第4讲 MATLAB数据挖掘算法（下）

卓金武
MathWorks 中国
steven.zhuo@mathworks.cn

课程介绍

- **第1讲：MATLAB快速入门**
- MATLAB快速入门实例
- MATLAB实用操作技巧
- MATLAB数据类型
- MATLAB程序结构
- MATLAB编程模式
- MATLAB学习理念

- **第2讲：MATLAB数据挖掘基础**
- MATLAB数据挖掘的过程
- 数据的可视化
- 数据的预处理
- 数据的探索
- 假设检验
- 数据回归

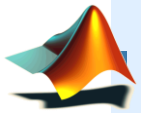
- **第3讲：MATLAB数据挖掘算法（上）**
- 回归算法
- 关联算法
- 聚类算法

- **第4讲：MATLAB数据挖掘算法（下）**
- 分类算法
- 预测算法
- 异常诊断算法

- **第5讲：MATLAB高级数据挖掘技术**
- MATLAB分类学习机
- 算法的高级使用方法
- 综合使用实例

- **第6讲：MATLAB数据挖掘项目实例**
- 故障诊断
- 生物信息学研究
- 量化投资

内容提要



分类算法

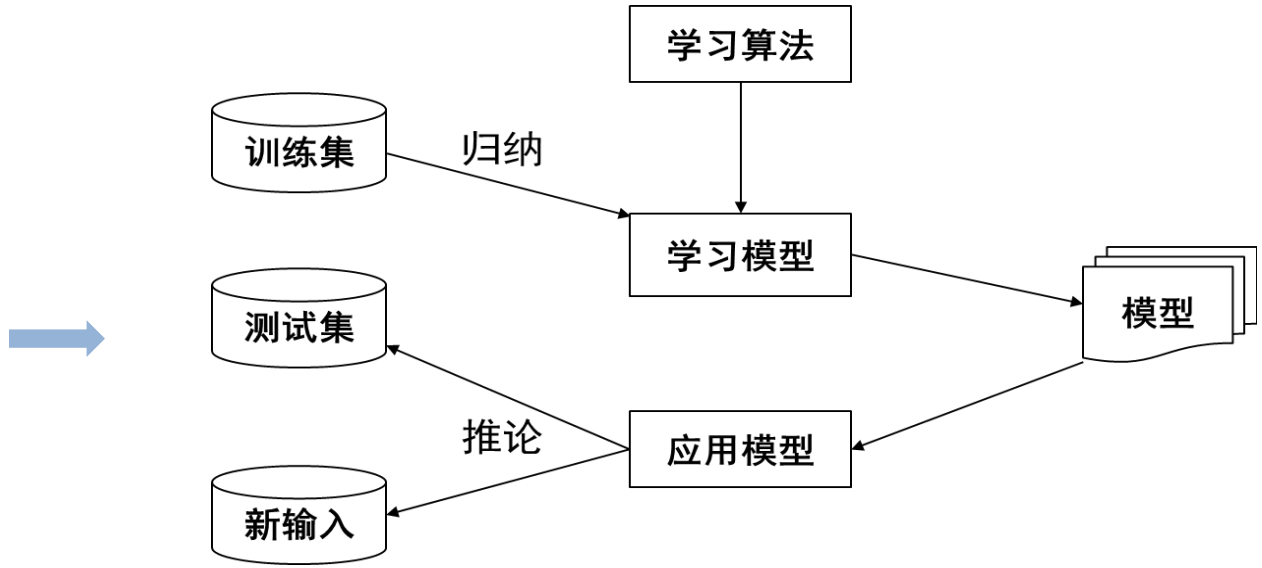
- 预测算法
- 异常诊断算法

分类的概念

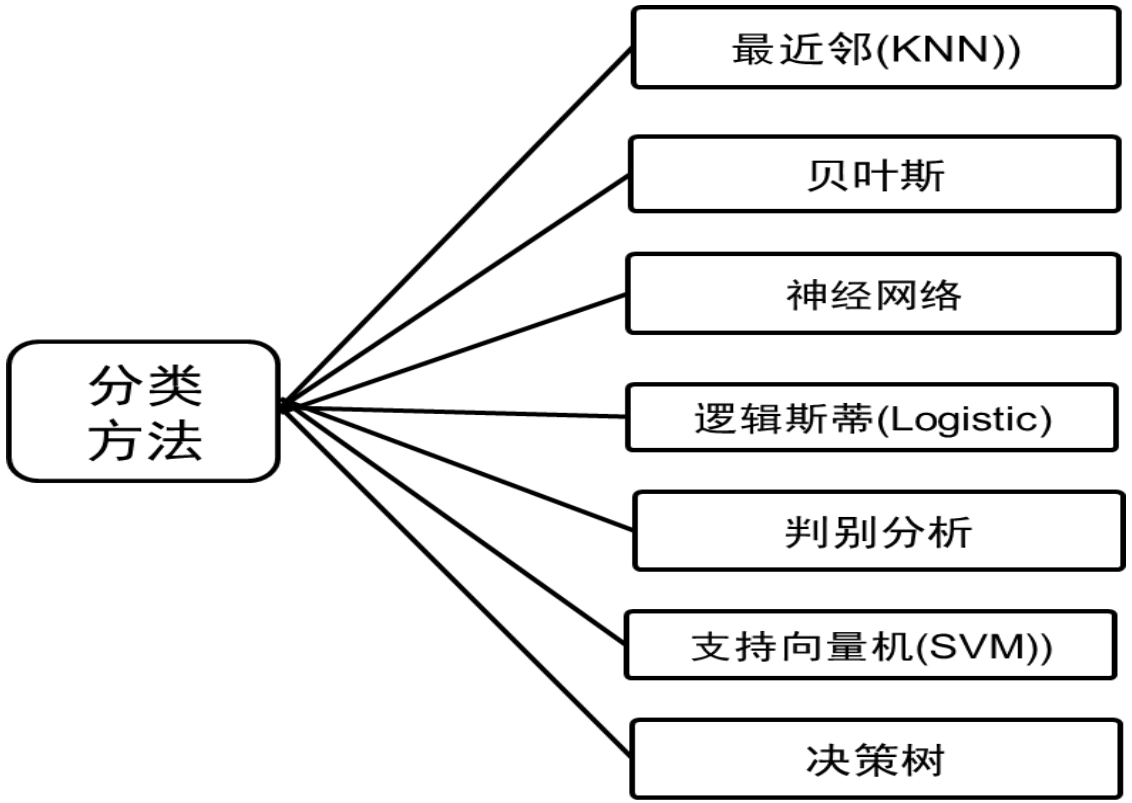
分类的定义：对现有的数据进行学习，得到一个目标函数或规则，把每个属性集x映射到一个预先定义的类标号y。

目标函数或规则也称分类模型（classification model），分类模型有两个主要作用：
一是描述性建模，即作为解释性的工具，用于区分不同类中的对象；
二是预测性建模，即用于预测未知记录的类标号。

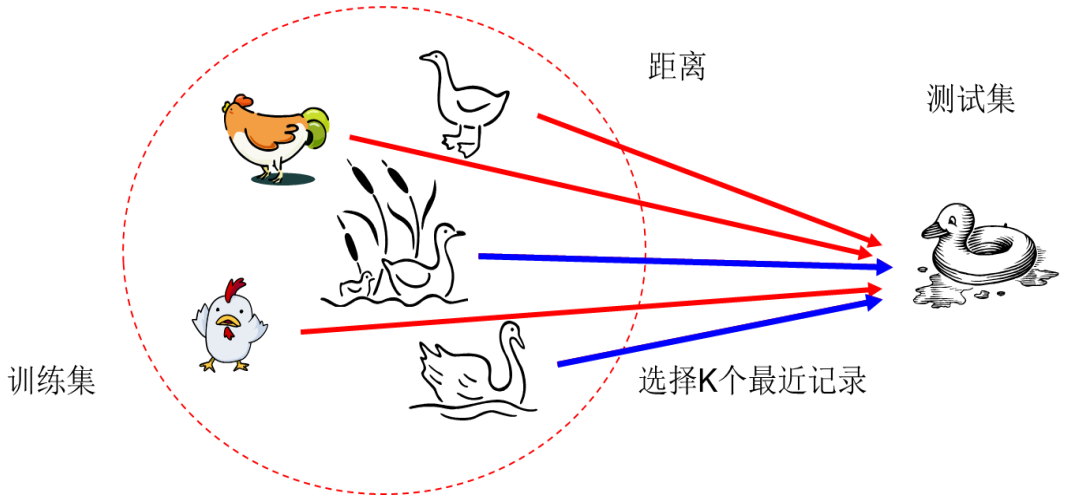
原理：需要一个训练集（training set），它由类标号已知的记录组成。使用训练集建立分类模型，该模型随后将运用于检验集（test set），检验集由类标号未知的记录组成。



