Data Mining

3.6 Regression Analysis

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- Straight-Line Linear Regression
- Multiple Linear Regression
- Other Regression Models
- References

Introduction

Introduction

- **Numerical prediction** is similar to classification
  - construct a model
  - use model to predict continuous or ordered value for a given input
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions
- Major method for prediction: **regression**
  - model the relationship between one or more independent or **predictor** variables and a dependent or **response** variable

Introduction

- In the context of data mining
  - the predictor variables are the attributes of interest describing the tuple that are known.
  - The response variable is what we want to predict
- Many texts use the terms “regression” and “numeric prediction”
- Some classification techniques can be adapted for prediction.
  - e.g. backpropagation, support vector machines, and *k-nearest-neighbor classifiers*

Introduction

- Regression analysis is a good choice when all of the predictor variables are continuous valued as well.
- **Regression analysis methods**
  - Linear regression
    - Straight-line linear regression
    - Multiple linear regression
  - Non-linear regression
  - Generalized linear model
    - Poisson regression
    - Logistic regression
  - Log-linear models
  - Regression trees and Model trees

Straight-Line Linear Regression

Linear Regression

- **Straight-line linear regression**:
  - involves a response variable y and a single predictor variable x

    $y = w_0 + w_1 x$

  where w_0 (y-intercept) and w_1 (slope) are **regression coefficients**

Linear regression

- **Method of least squares**: estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

- *D:* a training set
- *x*: consisting of values of predictor variable
- *y*: response variable
- */D/* : data points of the form($x1$, $y1$), ($x2$, $y2$),…, ($x/D/$, $y/D/$).
- $\bar{x}$ *:* the mean value of $x1$, $x2$, … , $x/D/$
- $\bar{y}$ *:* the mean value of $y1$, $y2$, … , $y/D/$

