Introduction to

商务智能

第5章 数值预测
Chapter 5 Numeric Prediction

Outline

- Basic concepts
- Typical techniques
- Model evaluation
- summary

What Is Numeric Prediction?

- Model continuous-valued functions, i.e., predicts unknown or missing values
- Prediction is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions

Classification vs. Numeric Prediction

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Independent variable
（自变量）

Dependent variable (因变量）

name	age	years	income
Mike	<=30	3	8120.5
Mary	<=30	2	6208
Bill	31…40	4	3060
Jim	>40	7	7050
Dave	>40	6	10300
Anne	31…40	7	20060
…	…	…	…

Continuous value

Prediction methods

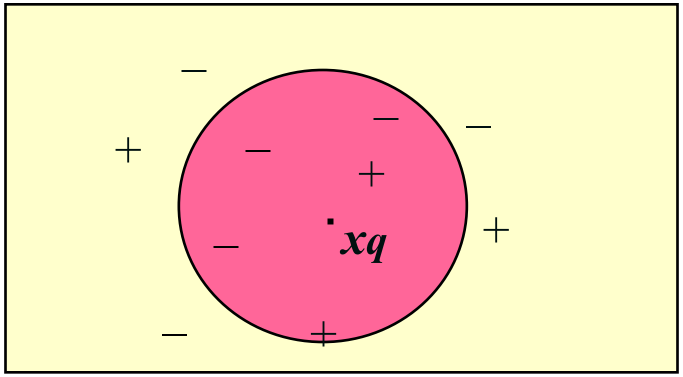
- Two kinds of methods
  - Lazy and Eager
- Major method for prediction: regression
  - Linear and multiple regression
  - Non-linear regression
  - Model tree, regression tree

K nearest Neighbors

The *k*-Nearest Neighbor Algorithm

- All instances correspond to points in the n-D space.
- The nearest neighbor are defined in terms of Euclidean distance.
- The target function could be discrete- or real- valued.
- For continuous-valued target functions， Calculate the mean values of the *k* nearest neighbors

$$v' = \frac{v - min_A}{max_A - min_A}$$



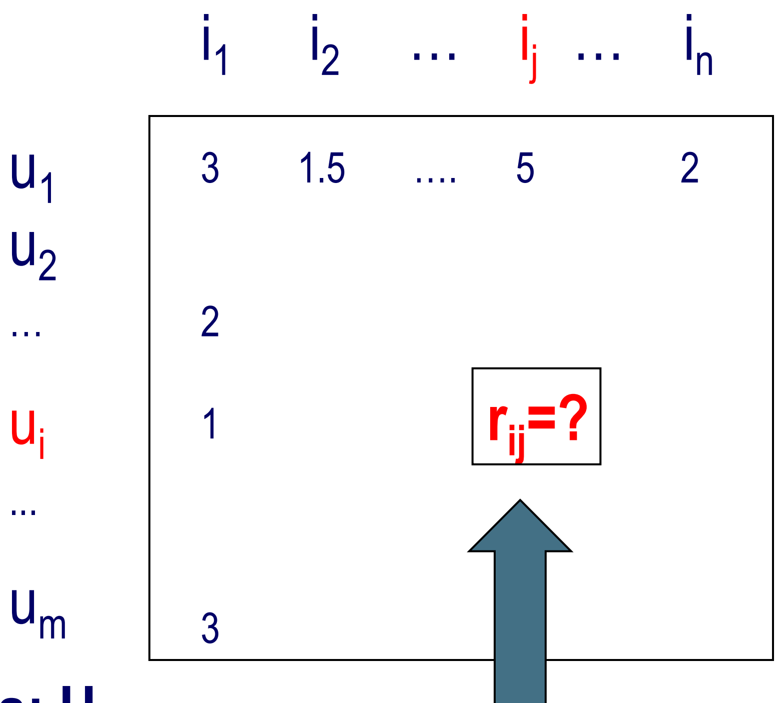
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

Application： Recommender Systems

- Collaborative Filtering (CF,协同过滤)
  - Look at users collective behavior
  - Look at the active user history
  - Combine!
- Content-based Filtering
  - Recommend items based on key-words

Collaborative Filtering: A Framework

Items: I



Users: U

Unknown function
$f: U \times I \rightarrow R$

The task:

Q1: Find Unknown ratings?

