Dhanvanth Voona
A20543395

Regression Math 564

Homework - 1

1) (a) The given scenario is $\boxed{\text{Regression}}$ problem because the CEO salary can not be categorised, instead it is a ~~discrete number~~ Continuous variable.

→ Since we are interested in identifying which factors affect target variable, It is a $\boxed{\text{inference}}$ problem

→ Here, number of observations = no. of firms in the given data

$\boxed{n = 500}$

No. of predictors = (Profit, no. of employees, industry)

$\boxed{p = 3}$

(b) → This scenario is a $\boxed{\text{klassification}}$ problem because it can be categorised with "success" or "failure".

→ Here the goal is to predict the outcome of a product it is $\boxed{\text{prediction}}$ problem

→ Here $\boxed{n = 20}$ [No. of previous products]

p = (Price charged, marketing budget, competition price & 10 other variables)

$\boxed{p = 13}$

(c) → The % change in the USD/Euro ~~ex~~ exchange rate is a continuous variable. ∴ It is $\boxed{\text{Regression}}$ problem

→ Here the goal is to predict % change in USD / Euro exchange rate ∴ It is a $\boxed{\text{Prediction}}$ problem

→ number of data points (n) = no. of weeks in a year

$\Rightarrow \boxed{n = 52}$ (✗)

p = (% change in US ~~Euro~~ market, % change in British market, & % change in German market)

$\boxed{p = 3}$

(2) $\boxed{\text{Parametric Methods}}$:

$\boxed{\text{Advantages}}$:- • The models are often straightforward to interpret. For example, consider a linear model with Parameters $\beta_1, \beta_2 \ldots \beta_p$. These parameters infer about the effect of each Predictor on target Variables which provides clear understanding of Predictor variables and its relationship with target variable.

• Since they make strong assumption about form of data, assuming the assumptions are true, they would require only small set of data.

Disadvantages :-

→ These methods are inflexible as they are constrained by predefined functional form. This makes it difficult to capture the complex & nonlinear data.

→ Highly dependent on assumption of chosen model. If the data is non-linearity in nature, a linear model assumption will fail to capture its true nature

Non-Parametric methods :-

Advantages :-

→ No assumptions are made on data distributions. So it can capture complex and non linear data

→ Since there is no prior assumption of function 'f', they avoid danger of choosing a model that is very different from true underlying function.

Disadvantages :-

→ They involve more complex calculation & require processing the entire dataset rather than a simplified model which increases the cost of computation

→ Unlike Parametric methods, they are hard to interpret.

3A To calculate:

least square line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Given Data

x_i	y_i	$x_i y_i$	x_i^2
1	10	10	1
2	14	28	4
4	12	48	16
6	13	78	36
7	15	105	49
8	12	96	64
10	13	130	100

$\boxed{\Sigma x = 38, \ \Sigma y = 89 \quad \Sigma x y_i = 495 \quad \Sigma x_i^2 = 270}$ $n = 7$

Given

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{7 \times 495 - (89 \times 38)}{7 \times 270 - 38^2} = \frac{3465 - 3382}{1890 - 1444}$$

$$\boxed{\hat{\beta}_1 = \frac{83}{446}}$$

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

$= \frac{89}{7} - \frac{83}{446}\left(\frac{38}{7}\right) = \frac{89}{7} - \frac{3154}{3122}$

$= \frac{39694 - 3154}{3122} = \frac{36540}{3122}$

$\boxed{\hat{\beta}_0 = \frac{18270}{1561}}$

So the least square line is

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \frac{18270}{1561} + \frac{83}{446} x$

$\boxed{\hat{y} = \frac{18270}{1561} + \frac{83}{446} x}$

when $x = 1$

$\hat{y} = 11.704 + 0.18 = 11.89$

$= \frac{5303}{446}$

4A

To show, for the least square line obtained from training data, we have

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$,

$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0.$

Firstly, $y_i \rightarrow$ Real value, $\hat{y}_i \rightarrow$ Predicted Value

So, $(y_i - \hat{y}_i)$ is error (residual) and we know that if the sum of residuals is equal to zero then it is unbiased model.

