Theory Assignment - 4

Ans:- Steps to calculate Information Gain:-

i) Calculate entropy of training set T given by
$$H(T) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

ii) For each attribute a that divides T into subsets T_i perform the following

a) Calculate the entropy of each subset T_i

b) Calculate the avg. entropy : $H(T, a) = \sum_i P_i H(T_i)$

c) Calculate the information gain : $I(T, a) = H(T) - H(T, a)$

To select the ~~highest gain~~ most informative attribute, choose the attribute with highest gain.

Following the above steps to obtain the order of the most to least informative attributes.

i) Entropy of Training Set, $H(T)$

$$H(T) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$
$$= -\frac{5}{9} \log_2\left(\frac{5}{9}\right) - \frac{4}{9} \log_2\left(\frac{4}{9}\right)$$
$$= 0.991$$

ii) Calculate average entropy & information gain for each attribute.

a) Attribute, a = Type of Call

$$H(\text{Type of call} = \text{Local}) = -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$$
$$= -\frac{1}{2} \log_2 2^{-1} - \frac{1}{2} \log_2 2^{-1}$$

$= \frac{1}{2} + \frac{1}{2} = 1$

$H(\text{Type of call} = \text{Intern}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.91829$

$H(\text{Type of Call} = \text{Long Dis.}) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)$

$= -1\log_2 1 - 0$

$= 0$

Therefore, $H(T, \text{Type of call}) = \frac{4}{9}\log_2\left(\frac{4}{9}\right)$.

Therefore, $H(T, \text{Type of Call}) = \left(\frac{4}{9} \times 1\right) + \left(\frac{3}{9} \times 0.042\right) +$

$= \left(\frac{4}{9} \times 1\right) + \left(\frac{3}{9} \times 0.91829\right) + \left(\frac{2}{9} \times 0\right)$

$= 0.7505$

$= \frac{4}{9} + 0.014 + 0$

$= 0.4584$

b) Attribute, a = Language Frequency

$H(\text{Lang. Freq} = \text{Fluent}) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$

$= -1\log_2 1 = 0$

$H(\text{Lang. Freq} = \text{Not Fluent}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= \frac{1}{3}\log_2 3 - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

~~0.52764~~ = 0.91829

$H(\text{Lang Freq} = \text{Accent}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= \frac{1}{2} + \frac{1}{2}$

$= 1$

$H(\text{Lang Freq} = \text{Foreign}) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$

$= 0$

Therefore, $H(T, \text{Lang Freq.}) = \left[\left(\frac{3}{9}\right) \times 0\right] + \left(\frac{3}{9} \times \text{0.91829}\right) + \left(\frac{2}{9} \times 1\right) + \left(\frac{1}{9} \times 0\right)$

$= \left(\frac{1}{3} \times 0.2764\right) + \left(\frac{2}{9}\right)$

~~0.3145~~ = ~~0.5283~~ 0.5283

c) Attribute, a = Ticket Type

$H(\text{Ticket Type} = \text{Short}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= \frac{1}{2} + \frac{1}{2} = 1$

$H(\text{Ticket Type} = \text{Long}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)$

= ~~0.2922~~ 0.97095

Therefore, $H(T, \text{Ticket Type}) = \left(\frac{4}{9} \times 1\right) + \left(\frac{5}{9} \times \text{0.97095}\right)$

= ~~0.6068~~ 0.98386

d) Attribute, a = age

$H(\text{age} = \text{Very young}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \log_2\left(\frac{1}{2}\right)$

$= \frac{1}{2} + \frac{1}{2} = 1$

$H(\text{age} = \text{Old}) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$

$= 0$

$H(\text{age} = \text{very old}) = -\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)$

$= 0$

$H(\text{age} = \text{middle}) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.91829$

$H(\text{age} = \text{young}) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$

$= 0$

Therefore, $H(T, \text{age}) = \left(\frac{2}{9} \times 1\right) + \left(\frac{1}{9} \times 0\right) + \left(\frac{2}{9} \times 0\right) + \left(\frac{3}{9} \times 0.91829\right) + \left(\frac{1}{9} \times 0\right)$

$= \frac{2}{9} + \left(\frac{1}{3} \times 0.91829\right) = 0.5283$

Calaclating the Information Grain for each attribute.

$I(T, \text{Type of Call}) = H(T) - H(T, \text{Type of Call})$

~~$= 0.9147 - 0.4384$~~ $= 0.991 - 0.7505$

~~$= 0.2983 - 0.4584$~~ $= 0.2405$

~~$= -0.1601$~~

$I(T, \text{Lang. Freq.}) = H(T) - H(T, \text{Lang. Freq.})$

~~$= 0.2983 - 0.3144$~~ $= 0.991 - 0.5283$

~~$= -0.0161$~~ $= 0.4627$

$I(T, \text{Ticket Type}) = H(T) - H(T, \text{Ticket Type})$

~~$= 0.2983 - 0.6068$~~ $= 0.991 - 0.98386$

~~$= -0.3085$~~ $= 0.00714$

$I(T, \text{age}) = H(T) - H(T, \text{age})$

~~$= 0.2983 - 0.3144$~~ $= 0.991 - 0.5283$

~~$= -0.0161$~~ $= 0.4627$

Hence, the rank of attributes based on information gain

Language frequency = Age > Type of Call > Ticket Type

Ans 2- Steps to find Principle Component.

i) Centralize the data.

ii) Calculate $p \times p$ covariance matrix / correlation matrix depending on situation M.

iii) Calculate the eigen values and the eigen vectors of the covariance matrix.

iv) Select m eigenvectors that correspond to the largest m eigen values.