Q2: Which items should we recommend to this user?

.

.

.

Collaborative Filtering Road Map

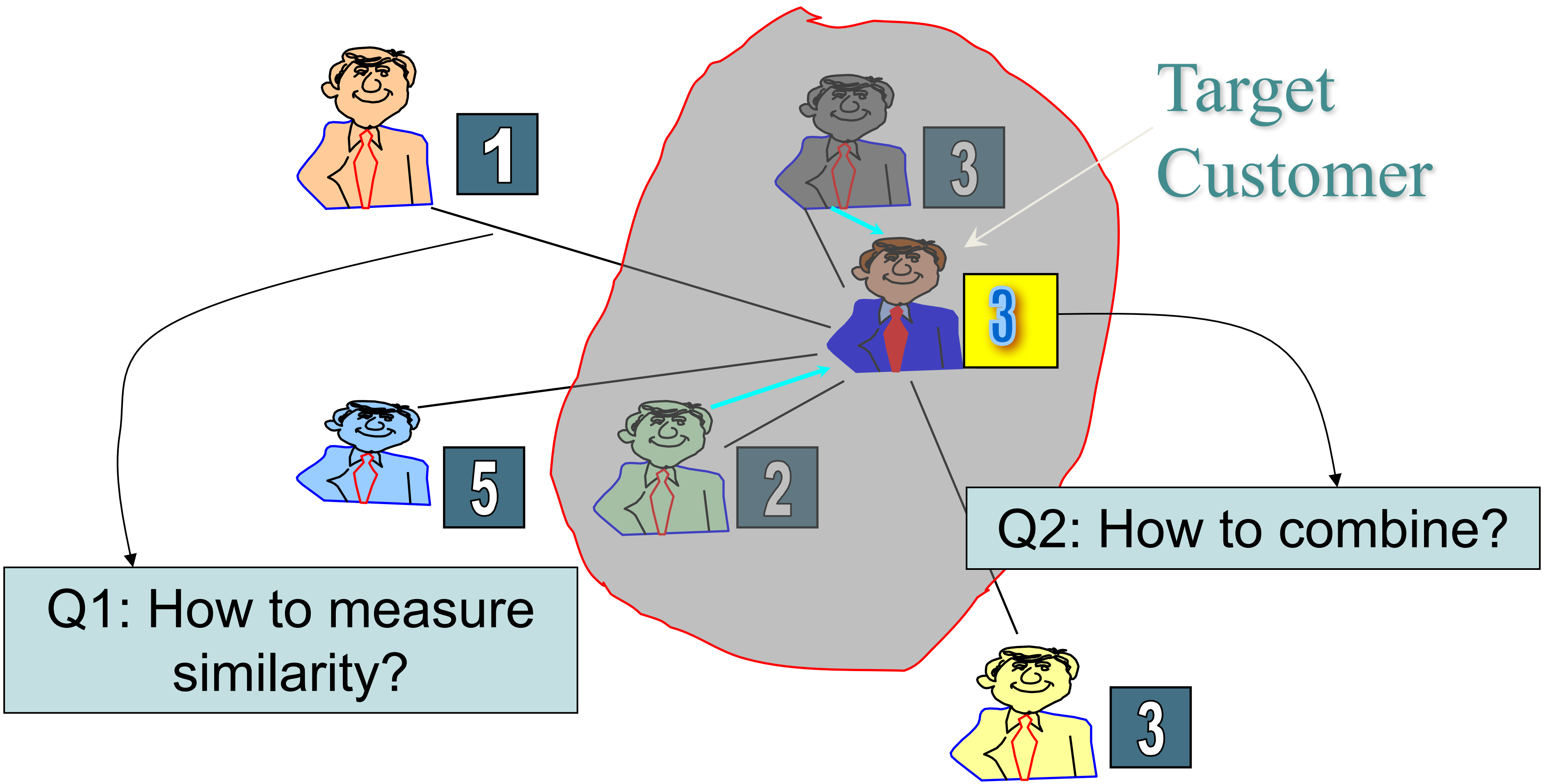
- User-User Methods
  - Identify like-minded users
  - Memory-based: KNN
- Item-Item Method
  - Identify buying patterns
  - Correlation Analysis