To prove that :

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

We know that $\hat{\beta}_0$ & $\hat{\beta}_1$ are constant, and if you add $\hat{\beta}_0$ for 'n' times it will be $n\hat{\beta}_0$.

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i} \quad \text{—— (1)}$$

We also know that, $\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$ —— (2)

$$\boxed{\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \& \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i} \quad \text{—— (3)}$$

Substitute (2) in (1)

$$\sum_{i=1}^{n} \hat{y}_i = (\bar{y} - \hat{\beta}_1 \bar{x}) n + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\boxed{\sum_{i=1}^{n} \hat{y}_i = (n\bar{y}) - \hat{\beta}_1 (n\bar{x}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i} \quad (4)$$

~~Subst~~ Substitute (3) in (4)

$$\cancel{n\bar{y}} = \cancel{n\bar{y}} - \cancel{\hat{\beta}(n\bar{x})} + \cancel{\hat{\beta}_1 \bar{x}}$$

$$\boxed{\sum_{i=1}^{n} \hat{y}_i = n\bar{y}} - \cancel{\hat{\beta}_1 (n\bar{x})} + \cancel{\hat{\beta}_1 (n\bar{x})}$$

$\Rightarrow$ Now: we have,

$$\sum_{i=1}^{n} \hat{y}_i = n\bar{y} \quad \& \quad \sum_{i=1}^{n} y_i = n\bar{y} \quad \therefore \text{(5)}$$

Substitute (5) in below equation.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = n\bar{y} - n\bar{y} = 0$$

5A

(a) We know that, Expected test Mean Square Error

$$E[Y-\hat{Y}]^2 = \text{var}(\hat{f}(x_0)) + \text{Bias}(\hat{f}(x_0))^2 + \text{Var}(\epsilon)$$

~~three~~ ~~three~~ Variance ~~for that~~

We know that variance is inherently a non negative quantity. and square bias is also non-negative. whereas $\text{Var}(\epsilon)$ is an irreducible error.

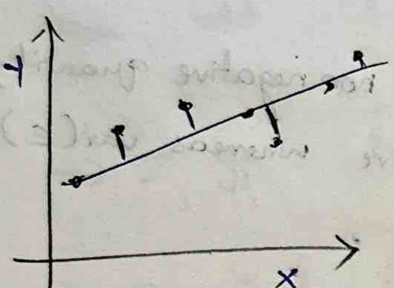
So, In order to reduce Expected test MSE, we need to ~~obtain~~ minimize the variance & Bias.

Variance: It refers to amount by which $\hat{f}$ should change if we estimate it using a different training dataset. So if a statistical method has high variance, it means it is over fitting and shows weak Performance for ~~unknown~~ ~~dat~~ other training set

Bias :- Refers to error that is introduced by approximating real world problem which may be extremely complicated by a simpler model.

(b) Scenario 1 :- Model is very simple like linear regression with only one feature.

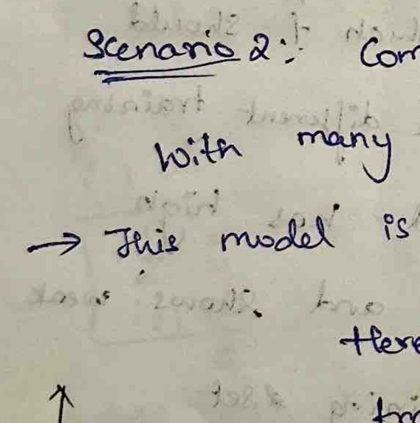
→ This model is more likely to have high bias and low variance.



Here the model show high error while ~~estimated~~ predicting y because the linear model assuming linear relationship for non-linear data. This underperformance with high bias causes underfitting model.

Scenario 2 :- Complex ~~need~~ model, like a polynomial model with many features.

→ This model is more likely to have low bias and high variance



Here blue line in the graph represents training set and polynomial model fitting. But Black lines represent change in training set and model poorly predicts for this. This underperformance with high variance causes overfitting model

To conclude,

High bias → ~~overfitting~~ underfitting

High variance → over fitting

Hence we need to choose a model with a tradeoff between bias and variance.