Overview 概述

Instructor: Zach Zhizhong ZHOU, Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

数据挖掘中的关键想法

□Classification 分类

□Prediction 预测

□Association Rules 关联规则

□Data Reduction 数据减缩

□Data Exploration 数据探索

□Visualization 数据可视化

Supervised Learning 有监督学习

□Goal: Predict a single “target” or “outcome” variable
目标是预测一个结果变量

□Training data, where target value is known
使用结果变量已知的数据进行训练

□Score to data where value is not known
结果变量未知的数据组：对结果变量打分

□Methods: Classification and Prediction
使用的方法：分类和预测

Unsupervised Learning 无监督学习

□Goal: Segment data into meaningful segments; detect patterns
目标是把数据分割成有意义的数据段，检查各数据段呈现出的模式

□There is no target (outcome) variable to predict or classify
没有结果变量可供预测或者分类

□Methods: Association rules, data reduction & exploration, visualization
方法：关联规则，数据减缩和探索，数据可视化

Supervised: Classification 有监督：分类

□Goal: Predict categorical target (outcome) variable
目标是预测类别（结果）变量

□Examples: Purchase/no purchase, fraud/no fraud, reditworthy/not creditworthy…
例子：购买/未购买，欺诈/无欺诈，资信好/资信差

□Each row is a case (customer, tax return, applicant)
每行都是一个实例（顾客、退税、信贷申请）

□Each column is a variable
每列都是一个变量

□Target variable is often binary (yes/no)
目标变量通常是二元的（是/否）

Supervised: Prediction 有监督：预测

□Goal: Predict numerical target (outcome) variable
目标是预测一个数值型的结果变量
Examples: sales, revenue, performance
例子：销量、利润、绩效

□As in classification:
与“分类”一样：

□Each row is a case (customer, tax return, applicant)
每行都是一个实例（顾客、退税、信贷申请）

□Each column is a variable
每列都是一个变量

□Taken together, classification and prediction constitute “predictive analytics” 分类和预测一起构成预测性分析。

Unsupervised: Association Rules 无监督：关联规则

□Goal: Produce rules that define “what goes with what”
生成“某某与某某一起出现”的关联性规则
Example: “If X was purchased, Y was also purchased”
例子：如果消费者购买了X，那么他也会购买Y。

□Rows are transactions 每行是一个交易数据

□Used in recommender systems – “Our records show you bought X, you may also like Y” 用在推荐系统中：我们的记录显示你买了X，那么也许你会喜欢Y。

□Also called “affinity analysis” 也称亲和度分析。

Unsupervised: Data Reduction 无监督：数据减缩

□Distillation of complex/large data into simpler/smaller data 将复杂的或者庞大的数据提炼成简单的或者较小的数据。

□Reducing the number of variables/columns (e.g., principal components) 降低列数目或者变量个数（如主成分分析法）

□Reducing the number of records/rows (e.g., clustering) 降低行数目或者数据记录数目（如聚类）

Unsupervised: Data Visualization 无监督：数据可视化

- Graphs and plots of data 数据图形化

- Histograms, boxplots, bar charts, scatterplots 图包括：直方图、箱线图、条形图、散点图

- Especially useful to examine relationships between pairs of variables 对于检查一对数据之间的关系尤为有用

Data Exploration 数据探索

- □Data sets are typically large, complex & messy 数据集通常是庞大、复杂和混乱的

- □Need to review the data to help refine the task 需要检视数据以便精炼任务

- □Use techniques of Reduction and Visualization 使用的技术有数据减缩和数据可视化

The Process of Data Mining
数据挖掘的过程

Steps in Data Mining 数据挖掘的步骤

1.Define/understand purpose 定义/理解目标

2.Obtain data (may involve random sampling) 获得数据（可能需要进行随机抽样）

3.Explore, clean, pre-process data 探索、清理和预处理数据

4.Reduce the data; if supervised DM, partition it 数据减缩；如果是有监督数据挖掘则对数据进行分块处理

5.Specify task (classification, clustering, etc.) 明确任务（分类，聚类等）

6.Choose the techniques (regression, CART, neural networks, etc.) 选择数据挖掘技术（回归，决策树，人工神经网等）

7.Iterative implementation and “tuning” 反复应用数据挖掘技术并调优（调整并优化）。

8.Assess results – compare models 评估结果 比较不同模型

9.Deploy best model 采用最佳模型

Obtaining Data: Sampling 获取数据：抽样

□Data mining typically deals with huge databases 数据挖掘通常要处理巨大的数据库。

□Algorithms and models are typically applied to a sample from a database, to produce statistically-valid results 算法和模型通常用在从数据库中抽样出来的数据上，以得到统计上有效的结果。