User-User Method

Intuition

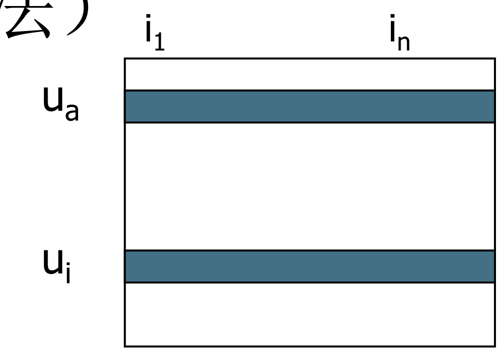
- Similar users have similar preferences
  - If $u \approx u'$, then for all o's, $f(u,o) \approx f(u',o)$
- User similarity (Zhang San vs. *Li Si*)
  - Suppose Zhang San and *Li Si* viewed similar movies in the past six months …
  - If Zhang San liked the paper, *Li Si* will like the paper

User-User Similarity: Intuition



How to Measure Similarity?

- Pearson correlation coefficient（相关系数法）



$$w_p(a,i) = \frac{\sum_{j \in \text{Commonly Rated Items}} (r_{aj} - \overline{r_a})(r_{ij} - \overline{r_i})}{\sqrt{\sum_{j \in \text{Commonly Rated Items}} (r_{aj} - \overline{r_a})^2 \sum_{j \in \text{Commonly Rated Items}} (r_{ij} - \overline{r_i})^2}}$$

$$\begin{array}{c|cccccc|}
 & i_1 & i_2 & \cdots & i_j & \cdots & i_n \\
\hline
\cdots & & & & & & \\
u_a & r_{a1} & r_{a2} & & \cdots & & r_{an} \\
\cdots & & & & & & \\
u_i & r_{i1} & r_{i2} & & \cdots & & r_{in} \\
\cdots & & & & & & \\
u_m & r_{m1} & r_{m2} & & \cdots & & r_{mn} \\
\hline
\end{array}$$

其中：用户i的偏好均值：

$$\overline{r_i} = \frac{1}{|I_i|} \sum_{k \in I_i} r_{ik}$$

I_i表示用户i的投票范围

How to predict?

- 用户a对项目j的预测偏好

i是a的邻居

$$r_{aj} = \overline{r_a} + \underbrace{\frac{\sum_i w(a,i)(r_{ij} - \overline{r_i})}{\sum_i w(a,i)}}_{\text{User a's estimated deviation}}$$

User i's deviation

User a's neutral

W(u1,u2)=0.997
W(u2,u4)=0.86

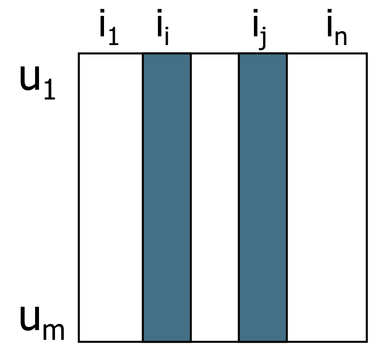
	i_1	i_2	i_3	i_4	i_5	i_6
u_1	1	2	2		3	
u_2	2	?	3	2	5	
u_3	3	4		5	1	2
u_4	2	3	4		2	1