常用的分类方法



K-近邻：概念与步骤



K-近邻分类方法通过计算每个训练样例到待分类样品的距离，取和待分类样品距离最近的K个训练样例，K个样品中哪个类别的训练样例占多数，则待分类元组就属于哪个类别。

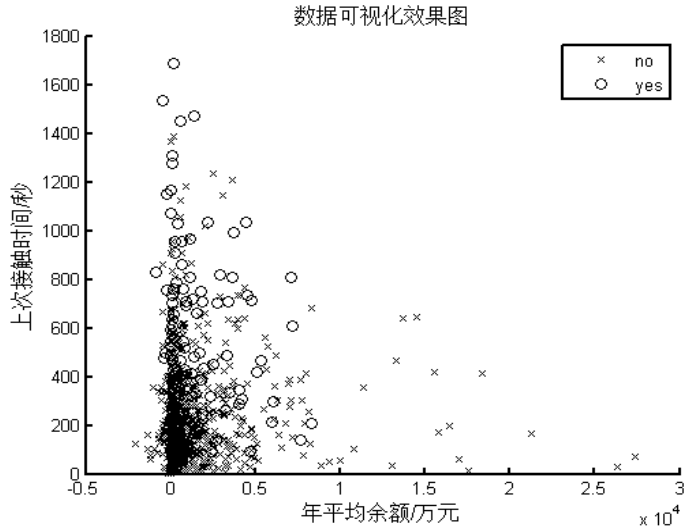
KNN算法具体步骤如下：
step1：初始化距离为最大值。
step2：计算未知样本和每个训练样本的距离dist。
step3：得到目前K个最邻近样本中的最大距离maxdist。
step4：如果dist小于maxdist，则将该训练样本作为K-最近邻样本。
step5：重复步骤step2、step3、step4，直到未知样本和所有训练样本的距离都算完。
step6：统计K个最近邻样本中每个类别出现的次数
step7：选择出现频率最大的类别作为未知样本的类别。

K-近邻：应用实例和MATLAB实现过程

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练KNN分类器

Matlab函数：
ClassificationKNN.fit

最近邻方法分类结果：

```
C_knn =
   352     8
    28    12
```

贝叶斯分类：概念与原理

贝叶斯分类是一类利用概率统计知识进行分类的算法，其分类原理是贝叶斯定理。

假设X，Y是一对随机变量，它们的联合概率P（X=x，Y＝y）是指X取值x且Y取值y的概率，条件概率是指一随机变量在另一随机变量取值已知的情况下取某一特定值的概率。例如，条件概率P（Y＝y|X=x）是指在变量X取值x的情况下，变量Y取值y的概率。X和Y的联合概率和条件概率满足如下关系：

$$P(X,Y)=P(Y\mid X)\times P(X)=P(X\mid Y)\times P(Y)$$

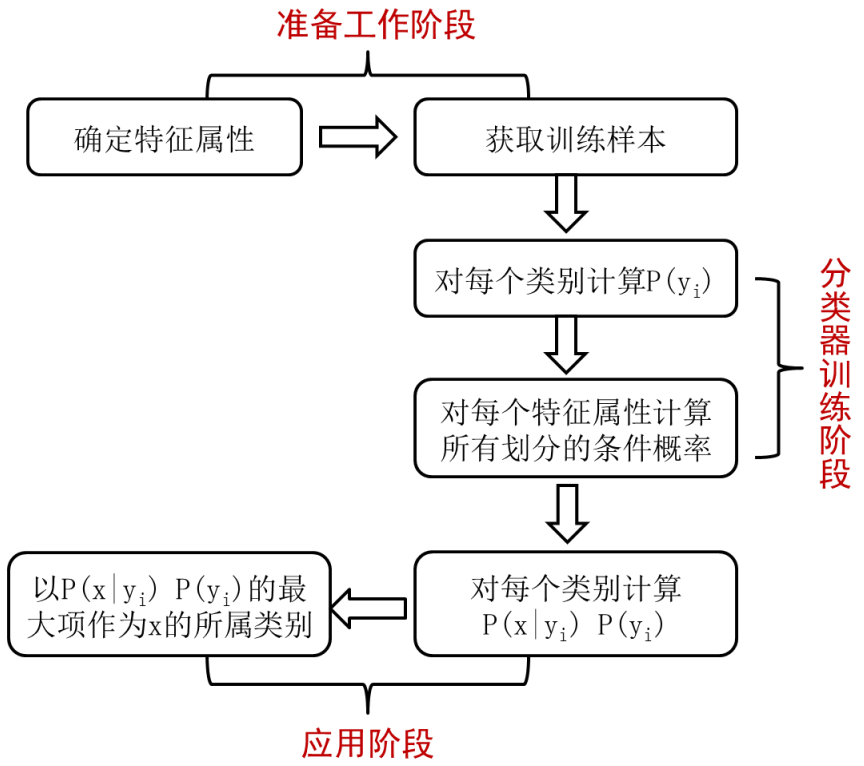
对此式变形，可得到下面公式，称为贝叶斯定理：

$$P(Y\mid X)=\frac{P(X\mid Y)\times P(Y)}{P(X)}$$

贝叶斯分类：朴素贝叶斯分类

朴素贝叶斯的思想基础：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

朴素贝叶斯分类分为三个阶段：
（1）准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的准备。
（2）分类器训练阶段，这个阶段的任务就是生成分类器。
（3）应用阶段，这个阶段的任务是使用分类器对待分类项进行分类。

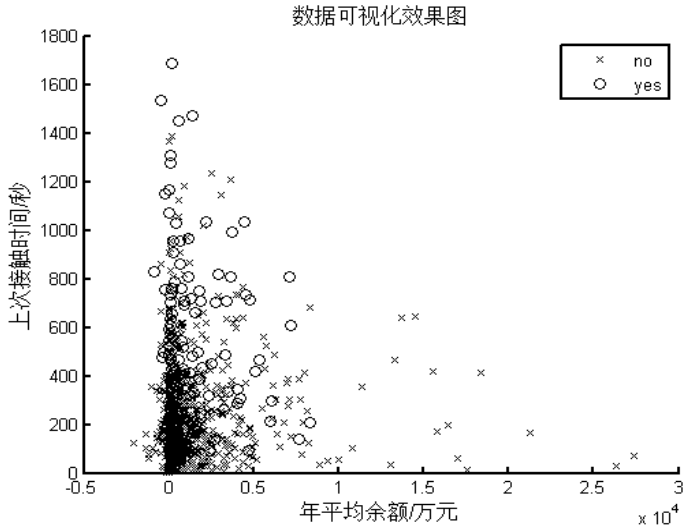


贝叶斯分类：应用实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练朴素贝叶斯分类器

Matlab函数：
NaiveBayes.fit

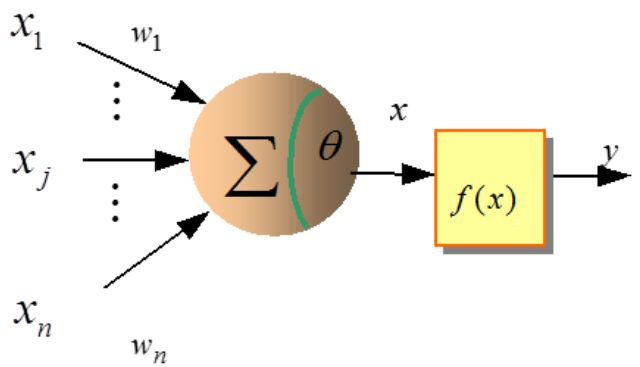
贝叶斯方法分类结果：

```
C_nb =
   305    55
    19    21
```

神经网络：原理

人工神经网络（Artificial Neural Networks，ANN）是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型。

一个简单的神经网络结构——感知器。感知器包含两种结点：几个输入结点，用来表示输入属性；一个输出结点，用来提供模型输出。神经网络结构中的结点通常叫作神经元或单元。在感知器中，每个输入结点都通过一个加权的链连接到输出结点。这个加权的链用来模拟神经元间神经键连接的强度。像生物神经系统一样，训练一个感知器模型就相当于不断调整链的权值，直到能拟合训练数据的输入输出关系为止。

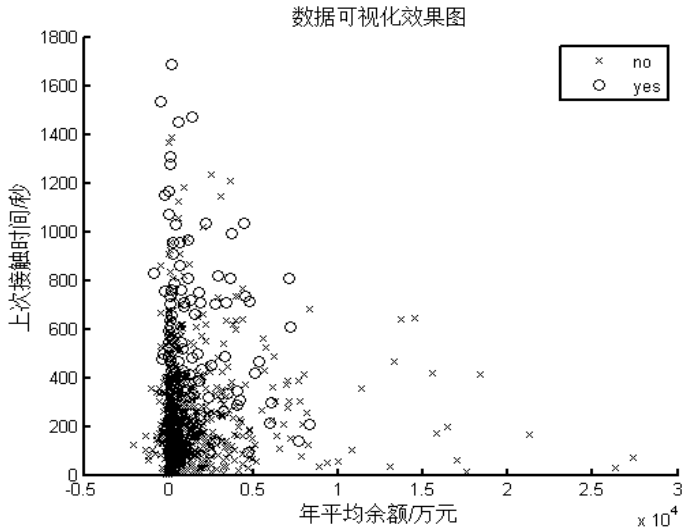


神经网络：分类实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练神经网络分类器

Matlab函数：
train

```
C_nn =
   348    12
    26    14
```

Logistic 分类：原理

Y是一个定性的变量，比如，Y=0或Y=1 ，这时就不能用通常的regress函数对y进行回归，而是使用Logistic回归。Logistic方法主要应用在研究某些现象发生的概率 ，比如股票涨还是跌，公司成功或失败的概率。 Logistic回归模型的基本形式为：

$$P(Y = 1 \mid x_1, x_2, \cdots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