□Once you develop and select a final model, you use it to “score” the observations in the larger database 当你建立和选择了最后的模型，你可以用该模型来对大数据库中每条观测值进行预测打分。

Rare event oversampling 罕见事件的过度抽样

□Often the event of interest is rare 我们感兴趣的事件经常是罕见的

□Examples: response to mailing, fraud in taxes, … 例子：回应邮件，报税中的欺诈

□Sampling may yield too few “interesting” cases to effectively train a model 普通的抽样可能导致抽样的数据集包含太少的我们感兴趣的实例，这样就无法有效训练一个模型。

□A popular solution: oversample the rare cases to obtain a more balanced training set 通常的解决方法：过度抽样罕见实例

□Later, need to adjust results for the oversampling 最后需要为过度抽样调整结果

Pre-processing Data
数据预处理

Types of Variables 数据类型

□ Determine the types of pre-processing needed, and algorithms used 数据类型决定需要预处理的类型和需要用到的算法

□ Main distinction: Categorical vs. numeric 类别型数据和数值型数据

□ Numeric 数字型

❿ Continuous 连续的实数

❿ Integer 整数

□ Categorical 类别型

❿ Ordered (low, medium, high) 排序（低 中 高）

❿ Unordered (male, female) 未排序（男性 女性）

Variable handling 处理数据

□Numeric 数值型数据

- ➉ Most algorithms can handle numeric data 绝大多数算法能直接处理数值型数据
- ➉ May occasionally need to “bin” into categories 有时可能需要把它们“打包”放入不同类别中

□Categorical 类别型数据

- ➉ Naïve Bayes can use as-is 朴素贝叶斯分类算法可以直接使用类别型数据
- ➉ In most other algorithms, must create binary dummies (number of dummies = number of categories – 1) 绝大多数的其他算法中，需要为类别型数据创建取值为0或者1的虚拟变量（虚拟变量个数等于类别个数减去1）

Detecting Outliers 检查是否有离群点

□An outlier is an observation that is “extreme”, being distant from the rest of the data (definition of “distant” is deliberately vague) 离群点是观测值远离其他数值（“远离”是一个非常模糊的概念）

□Outliers can have disproportionate influence on models (a problem if it is spurious) 离群点可能对模型造成不成比例的影响（当该离群点是虚假数据时就有问题）

□An important step in data pre-processing is detecting outliers 数据预处理需要检查并找出这些离群点

□Once detected, domain knowledge is required to determine if it is an error, or truly extreme. 找到之后需要用领域知识判断该离群点是一个错误还是真的就是远离其他数据的点。

Detecting Outliers 检查是否有离群点

□In some contexts, finding outliers is the purpose of the DM exercise (airport security screening). This is called “anomaly detection”.

□在一些情况下，找到离群点本身就是数据挖掘实践的目的（机场安检）。这称之为“异常检测”。

Handling Missing Data 处理缺失数据

□ Most algorithms will not process records with missing values. Default is to drop those records. 绝大多数算法不会处理含有缺失数据的记录，默认做法是丢弃这些记录。

□ Solution 1: Omission 解决方案1：忽略

⑩ If a small number of records have missing values, can omit them 如果少量记录有缺失数据，可以忽略它们。

⑩ If many records are missing values on a small set of variables, can drop those variables (or use proxies) 如果数量较少的数据记录当中少量变量有缺失数据，可以忽略这些变量（或者使用代理）

⑩ If many records have missing values, omission is not practical 如果许多记录有缺失数据，那么忽略它们就是不现实的。

□ Solution 2: Imputation 解决方案2：填充

⑩ Replace missing values with reasonable substitutes 用合理的数值取代缺失的数据

⑩ Lets you keep the record and use the rest of its (non-missing) information 这样就允许你保留下记录并使用余下的未缺失的信息

Normalizing (Standardizing) Data 标准化数据

□ Used in some techniques when variables with the largest scales would dominate and skew results 当最大尺寸的变量主导并歪曲结果时使用数据标准化

□ Puts all variables on same scale 使所有变量有相同的尺寸量纲

□ Normalizing function: Subtract mean and divide by standard deviation 函数：数值减去平均值然后除以标准方差

□ Alternative function: scale to 0-1 by subtracting minimum and dividing by the range 可选函数：减去最小值然后除以数值分布范围

