11/02/21

Lec-18

Regression Analysis

→ classification vs. prediction (regression)

→ Regression (defn)

→ (linear, multiple linear, non-linear)

* classification predicts categorical / or discrete value
* Regression " continuous value.

X ⟨ clump dia, clump thick, texture ⟩ → ML model → class label: A normal, B benign, C malignant

(classification problem)

min temp, max temp, humidity, fog, precipitation of last 7 days → ML model → predict temp, rainfall of tomorrow

⟨ min 26°c, max 25°c   20 cm ⟩

(Regression)

ML model :

curve fitting process

$$\boxed{y = f(x)}$$

(70 % Tr / 30 % Test)

I/p matrix

1–50	C1
51–100	C2

Tr: 1–35 C1, 36–70 C2

Test: 1–15 C1, 16–30 C2

c1 → 1 0 0
c2 → 0 1 0
c3 → 0 0 1

Tr

x	y
C1	1 0
C2	0 1

Linear
$y = mx + c$

Non linear
$y = ax^2 + bx + c$

Irregular shape curve.

→ **Regression (defn)** Sir Frances Galton [1822-1911]

Regression analysis is used to model the relationship between one or more independent or predictor variables and a dependent or response variable (which is continuous valued).

$$\boxed{y = b + wx}$$

dependent or response variable ← y; x → predictor variable

Linear regression → involves a single predictor variable
multiple " " → " two/or more " "

Other regression models →
- generalized linear model
- Poisson regression
- log linear model
- regression tree

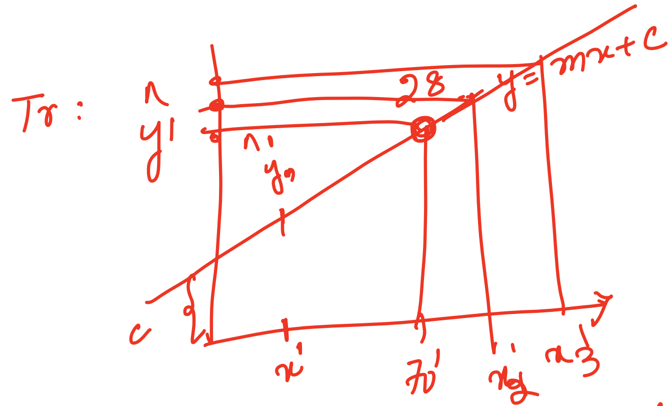
① **Linear regression:**

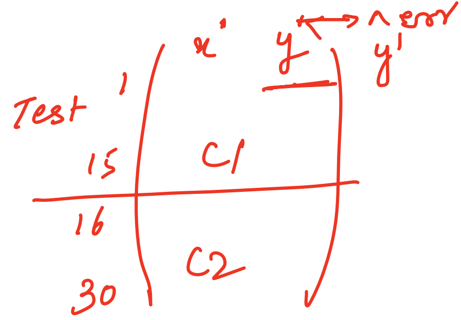
$$\boxed{y = b + wx}$$

where variance of y is assumed to be constant.

b and w are regression coefficients specifying Y intercept and slope of line resp.

x Exp in y	y Sal in $
—	—
—	—
—	—
—	—
—	—





	hum	Temp	$\hat{y}'$
$\rightarrow$	70	25	28

$T_1 \rightarrow (\hat{y}' - y')^2 = 3^2 = 9$

$T_2 \rightarrow (\quad)^2 = 1^2 = 1$

min $\sum$ err $(\hat{y}' - y')^2$

$\vdots$

T_{30}

minimize $\left\{ M.S.E = \dfrac{e_1 + e_2 + \dots + e_{30}}{30} \right\}$

→ These coefficients can be solved by method of least square (minimize squared error), which estimates the best fitting straight line as the one that minimizes the error between the actual data and estimate of the line.

$y = w_0 + w_1 x$ (assume w and b as wts. in above eqn.)

$c + mx$

regression coefficients are estimated as follows:

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

$$\boxed{w_0 = \bar{y} - w_1 \bar{x}} \quad (2)$$

$$\boxed{y = w_0 + w_1 x}$$

3

Numerical :

$x \rightarrow$ no. of years of work exp

$y \rightarrow$ salary data

Eg 1

x in (yrs)	y (sal in 1000 dollars)	A $(x_i - \bar{x})$	B $(y_i - \bar{y})$	$(x_i - \bar{x})^2$	A × B
3	30				
8	57				
9	64				
13	72				
3	36				
6	43				
11	59				
21	90				
1	20				
16	83				