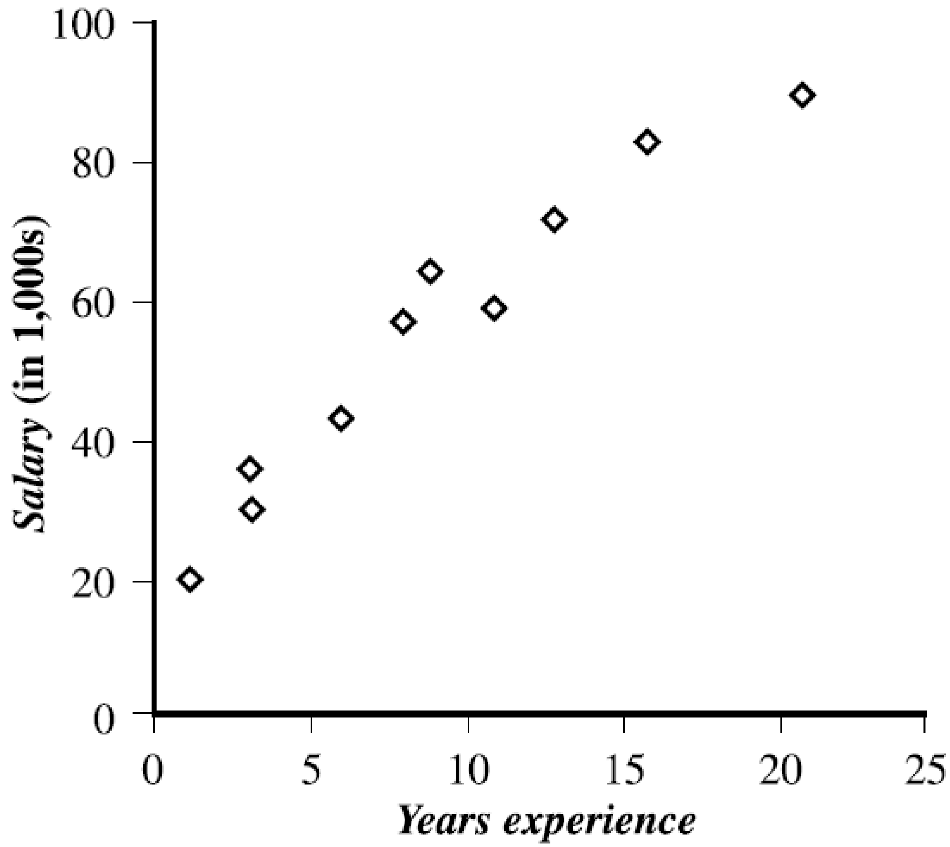
Example: Salary problem

- The table shows a set of paired data where x is the number of years of work experience of a college graduate and y is the corresponding salary of the graduate.

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

Linear Regression

- The 2-D data can be graphed on a *scatter plot.*
- The plot suggests a linear relationship between the two variables, *x and y.*



Example: Salary data

- Given the above data, we compute

$$\bar{x} = 9.1 \text{ and } \bar{y} = 55.4$$

- we get

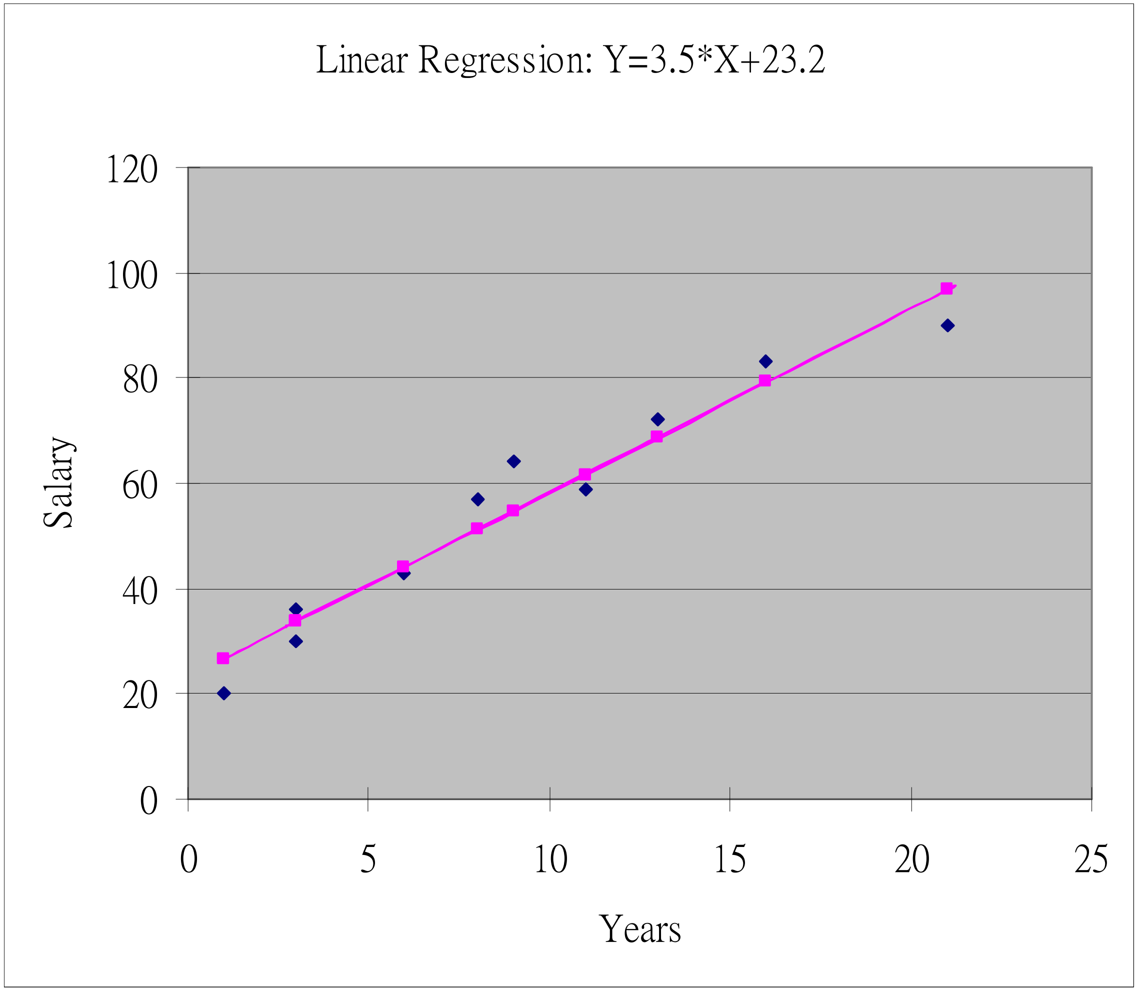
$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5$$