- ⑩ Useful when the data contain dummies and numeric 在数据中含有虚拟变量与数值变量时有用

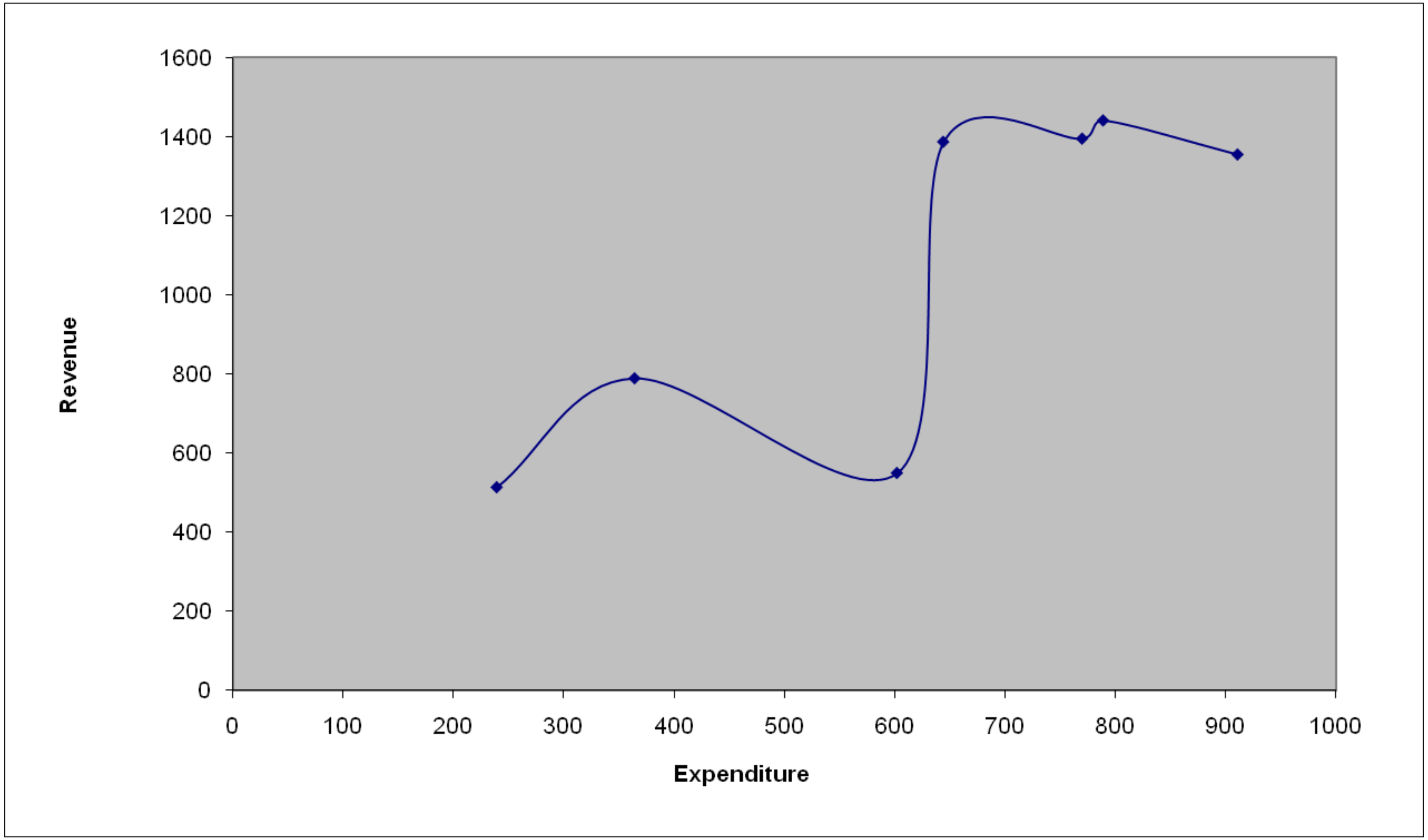
The Problem of Overfitting 过度拟合造成的问题

□ Statistical models can produce highly complex explanations of relationships between variables 统计模型可以生成高度复杂的对变量关系的解释。

□ The “fit” may be excellent 拟合结果可能是非常好的

□ When used with new data, models of great complexity do not do so well. 但是高度复杂的模型用来解释新的数据可能并不是非常好。

100% fit – not useful for new data



Overfitting (cont.) 过度拟合

Causes: 原因：

❿ Too many predictors 太多预测变量

❿ A model with too many parameters 太多参数

❿ Trying many different models 尝试了许多不同的模型

Consequence: Deployed model will not work as well as expected with completely new data. 过度拟合的模型用在新数据上表现将不如预期的好。

Partitioning the Data 数据分区

Problem: How well will our model perform with new data? 我们的模型在新数据下表现有多好?