$$r_{22} = \overline{r_2} + \frac{w(2,1)(r_{12} - \overline{r_1}) + w(2,4)(r_{42} - \overline{r_4})}{w(2,1) + w(2,4)}$$

$$=3+\frac{0.997 \times (2-2) + 0.86 * (3-2.4)}{0.997 + 0.86} = 3.27$$

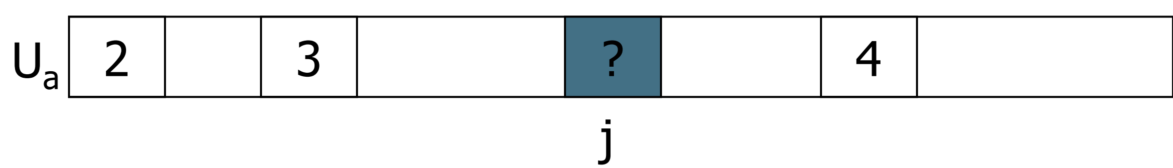
CF: Item-Item method

- Offline phase:
  - Calculate n(n-1) similarity measures
  - For each item
    - Determine its most k-similar items



- Online phase:
  - Predict rating for a given user-item pair as a weighted sum over similar items that he rated

$$r_{aj} = \frac{\sum_{i \in similar\ items} s_{ij}\, r_{ai}}{\sum_{i \in similar\ items} s_{ij}}$$



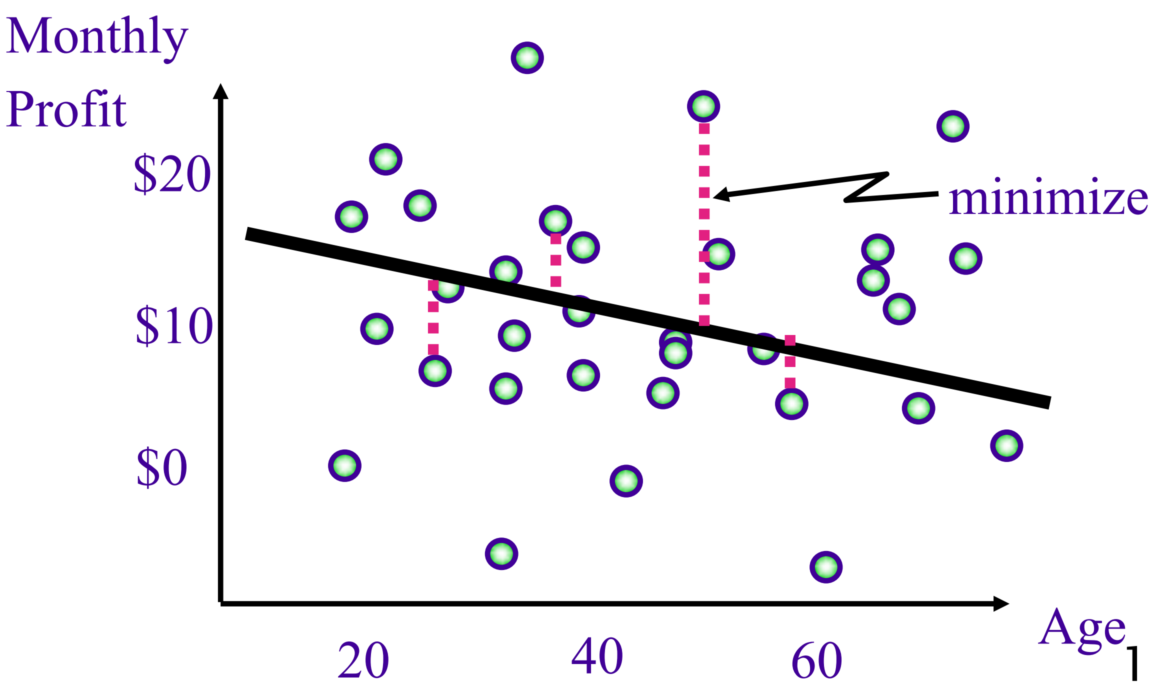
Regression

Regress Analysis

- <u>Linear regression</u>:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - Parameters: β_0 , β_1
  - Random variable: $\varepsilon \sim N(0, \sigma^2)$