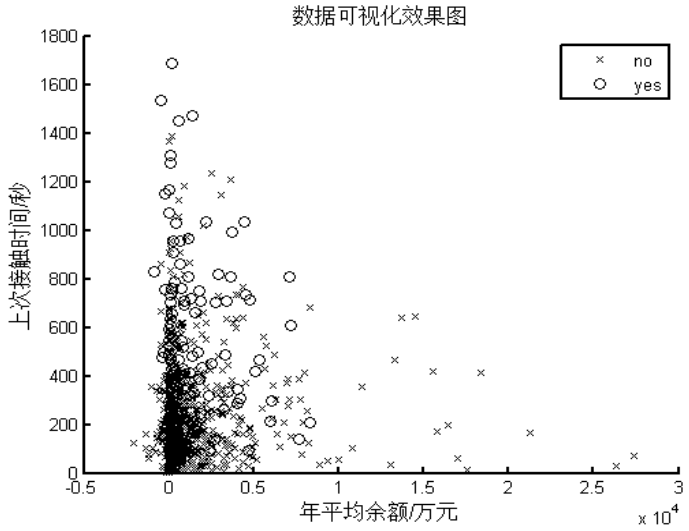
该式表示当自变量为 $x_1, x_2, \cdots, x_k$ 时， 因变量 Y 为1的概率。

Logistic 分类：应用实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练logistic分类器

Matlab函数：
fitglm

```
C_glm =
   345    15
    20    20
```

判别分析：原理

判别分析（Discriminant Analysis，简称DA）技术是根据观察或测量到的若干变量值判断研究对象如何分类的方法。具体地讲，就是已知一定数量案例的一个分组变量（grouping variable）和这些案例的一些特征变量，确定分组变量和特征变量之间的数量关系，建立判别函数（discriminant function），然后便可以利用这一数量关系对其他已知特征变量信息、但未知分组类型所属的案例进行判别分组。

判别分析的基本模型就是判别函数，它表示为分组变量与满足假设的条件的判别变量的线性函数关系，其数学形式为：

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k$$

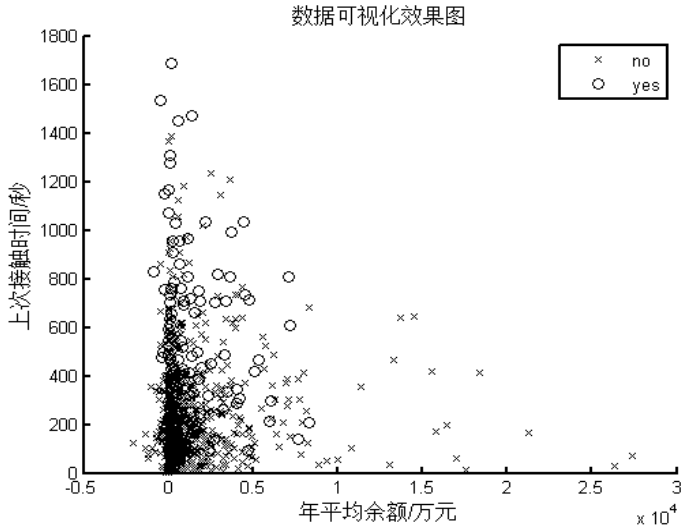
其中，y 是判别函数值，又简称为判别值（discriminant score）；x_i 为各判别变量；b_i 为相应的判别系数（dicriminant coefficient or weight），表示各判别变量对于判别函数值的影响，其中b_0是常数项。

判别分析：应用实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

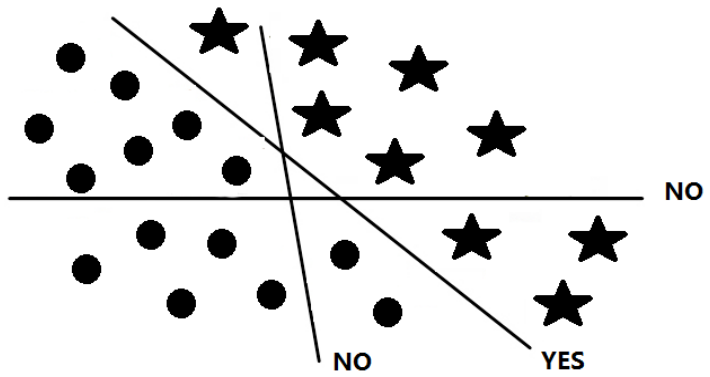
（4）训练判别分析分类器

Matlab函数：
ClassificationDiscriminant.fit

```
C_da =
   343    17
    21    19
```

支持向量机：原理

SVM构建了一个分割两类的超平面（这也可以扩展到多类问题）。在构建的过程中，SVM算法试图使两类之间的分割达到最大化，如下图所示。



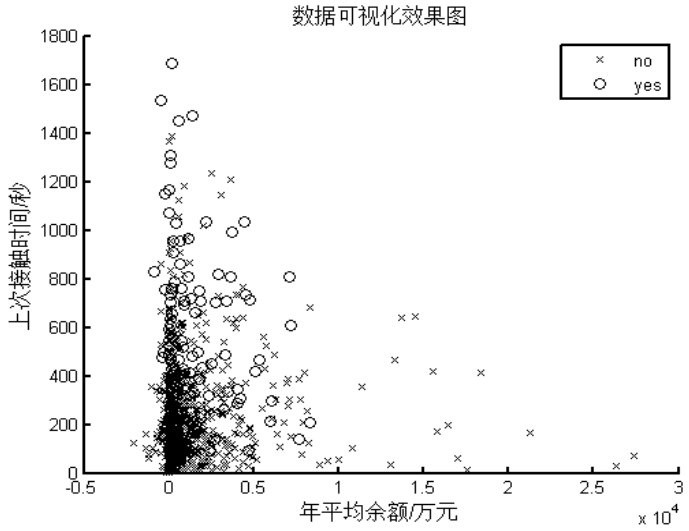
支持向量机的基本思想：与分类器平行的两个平面，此两个平面能很好地分开两类不同的数据，且穿越两类数据区域集中的点，现在欲寻找最佳超几何分隔平面使之与两个平面间的距离最大，如此便能实现分类总误差最小。

支持向量机：应用实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练SVN分类器

Matlab函数：
svmtrain

```
C_svm =
   276    84
     9    31
```

决策树：原理

决策树是一种监督式的学习方法，产生一种类似流程图的树结构。决策树对数据进行处理是利用归纳算法产生分类规则和决策树，再对新数据进行预测分析。树的终端节点”叶子节点（leaf nodes）”，表示分类结果的类别（class），每个内部节点表示一个变量的测试，分枝（branch）为测试输出，代表变量的一个可能数值。为达到分类目的，变量值在数据上测试，每一条路径代表一个分类规则。

1、选择适当的算法训练样本建构决策树
2、适当的修剪决策树
3、从决策树中萃取知识规则

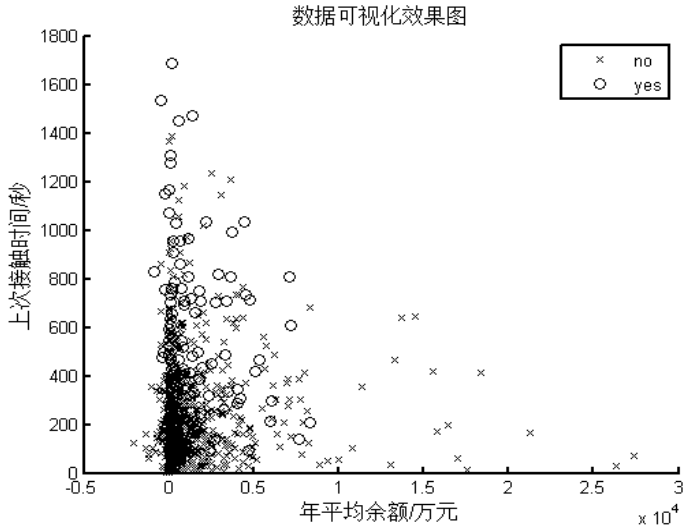
常用算法有：（1）ID3算法
（2）C4.5算法
（3）C5.0算法
（4）CART算法

决策树：应用实例

MATLAB中具体的实现步骤和结果如下：

（1）准备环境

（2）导入数据及数据预处理



（3）设置交叉验证方式

（4）训练决策树分类器

Matlab函数：
ClassificationTree.fit

```
C_t =
   326    34
    19    21
```

分类的评价：正确率

		预测的类	
		类1	类0
实际的类	类1	TP	FN
	类0	FP	TN

$$\text{正确率} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{错误率} = 1-\frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{灵敏性} = \frac{TP}{TP+FN}$$