Solution: Separate data into two parts

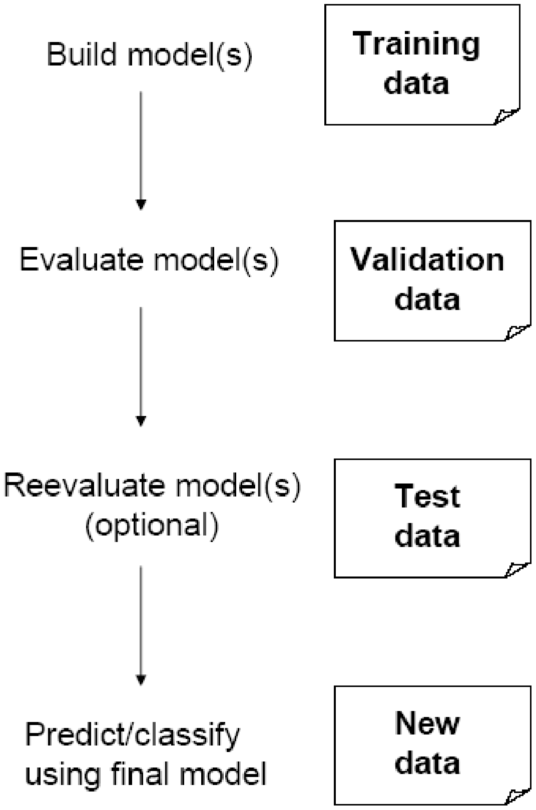
将数据切割成2块

⑩Training partition to develop the model

其中一块训练数据子集用于建模

⑩Validation partition to implement the model and evaluate its performance on “new” data

另外一块用于验证模型的数据子集用来评估模型在新数据中表现如何。



Build model(s)

Training data

Evaluate model(s)

Validation data

Reevaluate model(s) (optional)

Test data

Predict/classify using final model

New data

Addresses the issue of overfitting 这种方法用来解决过度拟合的问题。

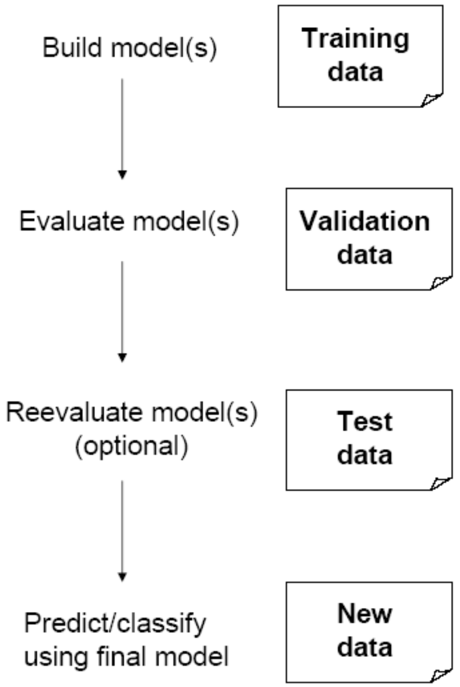
Test Partition 测试数据分区

- □ When a model is developed on training data, it can overfit the training data (hence need to assess on validation) 使用训练数据子集建模可能导致模型过度拟合训练数据子集（所以需要验证模型）

- □ Assessing multiple models on same validation data can overfit validation data 在验证数据子集上验证不同模型可能导致过度拟合验证数据。

- □ Some methods use the validation data to choose a parameter. This too can lead to overfitting the validation data 一些数据挖掘方法使用验证数据子集确定参数，这也可能导致过度拟合验证数据

- □ Solution: final selected model is applied to a test partition to give unbiased estimate of its performance on new data 解决方案：最终选择的模型用在一个测试数据分区上，可以评估该模型在新数据上的表现情况。



Build model(s)

Training data

Evaluate model(s)

Validation data

Reevaluate model(s) (optional)