steps

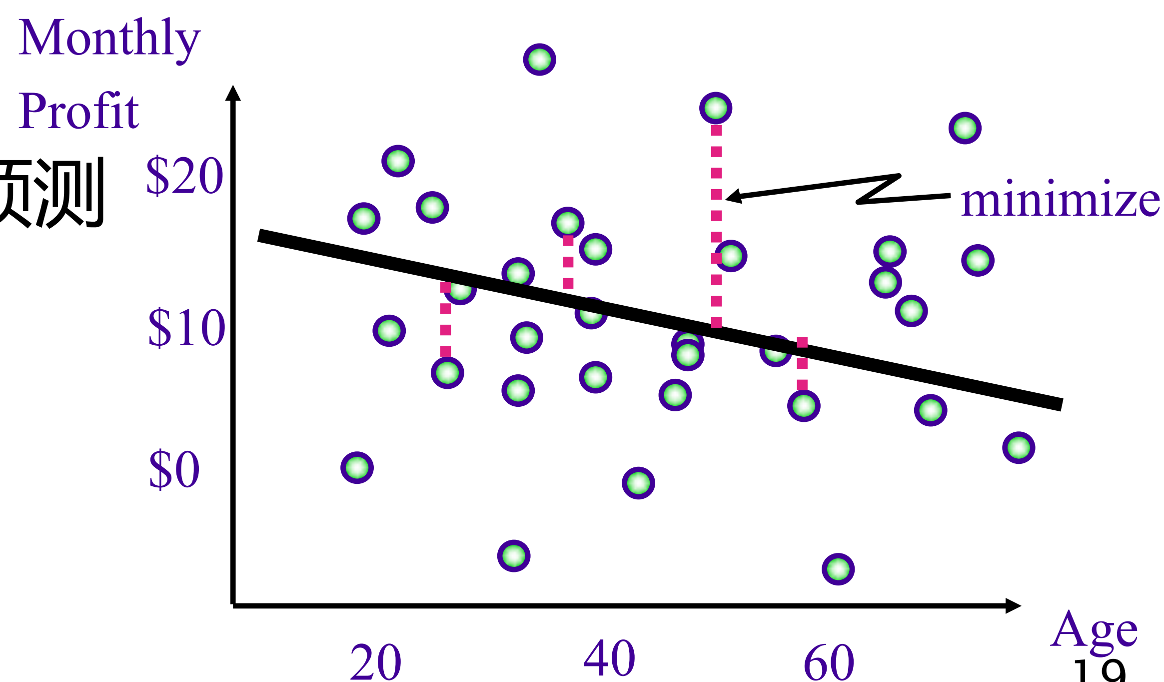
（1）构建包含因变量和自变量的训练集；

（2）通过散点图，确认因变量和自变量之间的近似线性关系；

（3）计算系数，构建模型；

（4）检验模型；

（5）利用模型进行预测

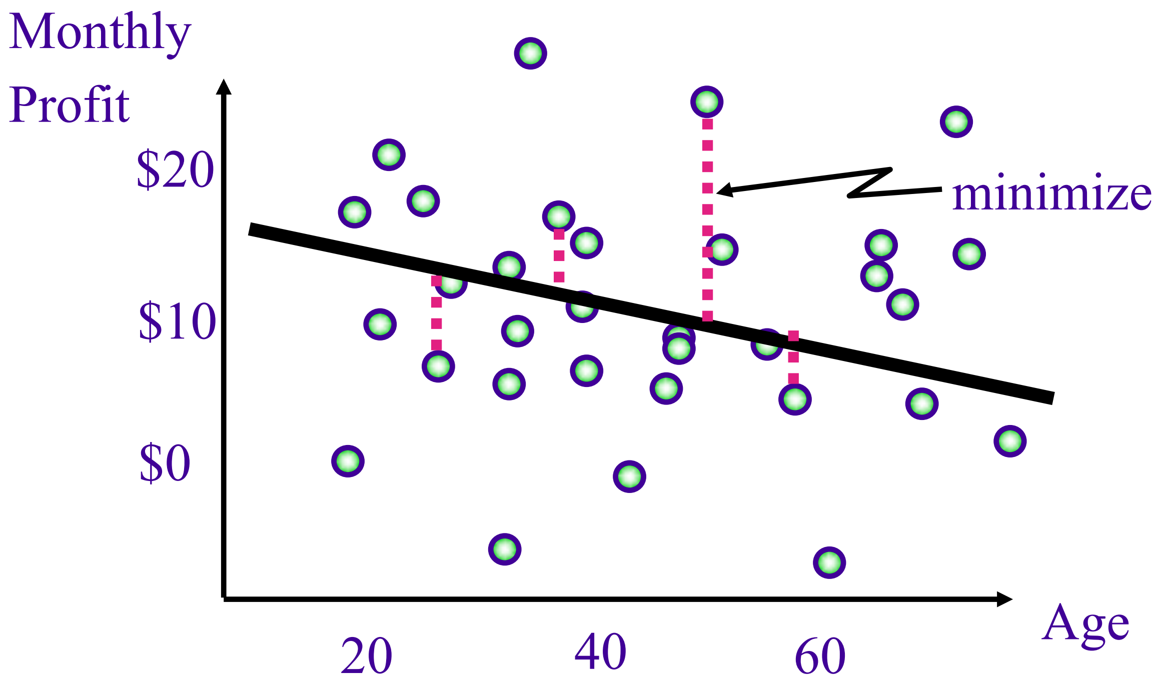


Regress Analysis

- Estimate parameters based on training dataset

$$\hat{y}_i = a + bx_i$$

  - Using the least squares criterion to the known values of (x1, y1) (x2,y2) ...



Estimating parameters

