Webmining 1

2022年7月11日 星期一 10:08

Cross - INdustry

- crisp - DM

{ regression
classification

1. crisp - DM

cross- industry standard process
for data mining

step I. understand purpose of the data mining study

step 2. understand data.

{ 1. perform statistical analysis
2. perform various types of visualizations

step 3. data consolidation { collect
select
Integrate

⇓

data cleaning { Impute missing values
Reduce noise in data
Eliminate inconsistencies

⇓

data Transformation { Normalize data
Discretize / aggregate data
Construct new attributes

⇓

data reduction { reduce number of varibles
reduce number of cases
balance ..

step 4 — model building

step 5 — testing & evaluation

{ regression → 偏差值
classification → 离群率
other evaluation

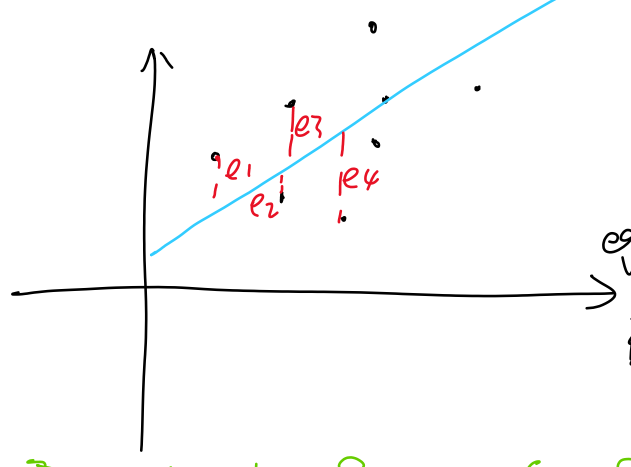
step 6 — deployment

Simple Linear Regression

$$Y \approx \beta_0 + \underbrace{\beta_1}_{\text{gradient}} X \Rightarrow Sales \approx \beta_0 + \beta_1 TV$$

Traning Data

$E = e_1^2 + e_2^2 + \dots + e_8^2$ (square errors)



$e_i : y_i - \hat{y}_i$ (RSS)

$E = e_1^2 + e_2^2 + \dots + e_8^2$

(make sure it's positive)

eg. manhanton distance

$E = |e_1| + |e_2| + \dots + |e_8|$

Residual Sum of Square

useful predictors

X1	X2	X3	X4..	y

R^2 : 0 ~ 1
不好 ✓ 好

F - STATISTics

越接近1，越无关，越大于1，越相关

Prob: (F-statistic): 1.47e-17

MSE : Mean Square

$$= \frac{1}{\text{degrees-of-freedom}} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

cading SCHEME (gender..)

gender $\xrightarrow{\text{numeric}}$ { 0
1

Ethnicity $\xrightarrow{\text{numeric}}$ { Caucasian 0
Asian 1 ×
African American 2

⟶ { 1 if in Caucassin
0 if not Caucasian