Test data

Predict/classify using final model

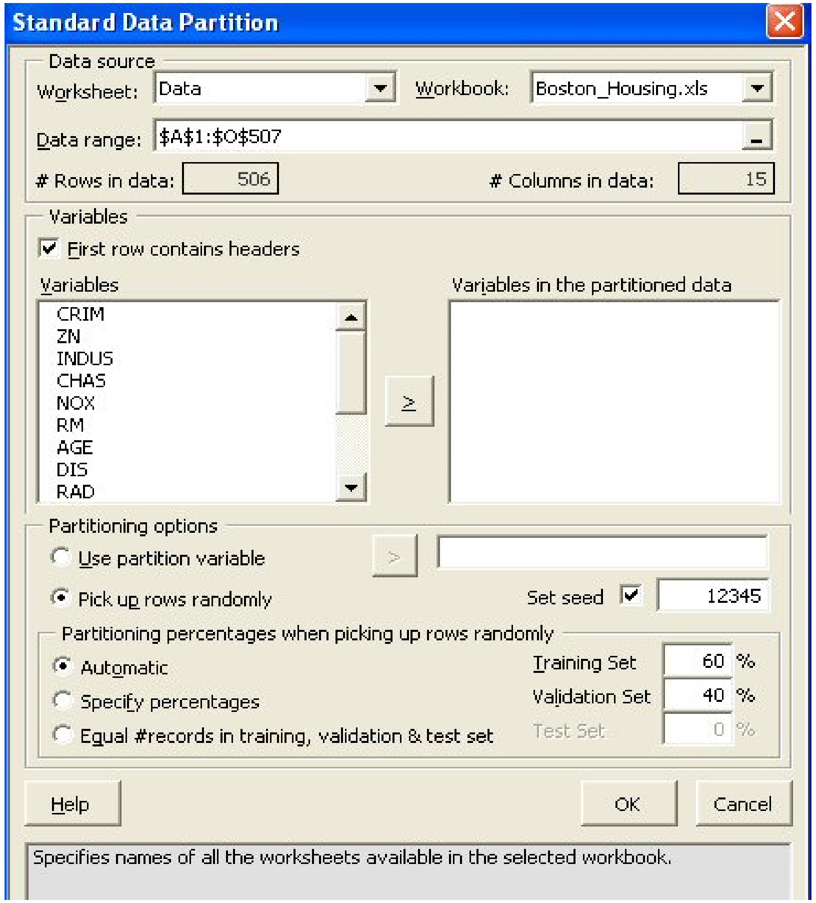
New data

Example – Linear Regression Boston Housing Data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000

Partitioning the data



Standard Data Partition

Data source

Worksheet: Data

Workbook: Boston_Housing.xls

Data range: $A$1:$O$507

Rows in data: 506

Columns in data: 15

Variables

☑ First row contains headers

Variables

CRIM
ZN
INDUS
CHAS
NOX
RM
AGE
DIS
RAD

Variables in the partitioned data

Partitioning options

○ Use partition variable

● Pick up rows randomly

Set seed ☑ 12345

Partitioning percentages when picking up rows randomly

● Automatic

○ Specify percentages

○ Equal #records in training, validation & test set

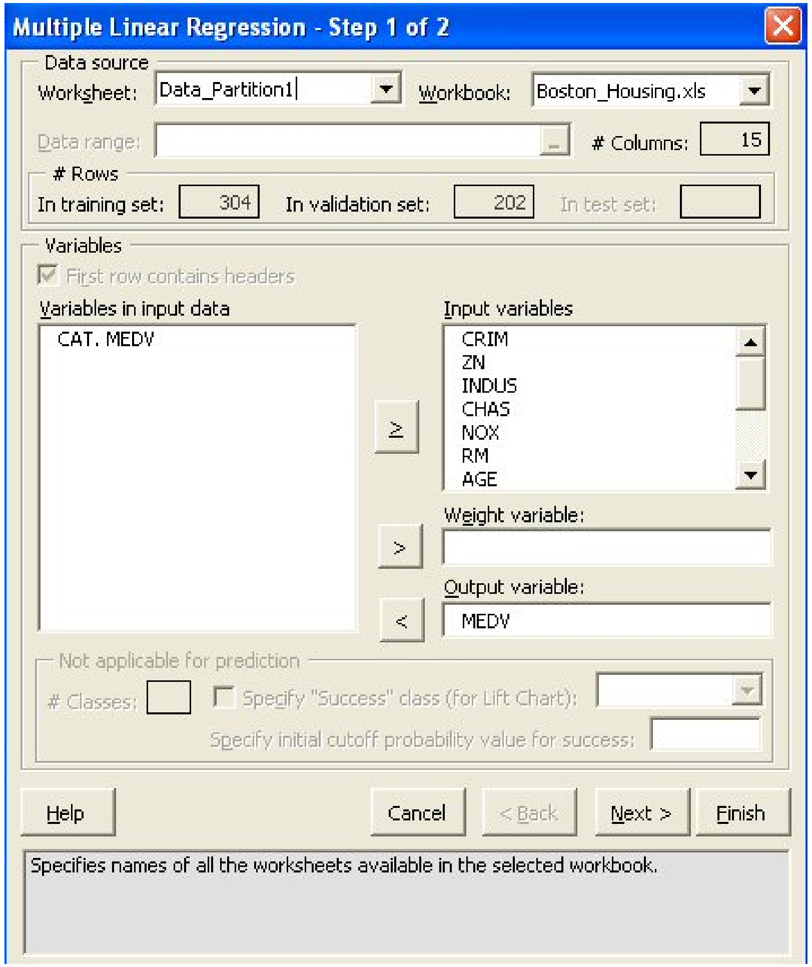
Training Set 60 %

Validation Set 40 %

Test Set 0 %

Specifies names of all the worksheets available in the selected workbook.