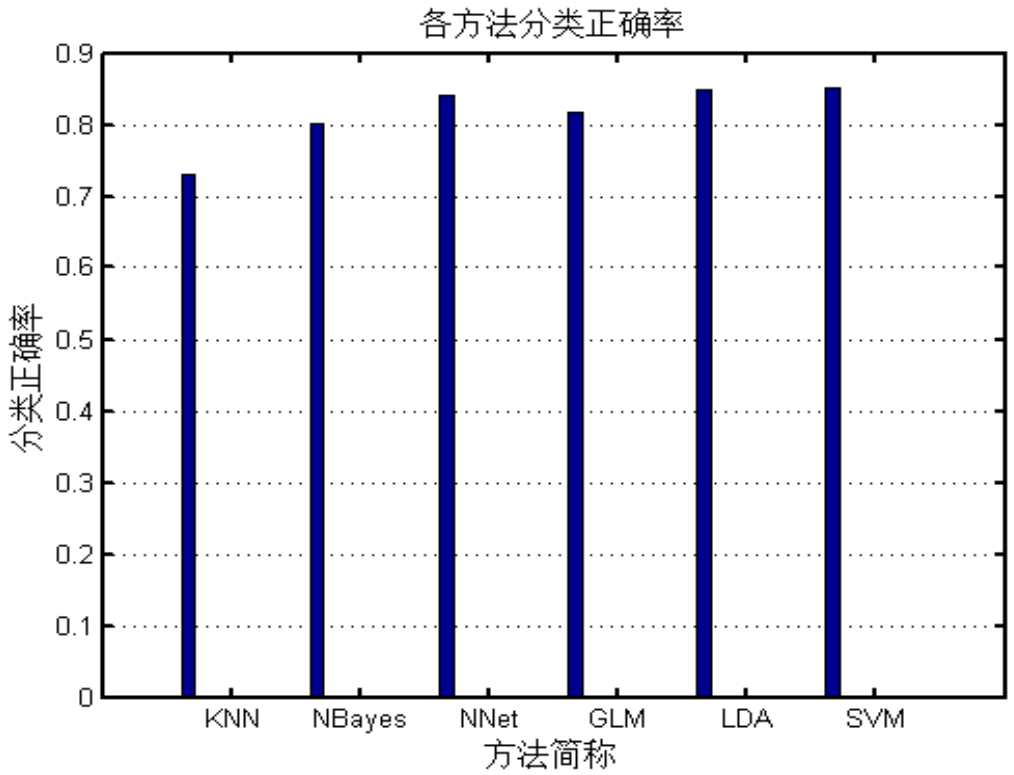
$$\text{特效性} = \frac{TN}{TN+FP}$$

$$\text{精度} = \frac{TP}{TP+FP}$$

$$\text{错正率} = \frac{FP}{TN+FP}$$

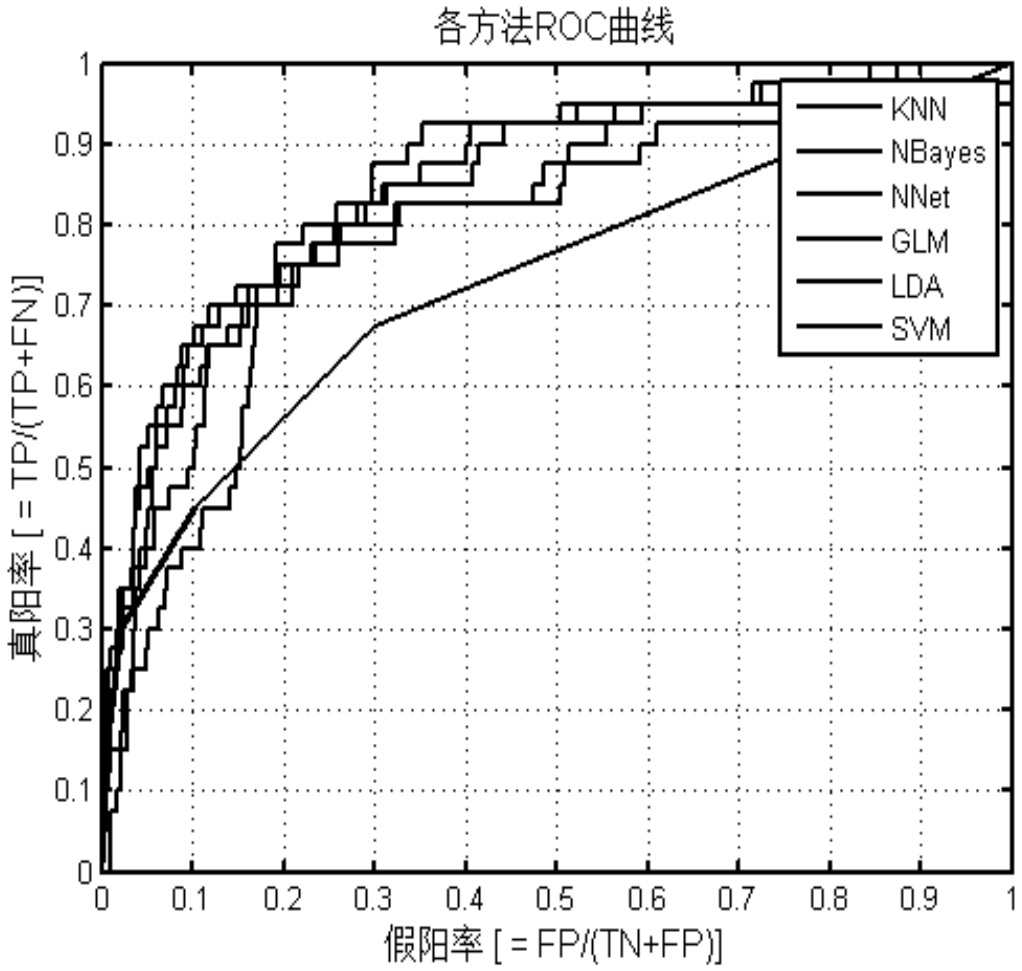
$$\text{负元正确率} = \frac{TN}{TN+FN}$$

$$\text{正元正确率} = \frac{FP}{TP+FP}$$



分类的评价：RCO曲线

ROC曲线图：其横轴是假正率，其纵轴是真正率，该图同时显示了一条对角线。ROC曲线离对角线越近，模型的准确率就越低。图中最上面的曲线所代表的神经网络模型（Neural）的准确率就要高于其下面的曲线所代表的逻辑回归模型（Reg）的准确率。



分类应用实例：股票分类

SID：表示股票的编号
X1-X8：为8个指标
Y：则为股票的涨跌状态

训练样本 =

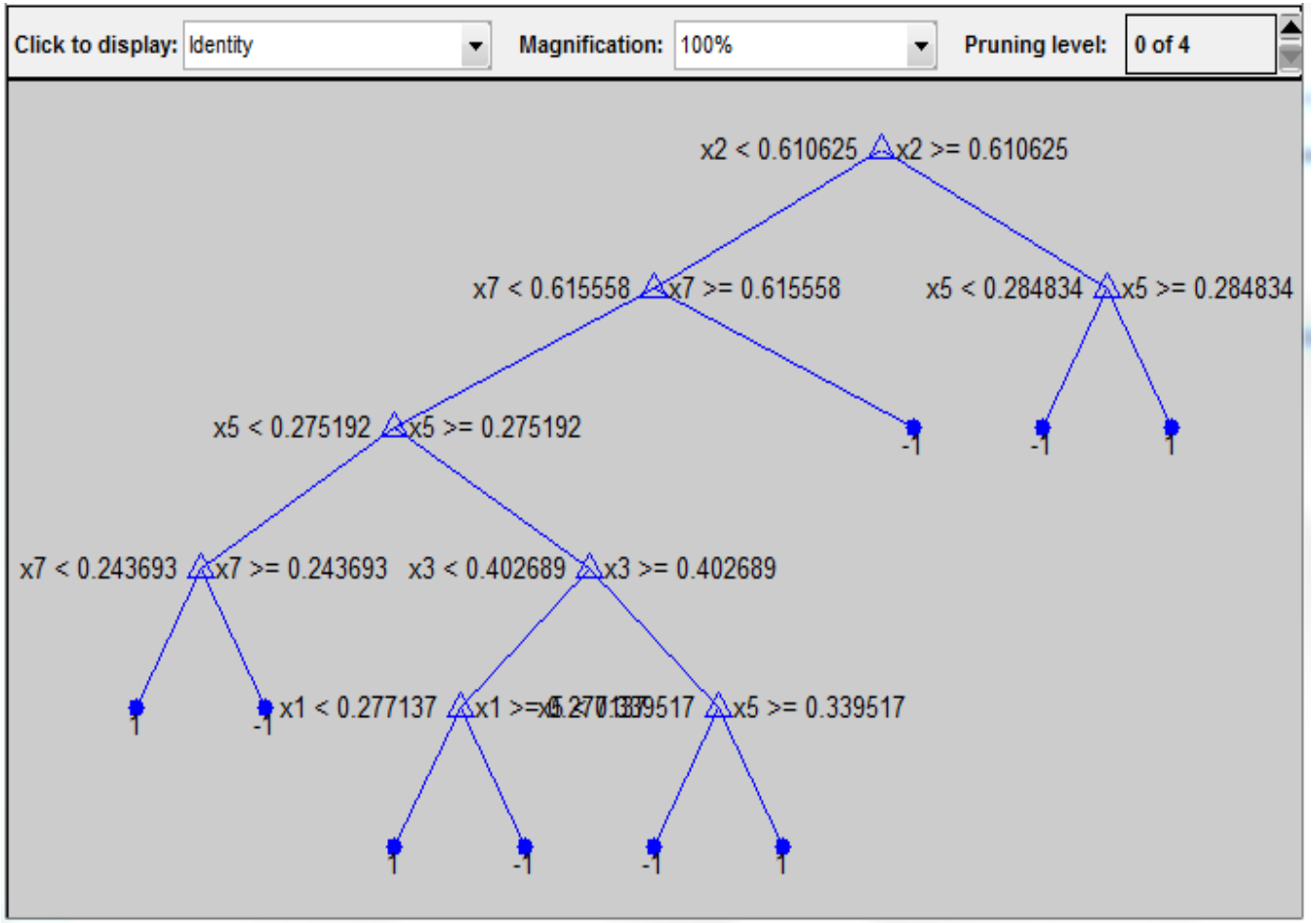
SID	X1	X2	X3	X4	X5	X6	X7	X8	Y
938	0	0.664186	0.568177	0.676557	0.807005	0.822039	0.159019	0.100741	1
955	0.369022	0.373625	0.444087	0.509025	0.594756	0.599438	0.287933	0.256512	1
957	1	1	1	1	0.841517	0.60381	0.549302	0.878456	1
957	0.719588	1	0.568177	0.676557	0.316724	0.630255	0.335306	0.631347	1
957	0.875241	1	0.568177	0.676557	0.620273	0.669086	0.453491	0.535071	1
973	0.7066	0.585696	1	1	0.681068	0.879814	0.353326	0.461259	1
977	0.093164	0.671944	0.302117	0.173961	0.237053	0.192934	0.128049	0.063319	1
1	0.203524	0.319353	0.361291	0.341493	0.073104	0.390368	0.456691	0.272403	-1
1	0.30148	0.336599	0.302117	0.173961	0.315328	0.452827	0.649412	0.502396	-1
1	0.331083	0.35282	0.302117	0.173961	0.755897	0.480425	0.778511	0.611119	-1
1	0.287717	0.298582	0.302117	0.173961	0.610018	0.368464	0.680474	0.606938	-1
1	0.362609	0.197105	0.361291	0.341493	0.185195	0.112184	0.680197	0.644573	-1
1	0.522551	0.229589	0.444087	0.509025	0.15945	0.181541	0.85895	0.751893	-1
1	0.465448	0.315421	0.361291	0.341493	0.214698	0.217401	0.85895	0.751893	-1
1	0.430687	0.373518	0.361291	0.341493	0.34743	0.389612	0.914562	0.804285	-1

