2020 - Mid

1. (a) Validation of Simple linear Regression Model:

① Coefficient of determination or $R^2 = \frac{SSR}{SST}$

② Hypothesis test for regression coefficient

③ Analysis of variance for overall model validity

④ Residual analysis to validate regression model assumpt

⑤ Outlier Analysis

(b) Lets use the OLS method (Ordinary least square)

we choose 5 random values of x_1 & x_2

i	x_1	x_2	y	dy	dx_1	dx_2
1	1	3	15	1	-1	1
2	2	1	11	-3	0	-1
3	1	2	12	-2	-1	0
4	3	3	19	5	1	1
5	3	1	13	-1	1	-1

$\bar{x}_1 = 2 \quad \bar{x}_2 = 2 \quad \bar{y} = 14$

$$\beta_1 = \frac{\Sigma dy\, dx_1}{\Sigma (dx_1)^2} = \frac{-1+0+2+5-1}{1+0+1+1+1} = \frac{5}{4} = 1.25$$

$$\beta_2 = \frac{\Sigma dy\, dx_2}{\Sigma (dx_2)^2} = \frac{1+3+0+5+1}{1+1+0+1+1} = \frac{10}{4} = 2.5$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

$$= 14 - 1.25 \times 2 - 2.5 \times 2 = 14 - 2.5 - 5 = 6.5$$

Thus $\boxed{\hat{y} = 6.5 + 1.25 x_1 + 2.5 x_2}$

(c) Null Hypothesis: $\exists\ x' - x = 1$ such that the sold price decreases by less than 75 times or increases

Alternate Hypothesis: $\forall\ x' - x = 1$ sold price decreases by atleast 75 times

x' = increased batting strike rate

x = prev. batting strike rate

(d) (i) VIF (Variance Inflation Factor) -> measures the severity of multicollinearity in regression analysis.

$VIF \in [1, \infty)$

(ii) Overfitting

Def-> A hypothesis h is said to overfit the training data if there is another hypothesis h' such that

$$error(h) < error(h') \text{ on training data}$$
$$error(h') < error(h) \text{ on test data}$$

=> due to more complexity in the model or many outlier due to noise in the data.

2. (b) (i)

$$\theta_1 = \underset{\theta_1}{\operatorname{argmax}} \left(\ln \left(\prod_{k=1}^{2} \theta_1 e^{-\theta_1 x_k} \right) \right)$$

$$= \underset{\theta_1}{\operatorname{argmax}} \left(\sum_{k=1}^{2} \left(\ln \theta_1 - \theta_1 x_k \right) \right)$$

$$\frac{\partial}{\partial \theta_1} = 0$$

$$\Rightarrow \sum_{k=1}^{2} \left(\frac{1}{\theta_1} - x_k \right) = 0$$

$$\Rightarrow \frac{1}{\theta_1} - 1 + \frac{1}{\theta_1} - 5 = 0 \quad (\because D_1 = \{1, 5\})$$

$$\Rightarrow \frac{2}{\theta_1} = 6 \quad \Rightarrow \boxed{\theta_1 = \frac{1}{3}}$$

similarly for θ_2

$$\frac{\partial}{\partial \theta_2} = 0 \quad \Rightarrow \sum_{k=1}^{3} \left(\frac{1}{\theta_2} - x_k \right) = 0$$

$$\Rightarrow \left(\frac{1}{\theta_2} - 3 \right) + \left(\frac{1}{\theta_2} - 6 \right) + \left(\frac{1}{\theta_2} - 9 \right) = 0$$

$(\because D_2 = \{3, 6, 9\})$

$$\Rightarrow \frac{3}{\theta_2} = 18 \quad \Rightarrow \boxed{\theta_2 = \frac{1}{6}}$$

(i) for $x \geq 0$ decision boundary is when:

$$p(x/\theta_1) = p(x/\theta_2)$$

$$\Rightarrow \theta_1 e^{-\theta_1 x} = \theta_2 e^{-\theta_2 x}$$

$$\Rightarrow \frac{1}{3} e^{-\frac{x}{3}} = \frac{1}{6} e^{-\frac{x}{6}}$$

$$\Rightarrow e^{-\frac{x}{3}+\frac{x}{6}} = \frac{1}{2}$$

$$\Rightarrow -\frac{x}{6} = \ln\frac{1}{2} \Rightarrow x = -6\ln\frac{1}{2} = 4.15$$

thus for $x > -6\ln\frac{1}{2}$ → class 2

otherwise class 1

$$\frac{\frac{1}{3} e^{-\frac{1}{3}\times 2^2}}{\frac{1}{6} e^{-\frac{1}{6}\times 2}} = \frac{0.0451}{0.0613}$$

(c)

Confusion matrix is used to summarize and represent the correct and incorrect predictions made by a classifier model.

→ More useful in binary classification but can be generalised for any no. of classes.

ROC = Receiver Operating Characteristics

↳ plot TPR vs FPR

True +ve rate $= \frac{TP}{TP+FN}$ False +ve rate $\frac{FP}{FP+TN}$

3. (a) Entropy (S) $= -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$

$= -\frac{4}{7} \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7}$

$= -\frac{4}{7} \times -0.807 - \frac{3}{7} \times -1.222$

$= 0.461 + 0.523 = \underline{\underline{0.984}}$

$$\text{Gain}(S, \text{city}) = \text{entropy}(S) - \sum_{v \in \text{values of city}} \frac{|S_v|}{|S|} \text{entropy}(S_v)$$

$= 0.984 - \left[\frac{3}{7} \left(-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) + \right.$

bangalore (2(+), 1(−)) → $\frac{3}{7}$

$\left. \frac{4}{7} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right]$ → chennai (2(+), 2(−)) → $\frac{4}{7}$

$= 0.984 - \frac{3}{7} (+0.389 + 0.528) - \frac{4}{7} \times 1$

$= 0.984 - 0.393 - 0.571$

$= \underline{0.002}$

emp 3+ 1−
stu 1+ 1−
unemp 0+ 1−

$Gain(S, job) = 0.984 - \left[\frac{4}{7}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{2}{7}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{7}(-0 - 1\log_2 1)\right]$

$= 0.984 - \frac{4}{7}(0.311 + 0.5) - \frac{2}{7}\times 1 - 0$

$= 0.984 - 0.463 - 0.285 = \underline{0.286}$

(b)

Expected value of $y(x) = 4x + 3$

$= \int_{-\infty}^{\infty} y(x) f(x)\, dx$

$= \int_{-1}^{2} (4x+3)\frac{x^2}{3}\, dx = \int_{-1}^{2}\left(\frac{4}{3}x^3 + x^2\right) dx$

$= \left[\frac{2}{3}x^4 + \frac{x^3}{3}\right]_{-1}^{2} = \frac{16}{3} + \frac{8}{3} - \frac{1}{3} - \frac{1}{3}$

$= \underline{8}$

(c) Issues in decision tree:

(i) Overfitting due to larger trees:

- Reduced error pruning
- Rule post pruning

} by splitting the train set to training & validation set

(ii) Extensions to other type of data like

	How to deal
→ continuous valued attributes	convert to discrete by partitioning the range using thresholds
→ Attributes with different cost	Modify ID3 algo to take into account the costs of attributes
→ Information gain might not always split the set effectively	→ Use of gain ratio instead