Multiple Linear Regression



Data source

Worksheet: Data_Partition1

Workbook: Boston_Housing.xls

Data range:

# Columns: 15

# Rows

In training set: 304

In validation set: 202

In test set:

Variables

☑ First row contains headers

Variables in input data

CAT. MEDV

Input variables

CRIM
ZN
INDUS
CHAS
NOX
RM
AGE

Weight variable:

Output variable: MEDV

Not applicable for prediction

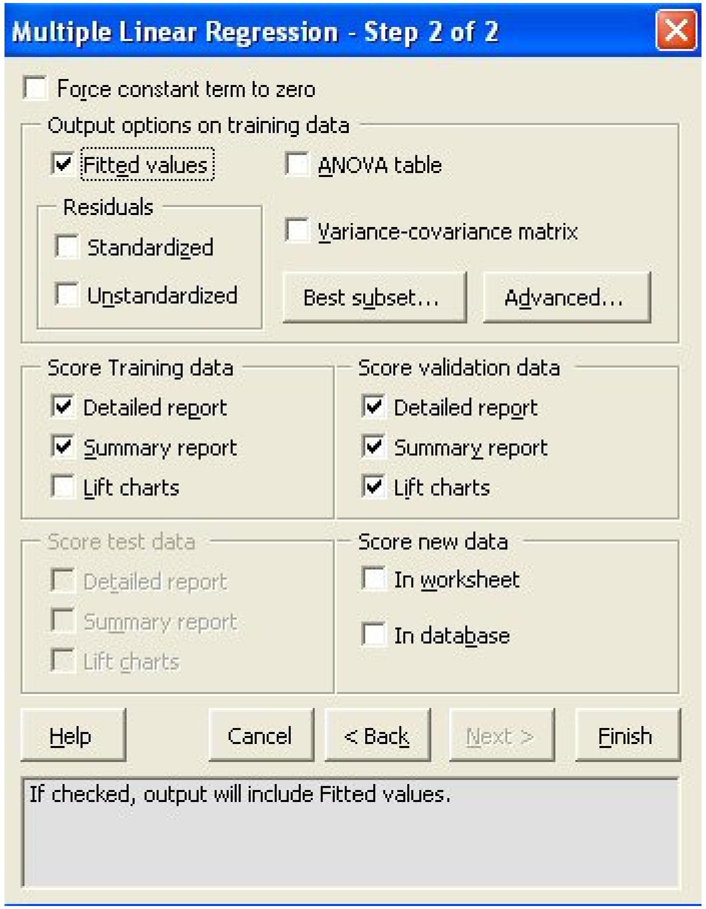
# Classes:

□ Specify "Success" class (for Lift Chart):

Specify initial cutoff probability value for success:

Specifies names of all the worksheets available in the selected workbook.