预测样本 =

SID	X1	X2	X3	X4	X5	X6	X7	X8
1	0.370518	0.228827	0.417467	0.456905	0.65828	0.269891	0.61134	0.440261
2	0.410009	0.223225	0.302097	0.244671	0.521301	0.531944	0.553721	0.322291
4	0.773221	0.372947	1	1	0.930767	0.894786	0.7914	0.487435
5	0.437815	0.38571	0.590378	0.669138	0.553453	0.458256	0.539316	0.331551
6	0.104282	0.140815	0.417467	0.456905	0.60773	0.358791	0.246194	0.248542
7	0.811114	0.462535	0.590378	0.669138	0.891488	0.814734	0.7914	0.771868
8	0.30237	0.353455	0.417467	0.456905	0.33083	0	0.060928	0.183124
9	0.746101	0.553203	0.590378	0.669138	0.503008	0.57995	0.7914	0.658269
10	0.391251	0.53078	0.417467	0.456905	0.308306	0.47891	0.703464	0.706443

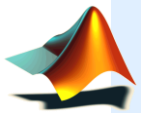
分类应用实例：股票分类过程及结果

选择决策树作为股票的分类器，在训练分类器后，程序可以显示该分类器的具体决策树结构图。通过该图，也可以看出，在这个问题中，哪些变量在哪些层次发生了作用，根据分类的依据，可以看出好股票的的指标具有哪些特点，这对于股票的技术分析师非常有帮助的。



内容提要

- 分类算法

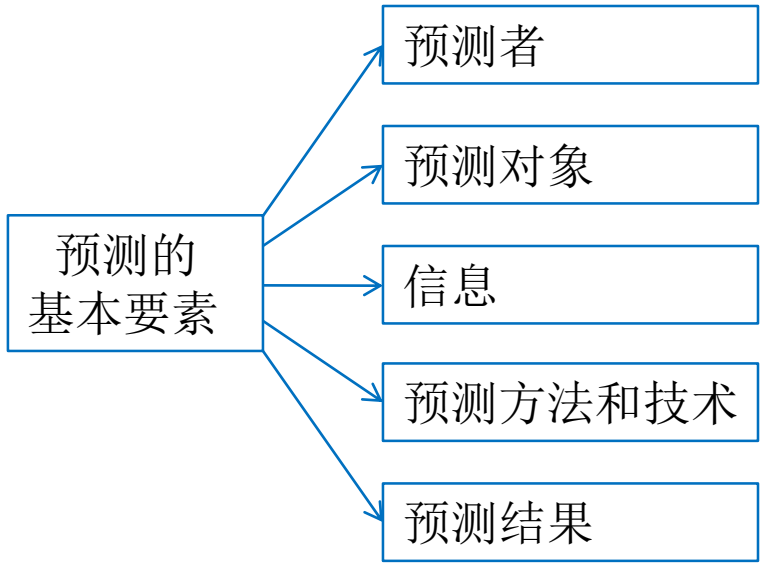


预测算法

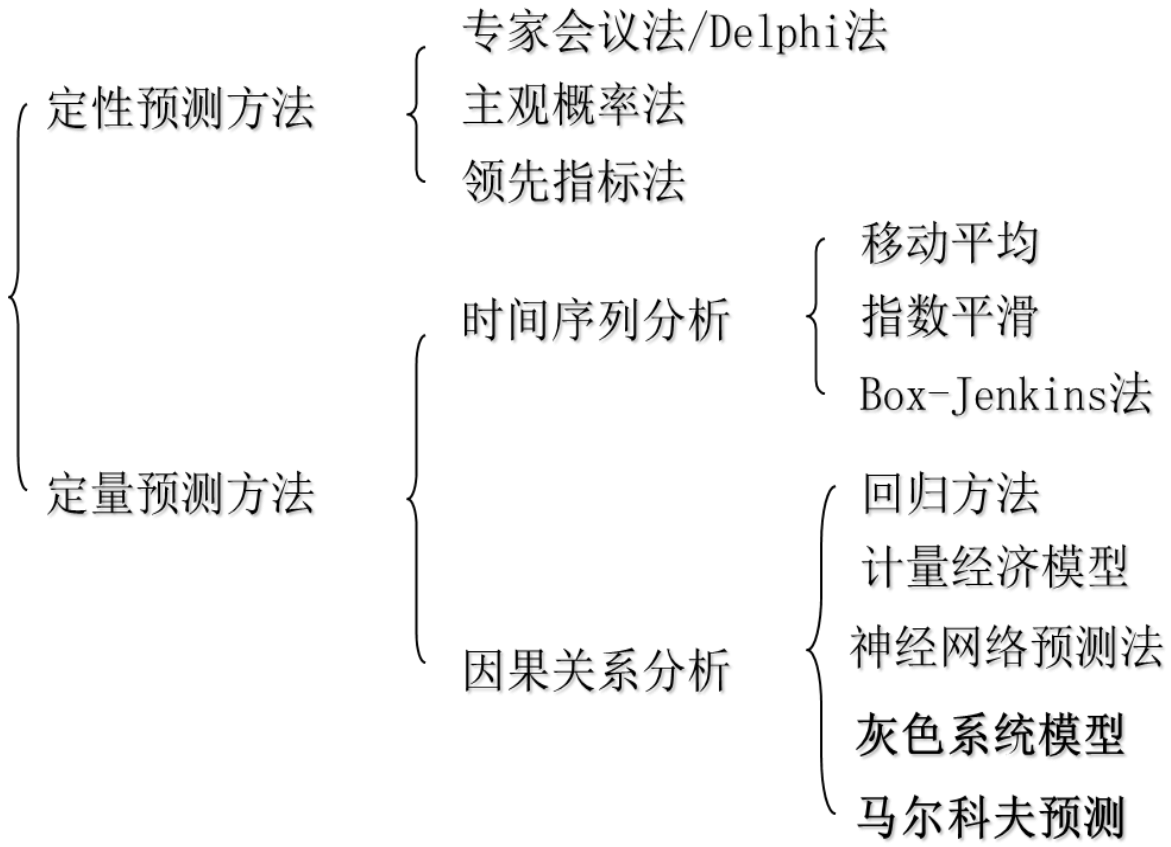
- 异常诊断算法

预测的概念

预测是指根据客观事物的发展趋势和变化规律对特定的对象未来发展趋势或状态作出科学的推断与判断，即预测就是根据过去和现在估计未来。



常用的预测方法


定性预测方法
专家会议法/Delphi法
主观概率法
领先指标法
定量预测方法
时间序列分析
移动平均
指数平滑
Box-Jenkins法
因果关系分析
回归方法
计量经济模型
神经网络预测法
灰色系统模型
马尔科夫预测