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

- the equation of the least squares line is estimated by

$$y = 23.6 + 3.5x$$

Example: Salary data



Multiple Linear Regression

Multiple linear regression

- Multiple linear regression involves more than one predictor variable
- Training data is of the form ($\mathbf{X}_1$, y_1), ($\mathbf{X}_2$, y_2),…, ($\mathbf{X}_{|D|}$, $y_{|D|}$)
- where the $\boldsymbol{X_i}$ are the n-dimensional training data with associated class labels, y_i
- An example of a multiple linear regression model based on two predictor attributes:

$$y = w_0 + w_1 x_1 + w_2 x_2$$

Example: CPU performance data

	Cycle time (ns) MYCT	Main memory (KB) Min. MMIN	Main memory (KB) Max. MMAX	Cache (KB) CACH	Channels Min. CHMIN	Channels Max. CHMAX	Performance PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
...							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

$$\begin{aligned}\text{PRP} = &-55.9 + 0.0489\ \text{MYCT} + 0.0153\ \text{MMIN} + 0.0056\ \text{MMAX} \\ &+ 0.6410\ \text{CACH} - 0.2700\ \text{CHMIN} + 1.480\ \text{CHMAX}.\end{aligned}$$

Multiple Linear Regression

- Various statistical measures exist for determining how well the proposed model can predict *y.* (described later)
- Obviously, the greater the number of predictor attributes is, the slower the performance is.
- Before applying regression analysis, it is common to perform attribute subset selection to eliminate attributes that are unlikely to be good predictors for *y.*
- In general, regression analysis is accurate for prediction, except when the data contain outliers.

Other Regression Models

Nonlinear Regression

- Sometimes we can get a more accurate model using a nonlinear model, For example,

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

- Some nonlinear models can be transformed into linear regression model. For example, the above function can be converted to linear with new variables: $x_2 = x^2$, $x_3 = x^3$

$$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$$

Generalized linear models

- **Generalized linear model** is foundation on which linear regression can be applied to modeling categorical response variables
- Common types of generalized linear models include
  - **Logistic regression**: Models the probability of some event occurring as a linear function of a set of predictor variables.
  - **Poisson regression**: models the data that exhibit a Poisson distribution

Log-linear models

- In the log-linear method, all attributes must be categorical
- Continuous-valued attributes must first be discretized.

