**第19周周日（2015.01.11）24时**
- **其他说明**
  - **电子版论文文件请务必按照“学号_班级_姓名.docx”命名，例如“2014211234_2014212103_张三.docx”，也请在邮件中留下姓名、学号及联系方式，以备论文有问题时能够联系到；**
  - 请在截止时间之前提交论文（不要在截止时间附近，以避免系统原因过期），过期将不再接收论文提交，成绩为0，请务必注意；
  - 每次提交论文后，作业邮箱都会有“已收到邮件”的自动回复，如未收到自动回复，表示发送不成功，请在截止时间内重新提交；
  - 论文评分的关注重点
    - 有效的课程建议和意见
    - 关注问题的新颖度
    - 个人分析和讨论的深度
    - 论文的整体工作量