Specifying Output



□ Force constant term to zero

Output options on training data

☑ Fitted values

□ ANOVA table

Residuals

□ Standardized

□ Unstandardized

□ Variance-covariance matrix

Score Training data

☑ Detailed report

☑ Summary report

□ Lift charts

Score validation data

☑ Detailed report

☑ Summary report

☑ Lift charts

Score test data

□ Detailed report

□ Summary report

□ Lift charts

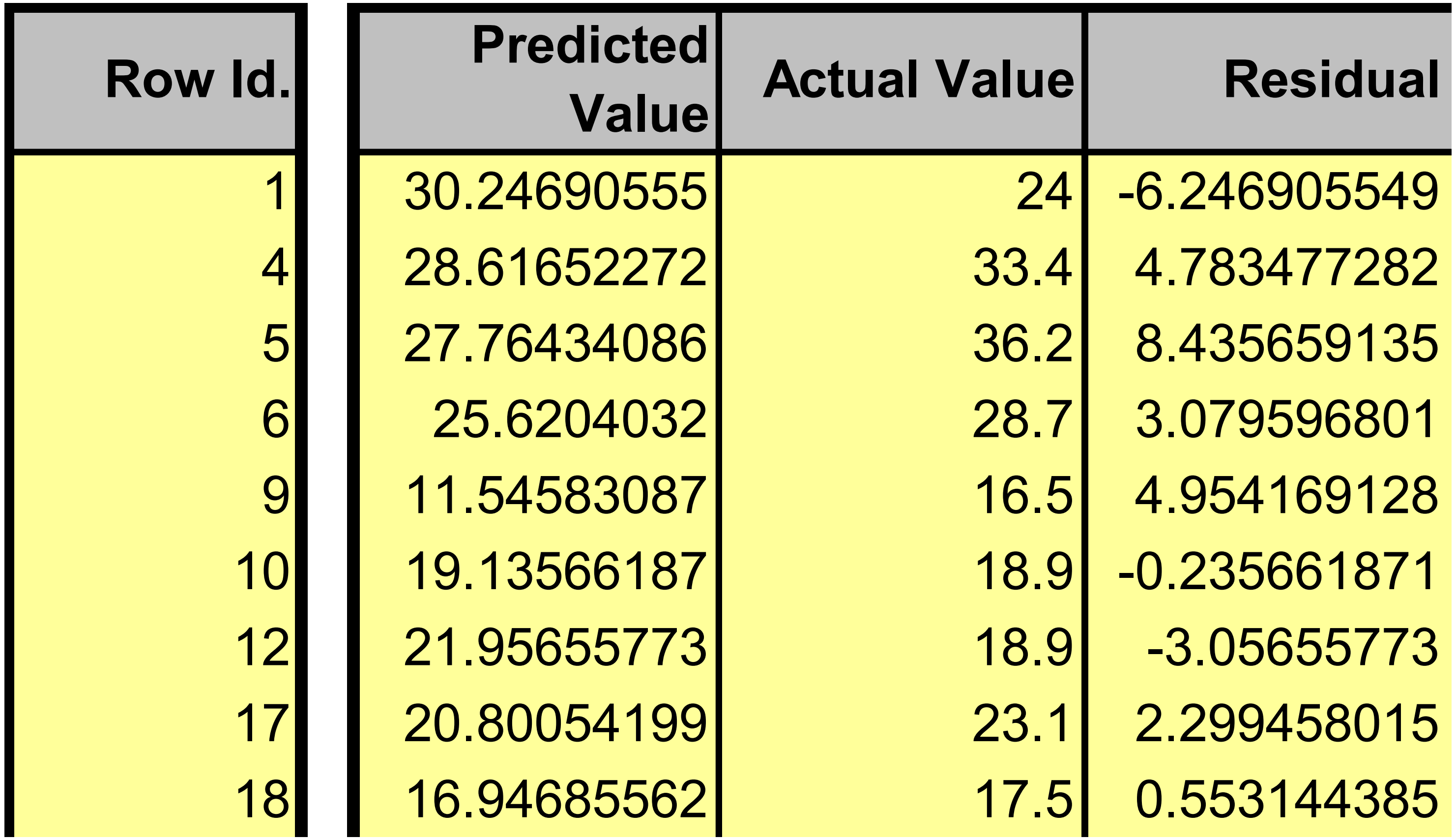
Score new data

□ In worksheet

□ In database

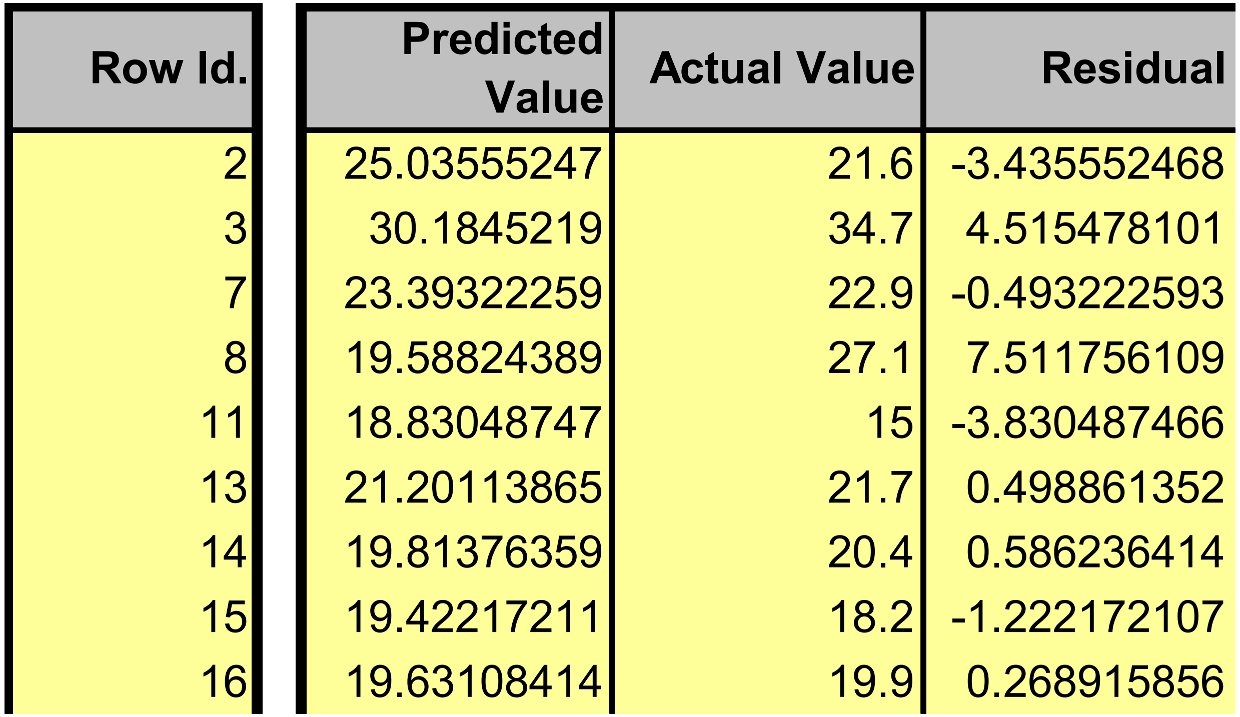
If checked, output will include Fitted values.