Regression Trees and Model Trees

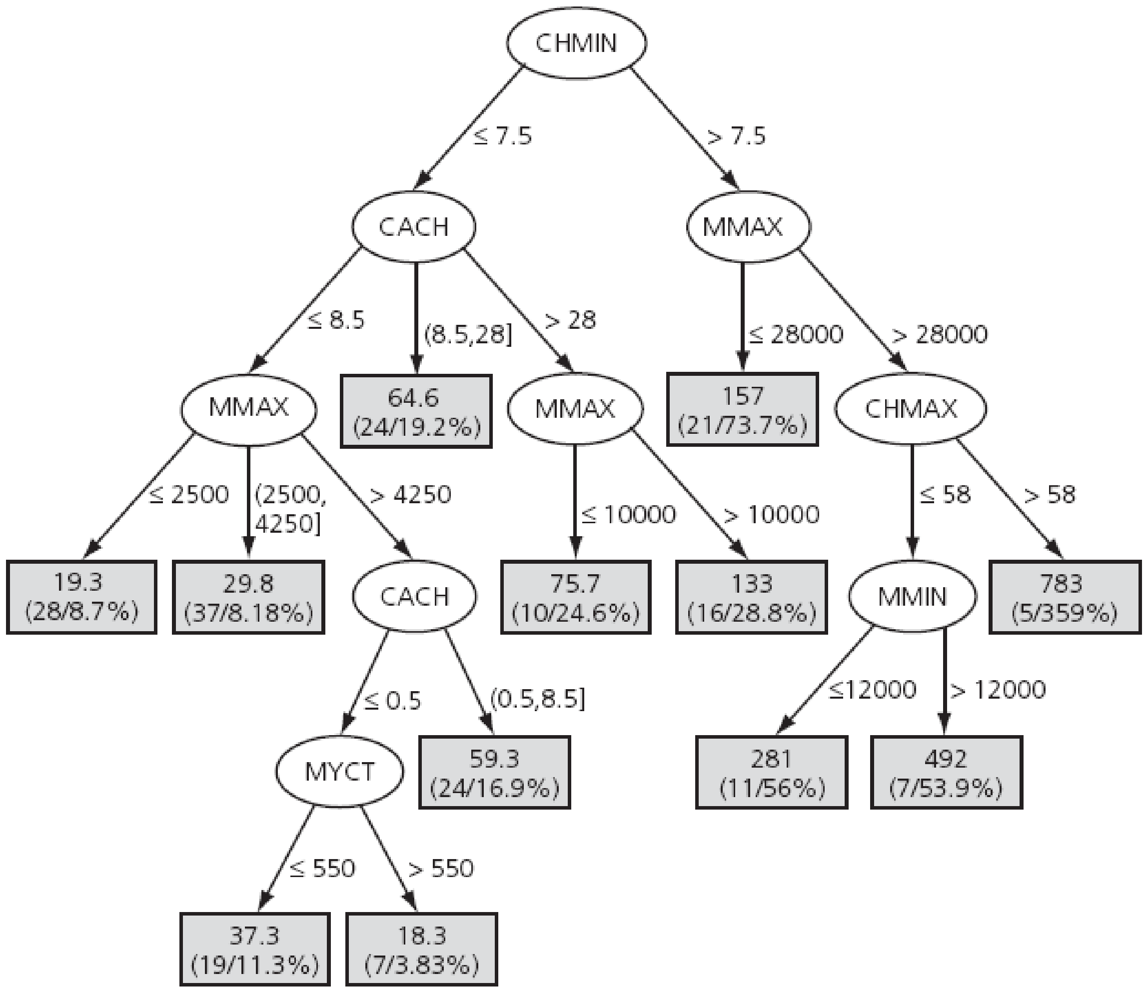
- Trees to predict continuous values rather than class labels
- **Regression and model trees** tend to be more accurate than linear regression when the data are not represented well by a simple linear model

Regression trees

- **Regression tree**: a decision tree where each leaf predicts a numeric quantity
- Proposed in CART system (Breiman et al. 1984)
  - CART: Classification And Regression Trees
- Predicted value is average value of training instances that reach the leaf