Following the above steps to get the principle components :-

X_1	X_2
10	-3
9	-1
8	-2
11	-4
7	0

$\bar{X}_1 = 9 \; ; \; \bar{X}_2 = -2$

$X_1 - \bar{X}_1$	$X_2 - \bar{X}_2$
1	-1
0	1
-1	0
2	-2
-2	2

i) Centralising data

$$\bar{X}_1 = \frac{10+9+8+11+7}{5} = 9$$

$$\bar{X}_2 = \frac{(-3)+(-1)+(-2)+(-4)+0}{5} = -2$$

ii) $$Cov(X_1, X_2) = \frac{\sum_{i=1}^{n} X_{1i} X_{2i}}{n-1}$$

$$= \frac{[1\times(-1)]+[0\times 1]+[(-1)*0]+[2\times(-2)]+[(-2)\times 2]}{5-1}$$

$$= \frac{-1-4-4}{4} = \frac{-9}{4} = -2.25$$

$$\text{cov}(X_1, X_1) = \frac{\sum_{i=1}^{n} x_{1i}^2}{n-1}$$

$$= \frac{1^2+0^2+(-1)^2+2^2+(-2)^2}{5-1}$$

$$= \frac{1+1+4+4}{4} = \frac{10}{4} = 2.5$$

Similarly, $\text{cov}(X_2, X_2) = 2.5$

Hence,

$$\text{cov}(X_1, X_2) = \begin{bmatrix} 2.5 & -2.25 \\ -2.25 & 2.5 \end{bmatrix}$$

Note: Correlation Matrix is not needed since scale of variables is similar.

iii) To find eigen values, we have

$$|A - \lambda I| = 0$$

where A = Covariance matrix, I = identity matrix & λ = eigen value.

Therefore, $A - \lambda I = \begin{bmatrix} 2.5 & -2.25 \\ -2.25 & 2.5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$= \begin{bmatrix} 2.5-\lambda & -2.25 \\ -2.25 & 2.5-\lambda \end{bmatrix}$$

Now, $\det(A - \lambda I) = 0$

$$\begin{vmatrix} 2.5-\lambda & -2.25 \\ -2.25 & 2.5-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (2.5-\lambda)^2 - (-2.25)^2 = 0$$

$$\Rightarrow (2.5)^2 + \lambda^2 - 5\lambda - 5.0625 = 0$$

$$\Rightarrow \lambda^2 - 5\lambda + 1.1875 = 0$$

$$\Rightarrow \lambda = \frac{5 \pm \sqrt{25 - 4 \times 1.1875}}{2}$$

$$= \frac{5 \pm \sqrt{25 - 4.75}}{2}$$

$$= \frac{5 \pm 4.5}{2}$$

$$= \frac{9.5}{2} \text{ OR } \frac{0.5}{2}$$

$$= 4.75 \text{ OR } 0.25$$

Calculating the eigon vector for the highest λ ie $\lambda = 4.75$

To obtain eigen vector, we have,

$$B\bar{X} = \bar{0}$$

where $B = A - \lambda I$

$$= \begin{bmatrix} 2.5 - \lambda & -2.25 \\ -2.25 & 2.5 - \lambda \end{bmatrix}$$

and X = eigen vector, $\bar{0}$ = zero matrix

Therefore, substituting $\lambda = 4.75$ above,

$$B = \begin{bmatrix} 2.5 - 4.75 & -2.25 \\ -2.25 & 2.5 - 4.75 \end{bmatrix}$$

$$B = \begin{bmatrix} -2.25 & -2.25 \\ -2.25 & -2.25 \end{bmatrix}$$

Now,

$$\begin{bmatrix} -2.25 & -2.25 \\ -2.25 & -2.25 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(-2.25X_1) + (-2.25X_2) = 0$$

$$-2.25(X_1 + X_2) = 0$$

$$X_1 = -X_2$$

Now, taking $X_1 = 1 \Rightarrow X_2 = -1$

Hence,

$$\bar{X} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$ is the 1st principle component

Similarly, for $\lambda = 0.25$

Eigen vector is given by

$$B\bar{X} = \bar{0}$$

where $B = A - \lambda I$

$$= \begin{bmatrix} 2.5 - \lambda & -2.25 \\ -2.25 & 2.5 - \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 2.25 & -2.25 \\ -2.25 & 2.25 \end{bmatrix}$$

Now,

$$\begin{bmatrix} 2.25 & -2.25 \\ -2.25 & 2.25 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$2.25X_1 - 2.25X_2 = 0$$

$$2.25(X_1 - X_2) = 0$$

$$X_1 = X_2$$

Taking $X_2 = 1$, we have $X_1 = 1$

Therefore, $X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is the 2nd principle component

Ans 3-

a) The smallest training RSS will be for the model following best subset approach because the model will be chosen after considering all the possible combination of models with k parameters.

b) Similarities between Ridge regression and Lasso regression.

i) Both methods are used to solve the problem of multicollinearity. Both methods are used when the number of parameters is large ($p >> n$) where n is the number of observations.

ii) Both methods introduce bias in order to reduce variance in the dataset.

iii) Both work by shrinking the coefficients penalizing the absolute size of the regression coefficients.

iv) Both have same assumptions as the least squared method except normality is not assumed.

Dissimilarities between Ridge Regression and LASSO.

i) It reduces complexity by coefficient shrinkage but doesn't get rid of the irrelevant features and only minimizes their impact on the training model.	i) It reduces complexity by coefficient shrinkage but makes the coefficients absolutely zero (exact 0) and hence does automatic feature selection. (Gets rid of irrelevant features).
ii) It is faster in computation.	ii) It is slower in computation.

Although LASSO is slower in computation, I would still prefer LASSO because it reduces complexity by shrinking the coefficients to zero and thus automatically helps in feature selection.