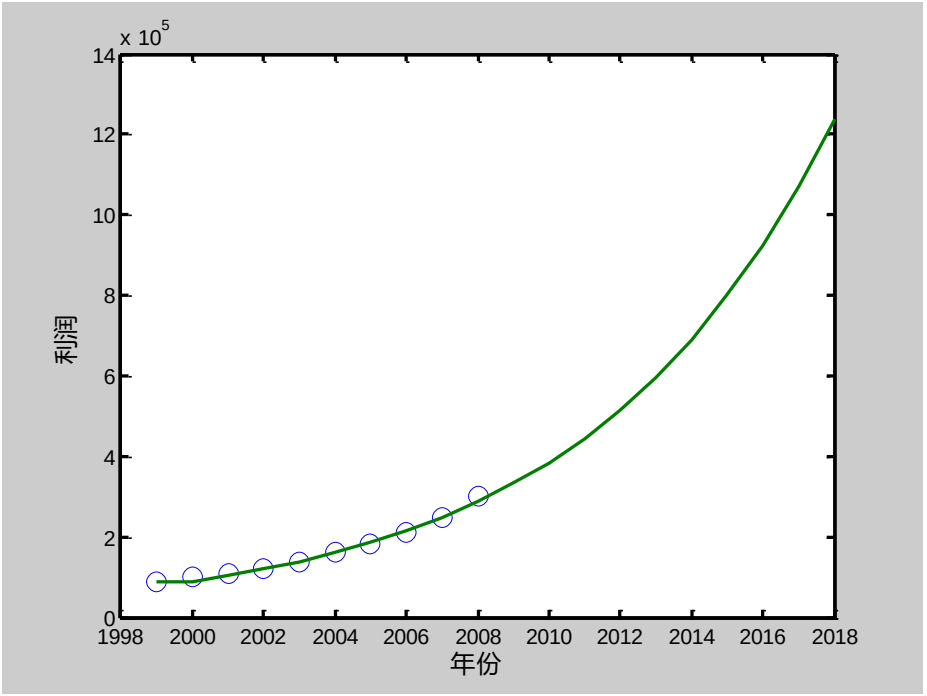
灰色预测的实例

步骤：
1、对原始数据进行累加；
2、构造累加矩阵B与常数向量；
3、求解灰参数；
4、将参数带入预测模型进行数据预测。

1999至2008年的利润数据：

灰色预测 →

下图中，蓝色圆圈为原始数据，绿色曲线为预测数据。



马尔科夫预测原理

马尔科夫过程是具有马尔科夫性质的离散随机过程。众所周知，事物总是随着时间而发展的，因此事物与时间之间有一定的变换关系。在一般情况下，人们要了解事物未来的发展状态，不但要看到事物现在的状态，还要看到事物过去的状态。安德烈·马尔可夫认为，还存在另外一种情况，人们要了解事物未来的发展状态，只需知道事物现在的状态，而与事物以前的状态毫无关系。马尔科夫过程的理论在近代物理、生物学、管理科学、经济、信息处理以及数字计算方法等方面都有重要应用。在此过程中，在给定当前信息或知识时，过去对于预测未来是无关的。

马尔科夫的定义：

设 $\{X_t, t \in T\}$ 为随机过程，若对任意正整数n及 $t_1 < t_2 < \cdots < t_n$ ，有：

$$P\{X_{t_1} = x_1, \cdots, X_{t_{n-1}} = x_{n-1}\} > 0$$

且条件分布：

$$P\{X_{t_n} \le x_n \mid X_{t_1} = x_1, \cdots, X_{t_{n-1}} = x_{n-1}\} = P\{X_{t_n} \le x_n \mid X_{t_{n-1}} = x_{n-1}\}$$

则称 $\{X_t, t \in T\}$ 为马尔科夫过程。

马尔科夫预测实例：大盘趋势判断（数据预处理）

A股指数数据

2012年11月6日	2205.43	2012年11月29日	2127.57
2012年11月7日	2205.17	2012年11月30日	2124.78
2012年11月8日	2169.23	2012年12月1日	2158.91
2012年11月9日	2166.64	2012年12月2日	2181.98
2012年11月12日	2177.38	2012年12月3日	2172.48
2012年11月13日	2144.47	2012年12月4日	2180.92
2012年11月14日	2152.37	2012年12月5日	2158.57
2012年11月15日	2125.99	2012年12月6日	2252.13
2012年11月16日	2109.69	2012年12月7日	2262.30
2012年11月17日	2112.02	2012年12月8日	2264.50
2012年11月18日	2103.54	2012年12月9日	2264.25
2012年11月19日	2125.96	2012年12月10日	2270.67
2012年11月20日	2110.54	2012年12月11日	2254.83
2012年11月21日	2122.92	2012年12月12日	2260.84
2012年11月22日	2112.52	2012年12月13日	2318.10
2012年11月23日	2084.98	2012年12月14日	2323.70
2012年11月24日	2066.51	2012年12月15日	2309.74
2012年11月25日	2055.99	2012年12月16日	2338.32
2012年11月26日	2073.24	2012年12月17日	2376.04
2012年11月27日	2051.96	2013年1月4日	2384.19
2012年11月28日	2068.08		

根据表中相邻两天的数据计算股价的增长率，如果增长超过1%记为上升，跌超过1%，则记为下降，其他情况记为持平。

股价变动状态

2012年11月7日	持平	2012年11月29日	上升
2012年11月8日	下降	2012年11月30日	持平
2012年11月9日	持平	2012年12月1日	上升
2012年11月12日	持平	2012年12月2日	上升
2012年11月13日	下降	2012年12月3日	持平
2012年11月14日	持平	2012年12月4日	持平
2012年11月15日	下降	2012年12月5日	下降
2012年11月16日	持平	2012年12月6日	上升
2012年11月17日	持平	2012年12月7日	持平
2012年11月18日	持平	2012年12月8日	持平
2012年11月19日	上升	2012年12月9日	持平
2012年11月20日	持平	2012年12月10日	持平
2012年11月21日	持平	2012年12月11日	持平
2012年11月22日	持平	2012年12月12日	持平
2012年11月23日	下降	2012年12月13日	上升
2012年11月24日	持平	2012年12月14日	持平
2012年11月25日	持平	2012年12月15日	持平
2012年11月26日	持平	2012年12月16日	上升
2012年11月27日	下降	2012年12月17日	上升
2012年11月28日	持平	2013年1月4日	持平

马尔科夫预测实例：大盘趋势判断（实现过程）

1、建立价格波动状态转移矩阵

$$P=\begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{bmatrix}=\begin{bmatrix} \frac{2}{8} & \frac{6}{8} & 0 \\ \frac{5}{26} & \frac{15}{26} & \frac{6}{26} \\ \frac{1}{6} & \frac{5}{6} & 0 \end{bmatrix}=\begin{bmatrix} 0.25 & 0.75 & 0 \\ 0.1923 & 0.5769 & 0.2308 \\ 0.1667 & 0.8333 & 0 \end{bmatrix}$$

2、马尔科夫过程的平稳分布与稳态条件下的解

$$\begin{cases} \pi_i=\sum_{i\in I}\pi_i p_{ij} \\ \sum_{j\in I}\pi_j=1,\pi_j\geq 0 \end{cases}\longrightarrow\begin{cases} z_1=0.25z_1+0.1923z_2+0.1667z_3 \\ z_2=0.75z_1+0.5769z_2+0.8333z_3 \\ z_3=0.2308z_2 \\ z_1+z_2+z_3=1 \end{cases}\longrightarrow$$

z1=5.88%

z2=76.47%

z3=17.65%

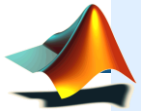
1、我国A股股价在2013年未来的变化趋势中，有5.88%的概率A股股价增长会高于1%，同时有76.47%的概率股价会在-1%~1%之间小范围徘徊，还会有17.65%的概率股价会下跌1%以上。

2、考虑到目前我国经济增速放缓的现状，以及在后金融危机时代贸易出口的困境，可以认为上述结论还是比较准确的。2012年我国沪市指数在历史较低点徘徊，所以可以认为未来下跌的空间并不是很大，此外由于经济增长乏力，导致大部分投资者看空未来经济，这会直接反映在未来股票市场的价格上，因此在2013年预计A股指数不会有太大的涨幅。