Example: CPU performance problem

- Regression tree for the CPU data



Example: CPU performance problem

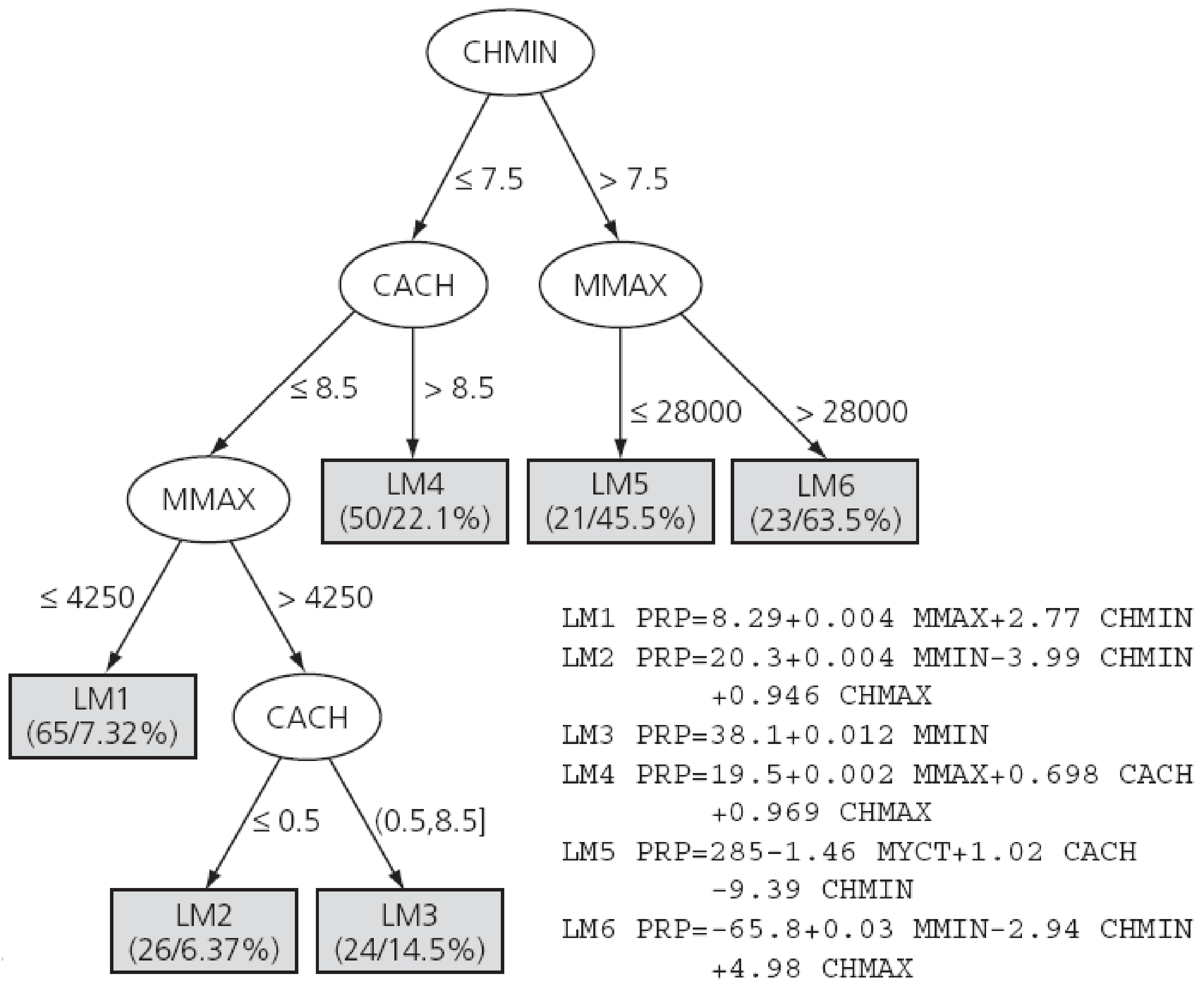
- We calculate the average of the absolute values of the errors between the predicted and the actual CPU performance measures
- It turns out to be significantly less for the tree than for the regression equation.

Model tree

- **Model tree**: Each leaf holds a regression model
- A multivariate linear equation for the predicted attribute
- Proposed by Quinlan (1992)
- A more general case than regression tree

Example: CPU performance problem

- Model tree for the CPU data



References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 6)

- I. H. Witten and E. Frank, **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd Edition, Elsevier Inc., 2005. (Chapter 6)

The end