$\bar{x} = ?$ 9.1 $\quad \bar{y} = ?$ 55.4

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

$$\boxed{w_1 = 3.5}$$

$$\boxed{w_0 = 23.6}$$

Eqn of best fit line $= \boxed{y = 23.6 + 3.5x}$

find if $x = 10$ yrs Exp ??

$$\boxed{y = 23.6 + 3.5x}$$

$\hat{y}_1 = 23.6 + 3.5\,(x)\ 3$

$= 34.1$ (estimated)

$y_1 = 30$ (actual)

$err_1^2 = (y_1 - \hat{y}_1)^2 = (30 - 34.1)^2 =$ ___

$err_2^2 = (y_2 - \hat{y}_2)^2 =$ ___

⋮

$$rmse = \sqrt{\frac{err_1^2 + \dots + err_{10}^2}{10}}$$

for Eg. $=$ 5.11 rmse

y if $x = 10$ yrs??

$y = 23.6 + 3.5x$

$= 23.6 + 3.5 \times 10$

$= (58.6) \times 1000\$$

$= 58600\ \$$

Basic concept :

Let D denote a data set that contains N observations

$$D = \{ (x_i, y_i) \mid i = 1, 2, \ldots N \}$$

Each x_i corresponds to the set of attributes of the ith observation also known as explanatory variables and y_i corresponds to the target or response variable.

→ Regression is the task of learning a target function f that maps each attribute set x into a continuous valued output y.

→ The goal of regression is to find a target function that can fit the input data with min. error.

→ The error function for a regression task can be expressed in terms of the sum of absolute or squared error.

$$\text{Absolute error} = \sum_i |y_i - f(x_i)| \quad \text{— (1)}$$

$$\text{Squared error} = \sum_i (y_i - f(x_i))^2 \quad \text{— (2)}$$

Least square method –

Suppose we wish to fit the following linear model to the observed data :

$$f(x) = w_1 x + w_0$$

where w_0 and w_1 are parameters of the model and are called the regression coefficients.

→ A standard approach for doing this is to apply the method of least-squares ; which attempts to find the parameters (w_0, w_1) that minimize the sum of squared error

$$SSE = \sum_{i=1}^{N} [y_i - f(x_i)]^2 = \sum_{i=1}^{N} [y_i - w_1 x - w_0]^2$$

which is also known as the <u>residual sum of squares</u>.

→ This optimization problem can be solved by taking the partial derivative of E wrt w_0 and w_1, setting them to 0, and solving the corresponding system of linear equations.

$$\frac{\partial E}{\partial w_0} = -2 \sum_{i=1}^{N} [y_i - w_1 x_i - w_0] = 0$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{i=1}^{N} [y_i - w_1 x_i - w_0] x_i = 0$$

solving the eqns. we get the following expression

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$$

$$\hat{w}_1 = \frac{\sigma_{xy}}{\sigma_{xx}}$$

where $\bar{x} = \frac{\sum x_i}{N}$ $\quad \bar{y} = \frac{\sum y_i}{N}$

$$\sigma_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_{xx} = \sum_i (x_i - \bar{x})^2$$

$$\sigma_{yy} = \sum_i (y_i - \bar{y})^2$$

Thus, linear model that results in the min. squared error is given by

$$f(x) = \bar{y} + \frac{\sigma_{xy}}{\sigma_{xx}}\left[x - \bar{x}\right]$$

Q. Calculate the reg. coefficient and obtain the line of reg. for the following data

x	y	x^2	y^2	$x.y$	$\hat{y}$	$(y-\hat{y})^2$
1	9	1	81	9		
2	8	4	64	16		
3	10	9	100	30		
4	12	16	144	48		
5	11	25	121	55		
6	13	36	169	78		
7	14	49	196	98		

$\Sigma X = 28$ $\Sigma Y = 77$ $\Sigma X^2 = 140$ $\Sigma Y^2 = 875$ $\Sigma XY = 334$

$$\boxed{Y = 0.929X + 7.284}$$

$SSE = 3.866$

$MSE = \frac{3.866}{7}$

$rmse = \sqrt{\frac{3.866}{7}}$

$= 0.79$

$\boxed{e = Y - \hat{Y}}$ (e is known as residual

$Y \rightarrow$ observed value

$\hat{Y} \rightarrow$ fitted value

→ The magnitude of residuals provide a good indication of how useful the regression line is for predicting y values from X values.

→ To summarize numerous error with a single numeric measure, the standard error of estimate denoted as (S_e) is mostly used which essentially measures the standard deviation of residuals.

$$S_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

→ denominator $(n-2)$ denotes the no. of parameters to be estimated from the sample size n. (Here the parameters are slope and intercept).