预测过程

内容提要

- 分类算法
- 预测算法



异常诊断算法

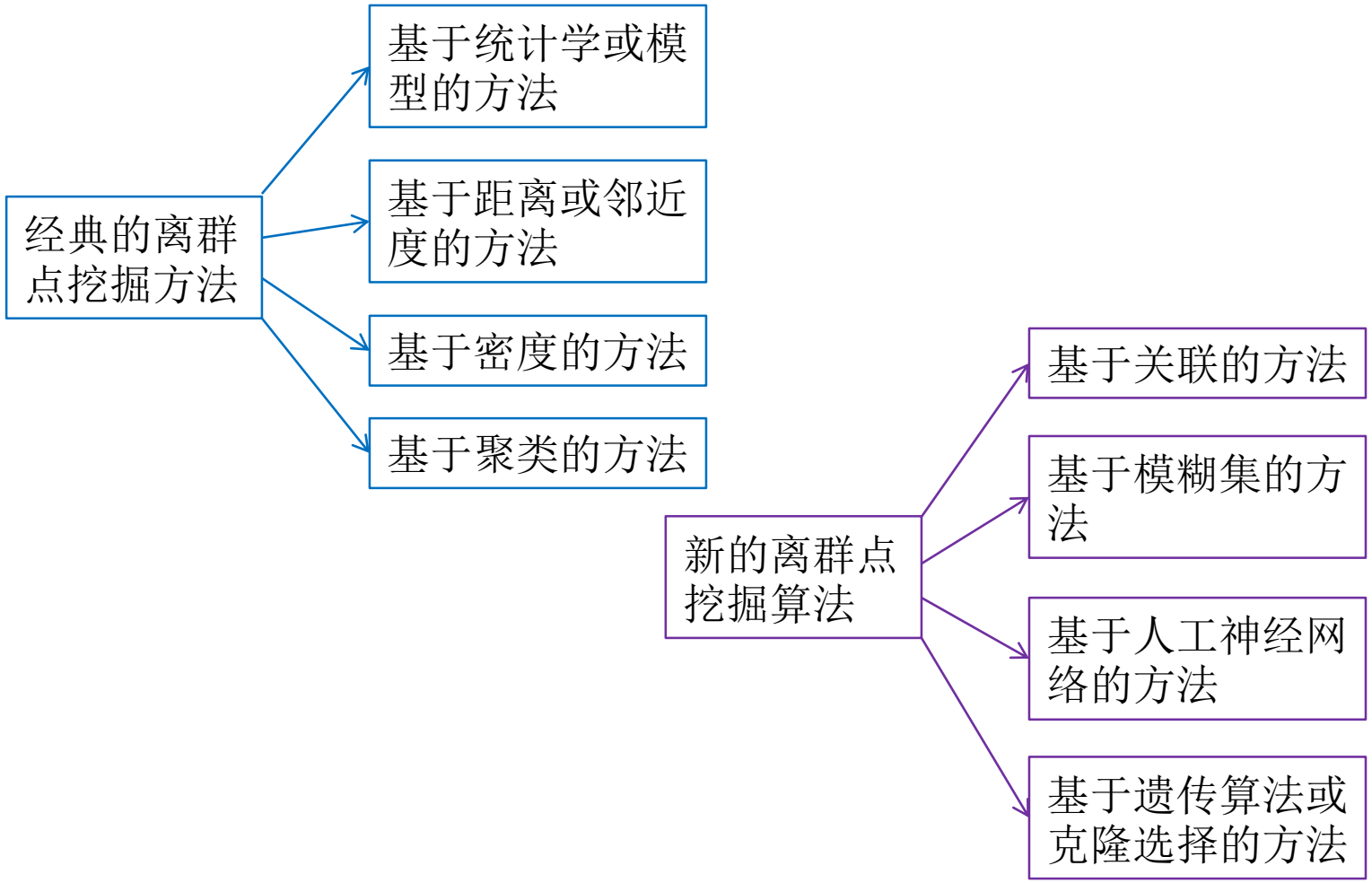
异常（离群点）诊断的定义

离群点（outlier）是指数值中，远离数值的一般水平的极端大值和极端小值。

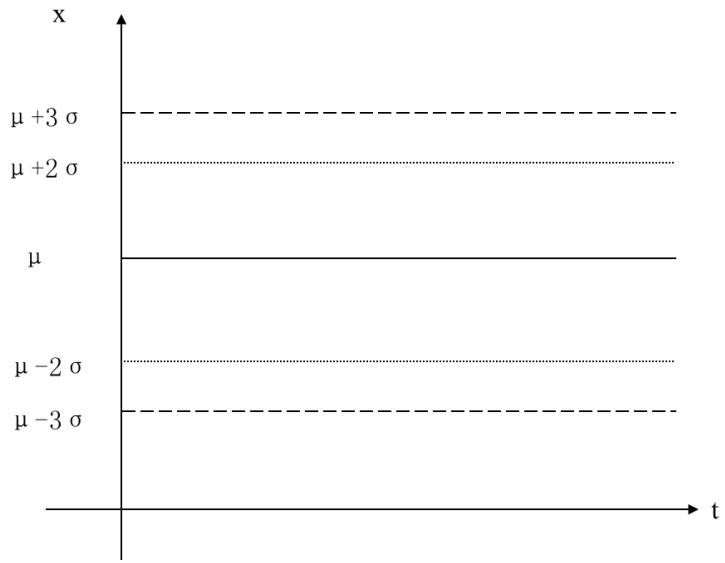
离群点诊断：给出n个数据点或对象的集合，及预期的离群点的数目k，发现与剩余的数据相比是显著差异的、异常的或不一致的前k个对象。

形成离群点的主要原因：
（1）首先可能是采样中的误差，如记录的偏误、工作人员出现笔误、计算错误等，都有可能产生极端大值或者极端小值。
（2）其次可能是被研究现象本身由于受各种偶然非正常的因素影响而引起的。

离群点诊断方法分类


经典的离群点挖掘方法
基于统计学或模型的方法
基于距离或邻近度的方法
基于密度的方法
基于聚类的方法
新的离群点挖掘算法
基于关联的方法
基于模糊集的方法
基于人工神经网络的方法
基于遗传算法或克隆选择的方法

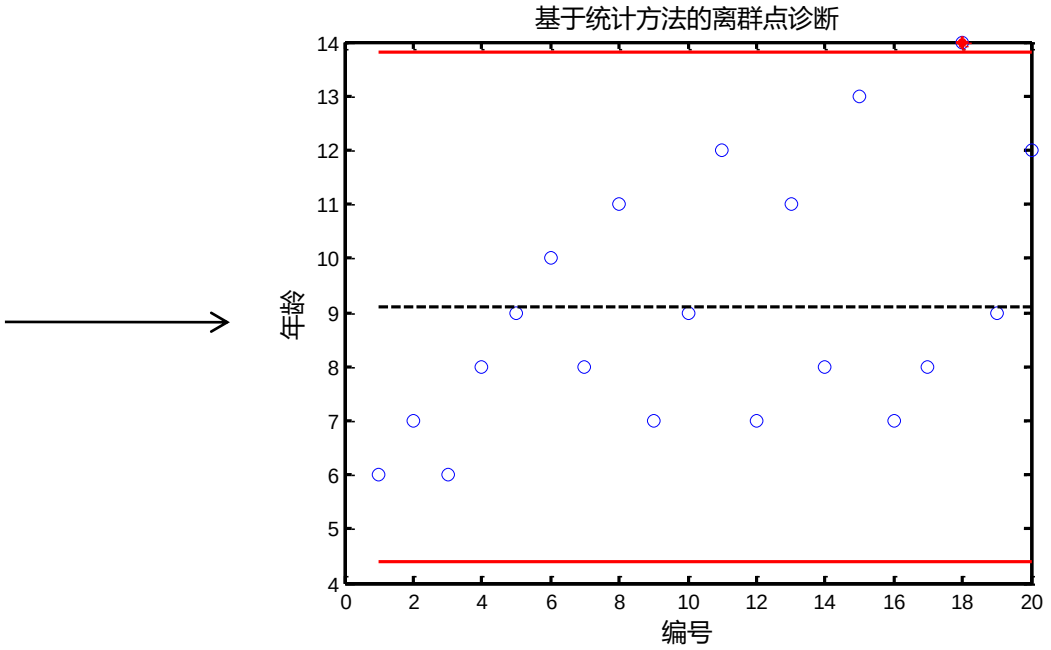
基于统计的离群点诊断：原理



（1）若此点在上、下警告线之间的区域内，则数据处于正常状态；
（2）若此点超出上、下警告线，但仍在上、下控制线之间的区域内，提示质量开始变劣，可能存在“离群”倾向；
（3）若此点落在上、下控制线之外，表示数据已经“离群"，这些点即被诊断出的离群点。

基于统计的离群点诊断：应用实例

20名儿童开始上学的年龄
={6,7,6,8,910,8,11,7,9,12,7,11,8,13,7,8,14,9,12}



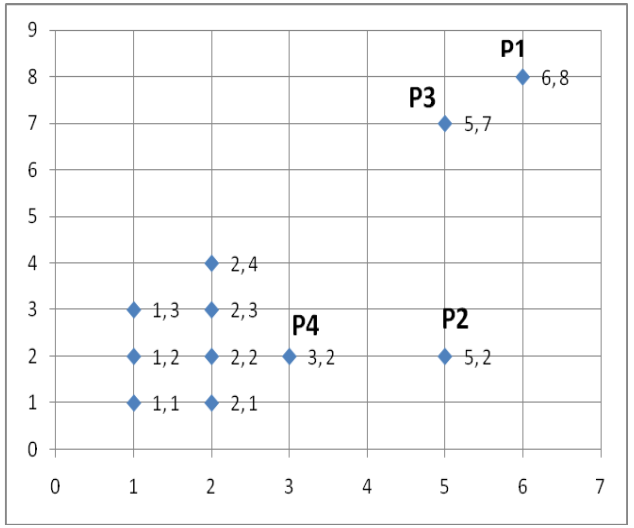
由上图可知，离群点为14。

基于距离的离群点诊断：原理

基于距离的离群点检测方法的基本思想是如果某个对象远离大部分其他对象，那么该对象是离群的。

基于距离方法的两种不同策略：
（1）采用给定邻域半径，依据点的邻域中包含的对象多少来判定离群点。如果一个点的邻域内包含的对象少于整个数据集的一定比例则标识它为离群点，也就是将没有足够邻居的对象看成是基于距离的离群点。
（2）利用k最近邻距离的大小来判定离群。使用k-最近邻的距离度量一个对象是否远离大部分点，一个对象的离群程度由到它的k-最近邻的距离给定。这种方法对k的取值比较敏感。k太小(例如1)，则少量的邻近离群点可能导致较低的离群程度。k太大，则点数少于k的簇中所有的对象可能都成了离群点。