- least squares (最小二乘法)
  - 残差平方和,
  - minimize SS_E

$$SS_E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

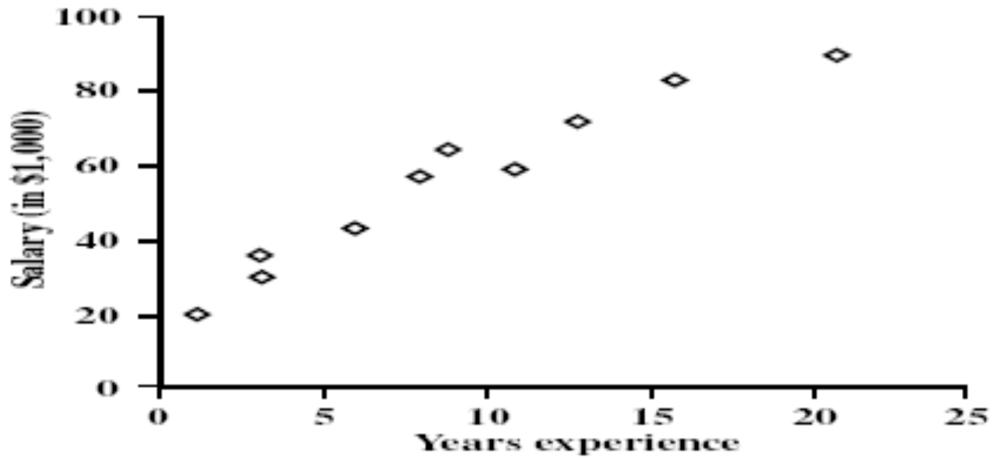
$$b = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - b \times \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - $\bar{x}$和$\bar{y}$: mean values of X and Y
  - s_{xx} 称为x的校正平方和, s_{xy} 称为校正交叉乘积和
  - s_{yy}称为y的校正平方和。