Prediction of Training Data



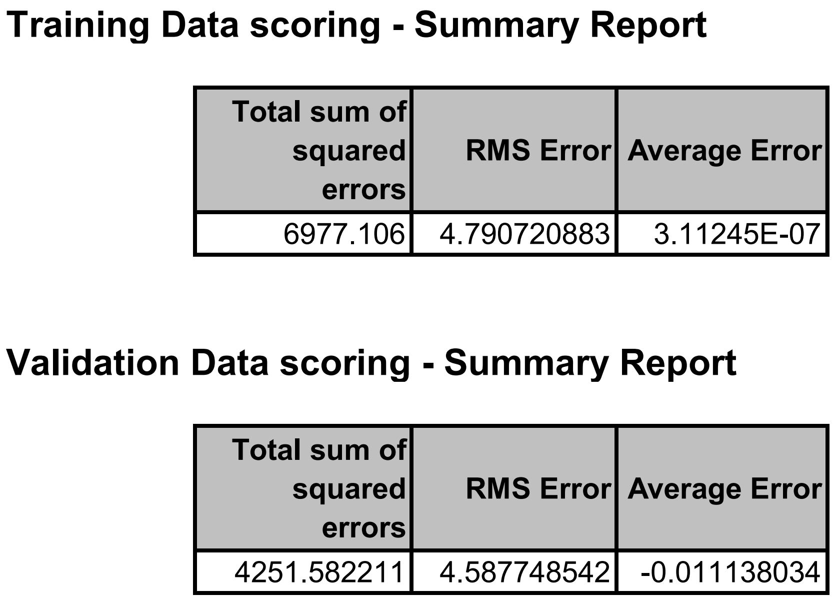
Row Id.	Predicted Value	Actual Value	Residual
1	30.24690555	24	-6.246905549
4	28.61652272	33.4	4.783477282
5	27.76434086	36.2	8.435659135
6	25.6204032	28.7	3.079596801
9	11.54583087	16.5	4.954169128
10	19.13566187	18.9	-0.235661871
12	21.95655773	18.9	-3.05655773
17	20.80054199	23.1	2.299458015
18	16.94685562	17.5	0.553144385

Prediction of Validation Data



Row Id.	Predicted Value	Actual Value	Residual
2	25.03555247	21.6	-3.435552468
3	30.1845219	34.7	4.515478101
7	23.39322259	22.9	-0.493222593
8	19.58824389	27.1	7.511756109
11	18.83048747	15	-3.830487466
13	21.20113865	21.7	0.498861352
14	19.81376359	20.4	0.586236414
15	19.42217211	18.2	-1.222172107
16	19.63108414	19.9	0.268915856

Summary of errors



Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6977.106	4.790720883	3.11245E-07

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4251.582211	4.587748542	-0.011138034

RMS error

Error = actual - predicted

RMS = Root-mean-squared error = Square root of average squared error

In previous example, sizes of training and validation sets differ, so only RMS Error and Average Error are comparable

Using Excel and XLMiner for Data Mining

□Excel is limited in data capacity

□However, the training and validation of DM models can be handled within the modest limits of Excel and XLMiner

□Models can then be used to score larger databases

□XLMiner has functions for interacting with various databases (taking samples from a database, and scoring a database from a developed model)

Summary

□Data Mining consists of supervised methods (Classification & Prediction) and unsupervised methods (Association Rules, Data Reduction, Data Exploration & Visualization) 数据挖掘包括有监督学习方法（分类和预测）和无监督学习方法（关联规则、数据减缩、数据探索和可视化）

□Before algorithms can be applied, data must be characterized and pre-processed 使用算法前，数据必须被特征化并被预处理

□To evaluate performance and to avoid overfitting, data partitioning is used 为了评估模型的表现以及避免过度拟合，应使用数据分区方法。

□Data mining methods are usually applied to a sample from a large database, and then the best model is used to score the entire database 数据挖掘方法通常用于大数据库的一个样本，最好的模型则用于整个数据库。