基于距离的离群点诊断：应用实例



对P1点进行分析。k=2；最近邻的点为P3、P2，distance(P1，P2)与distance(P1，P3)分别为6.08、1.41，平均距离为：

$$OF1(P1,k)=\frac{dis\tan ce(P1,P2)+dis\tan ce(P1,P3)}{2}=\frac{6.08+1.41}{2}=3.745$$

对P2点进行分析。k=2；最近邻的点为P3、P4，同理有：

$$OF1(P2,k)=\frac{dis\tan ce(P2,P3)+dis\tan ce(P2,P4)}{2}=\frac{5+2}{2}=3.5$$

因为OF1(P1,K)> OF1(P2,K),因此，P1点更有可能是离群点。

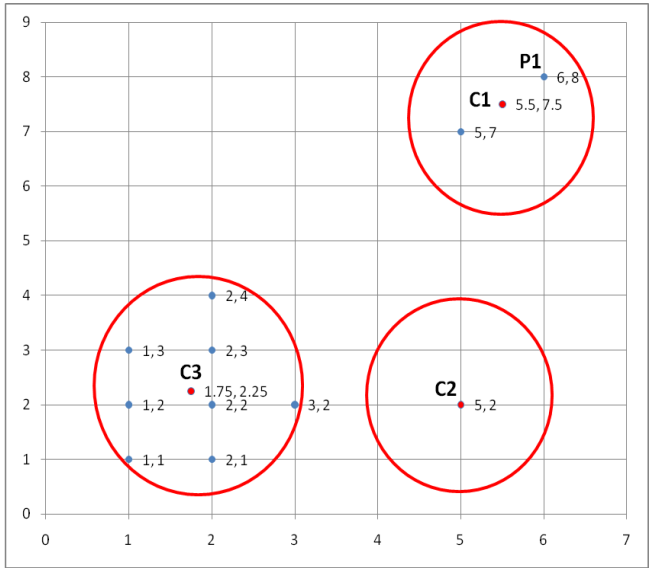
基于聚类的离群点诊断：原理

聚类分析是用来发现数据集中强相关的对象组，而离群点诊断是发现不与其他对象组强相关的对象。因此，离群点诊断和聚类是两个相对立的过程。如果聚类的结果中，某个簇的点比较少，且中心距离距离其他簇又比较远，则该簇中的点是离群点的可能性就比较大，所以从这个角度将聚类方法用于离群点诊断也是很自然的想法。

定义：假设据集*D*被聚类算法划分为*k*个簇 $C=\{C_1,C_2,\cdots,C_k\}$，对象*p*的离群因子(Outlier Factor)OP3(p)定义为*p*与所有簇间距离的加权平均值：

$$OF3(p)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(p,C_j)$$

基于聚类的离群点诊断：应用实例



求得所有对象的离群因子见下表：

x	Y	*OF3*
1	2	2.2
1	3	2.3
1	1	2.9
2	1	2.6
2	2	1.7
2	3	1.9
6	8	5.9
2	4	2.5
3	2	2.2
5	7	4.8
5	2	3.4

所有点的离群因子平均值*Ave_OF*=2.95，标准差*Dev_OF*=1.3，假设$\beta=1$；则阈值：

E=*Ave_OF* + **Dev_OF*=2.95+1.3=4.25*

离群因子大于4.25的对象可视为离群点，P1与P2都是离群点，但相对而言，P1更有可能成为离群点。

对于P1有：

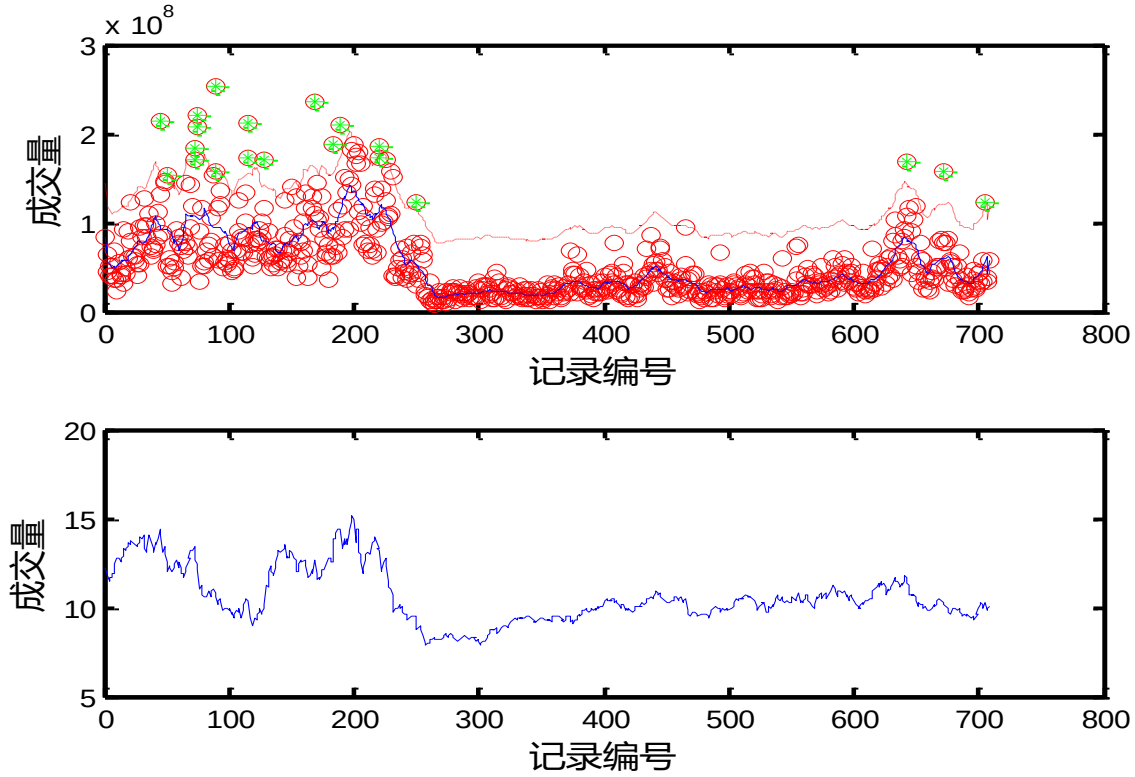
$$OF3(p_1)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(p_1,C_j)=\frac{8}{11}\sqrt{(6-1.75)^2+(8-2.25)^2}+\frac{1}{11}\sqrt{(6-5)^2+(8-2)^2}+\frac{3}{11}\sqrt{(6-5.5)^2+(8-7.5)^2}=5.9$$

对于P2有：

$$OF3(p_2)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\cdot d(p_2,C_j)=\frac{8}{11}\sqrt{(5-1.75)^2+(2-2.25)^2}+\frac{1}{11}\sqrt{(5-5)^2+(2-2)^2}+\frac{3}{11}\sqrt{(5-5.5)^2+(2-7.5)^2}=3.4$$

可见，点P1较P2更可能成为离群点。

异常诊断应用实例：股票买卖择时



上图为对一支股票的成交量进行离群点诊断而得到结果，图中虚线上面的点就是发现的异常点，所用的诊断技术为基于统计的技术。在实践中，当发现离群点后还要进一步判断是买入机会还是卖出机会，这时的判断更直接，也更有意义。

MATLAB 学习资源

- **www.mathworks.com**
  - 录制的讲座
  - 行业解决方案
  - MATLAB central
- **www.ilovematlab.cn**
  - 问题交流
  - 图书

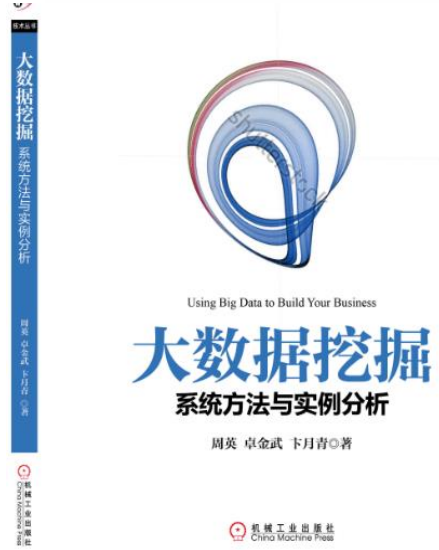
    《大数据挖掘：系统方法与实例分析》

- 购买正版**MATLAB**

  电话：010-59827000

- 答疑方式

  邮箱： 70263215@qq.com



关注MATLAB微信公众号，获取更多官方资讯！

关注MATLAB官方微信平台，发送你感兴趣的关键词，即可查看**MathWorks在线资源**。

MATLAB

实践与资源

第1讲：数据与程序 http://pan.baidu.com/s/1boGzSwn

第2讲：数据与程序 http://pan.baidu.com/s/1dELf87f

第3讲：数据与程序 http://pan.baidu.com/s/1c1Fcu5M

第4讲：数据与程序 http://pan.baidu.com/s/1jIbll2Y

谢谢大家！