Linear Regression

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



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

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

Model validation

$$SS_E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 回归平方和SS_R

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- 总离差平方和SS_T: 将y的均值作为总体估计值时的误差

$$SS_T = \sum_{i=1}^{n} (\mathrm{y}_i - \bar{y})^2$$

- $SS_T=SS_E+SS_R$
  - 总离差平方和中被回归模型解释的部分为回归平方和

- 拟合优度检验
  - R^2, adjusted R square $\bar{R}^2$
  - n为样本个数，k为自变量的个数

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$$\bar{R}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

回归模型的显著性检验

- Hypothesis

$$H_0\text{：} b=0; \qquad H_1: b\neq 0$$

- 可以证明在H_0成立的情况下由下式定义的F符合F(1,n-2)分布

$$F=\frac{SS_R}{SS_E/(n-2)}$$

- 给定显著性水平α，查自由度为（1，n-2）的F分布临界值表得临界值$F_\alpha(1,n-2)$，若由上式计算的$F_0 > F_\alpha(1,n-2)$则因变量和自变量之间的线性关系显著，假设H_0被拒绝

回归系数的显著性检验

- 为了检验回归模型中每个回归系数的显著性，可以推导出系数a和b的样本方差

$$S_b^{\ 2} = \frac{SS_E/(n-2)}{S_{xx}} \qquad S_a^{\ 2} = \frac{SS_E}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

- 可以证明$t_b=b/S_b$和$t_a=a/S_a$均符合自由度为(n-2)的t分布
- 其中重要的是检验系数b是否为0。因此需要检验假设

$$H_0:\ b=0; \qquad H_1:\ b\neq 0$$

- 给定显著性水平α，查自由度为(n-2)的t分布表，得到$t_\alpha(n-2)$若$t_b > t_\alpha(n-2)$，则拒绝假设H_0，即回归系数b显著
- 同时可以计算出P 值（p value），一般以$P < 0.05$ 为显著， $P < 0.01$ 为非常显著。

Regression Analysis

- Multiple regression:

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k+\varepsilon$$

$\varepsilon \sim N(0, \sigma^2)$

- n observation: $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$，i=1,2, …n。Based on these observations, estimate parameters：

$$\hat{y}_i = b_0 + b_1 x_i + \ldots + b_k x_k$$

- b_0、b_1、…b_k是β_0、β_1、…β_k的最小二乘估计

Regression Analysis

- Many nonlinear functions can be transformed into the above:
- $\boldsymbol{y=ax^b}$: $\lg y=\lg a+b\lg x$
- $y=ae^{bx}$，可以通过两边取对数变换为$\ln y=\ln a+bx$
- $y=a+b\lg x$，设 $X=\lg x$，则有$y= a+b\,X$.
- Commonly solved by using of statistical software packages, such as SAS, SPSS, and S-Plus
- Weka: weka.classify.functions.LinearRegression
  - Data: cpu

Evaluating Numeric Prediction

- Similar to classification
  - Training set, test set
  - cross-validation
- Different from classification
  - Quality measure by error rate is not appropriate
  - 均方误差(mean-squared error)
  - 均方根误差(root mean-squared error)
  - 平均绝对误差(mean absolute error)
  - 相对平方误差(relative squared error)
  - 相对绝对误差(relative absolute error)

Evaluating Numeric Prediction

- 对于test set中的每个样本（$x_{i1}, x_{i2}, \ldots, a_i$), 其预测值为$p_i$

mean-squared error

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$$

root mean-squared error

$$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$$

mean absolute error

$$\frac{|p_1-a_1|+\ldots+|p_n-a_n|}{n}$$

relative squared error

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}, \text{ where } \bar{a}=\frac{1}{n}\sum_i a_i$$

root relative squared error

$$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$$

relative absolute error

$$\frac{|p_1-a_1|+\ldots+|p_n-a_n|}{|a_1-\bar{a}|+\ldots+|a_n-\bar{a}|}$$