Data Mining: Assignment Week 8: Regression

(Each question carries 1 mark)

1. Regression is used in:

A. predictive data mining

B. exploratory data mining

C. descriptive data mining

D. explanative data mining

Ans: A

Explanation: Regression is used for prediction.

2. In the regression equation Y = 21 - 3X, the slope is

A. 21
B. -21
C. 3
D. -3

Ans: D
Explanation: slope intercept form of a line is y=mx+c.

3. The output of a regression algorithm is usually a:

A. real variable

B. integer variable

C. character variable

D. string variable

Ans: A

Explanation: Regression outputs real variables.

4. Regression finds out the model parameters which produces the least square error between -

A. input value and output value

B. input value and target value

C. output value and target value

D. model parameters and output value

Ans: C

Explanation: Regression finds out the model parameters which minimises the error between the output value and the target value

5. The linear regression model $y = a_0 + a_1x$ is applied to the data in the table shown below. What is the value of the sum squared error function $S(a_0, a_1)$, when $a_0 = 1$, $a_1 = 2$?

x	y
1	1
2	1
4	6
3	2

A. 0.0
B. 27
C. 13.5
D. 54

Ans: D
Explanation: y' is the predicted output.

y' = 1+2x

x	y	y'
1	1	3
2	1	5
4	6	9
3	2	7

sum of squared error = $(1\text{-}3)^2 + (1\text{-}5)^2 + (6\text{-}9)^2 + (2\text{-}7)^2 =$ **54**

6. Consider x_1, x_2 to be the independent variables and y the dependent variable, which of the following represents a linear regression model?

A. $y = a_0 + a_1/x_1 + a_2/x_2$

B. $y = a_0 + a_1x_1 + a_2x_2$

C. $y = a_0 + a_1x_1 + a_2x_2^2$

D. $y = a_0 + a_1x_1^2 + a_2x_2$

Ans: B

Explanation: In option B y is linear in x.

7. Find all the eigenvalues of the following matrix A.

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 3 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

   A. 1,3
   B. 2,3
   C. 1,2,3
   D. Eigenvalues cannot be found.

Ans: C
Explanation: If A is an n × n triangular matrix (upper triangular, lower triangular, or diagonal), then the eigenvalues of A are entries of the main diagonal of A. Therefore, eigenvalues are 1,2,3.

8. In the figures below the training instances for classification problems are described by dots. The blue dotted lines indicate the actual functions and the red lines indicate the regression model. Which of the following statement is correct?
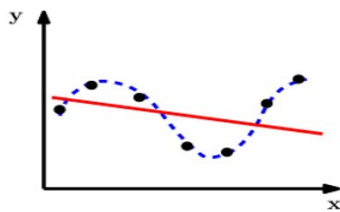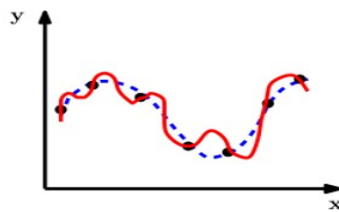
Figure 1   Figure 2

A. Figure 1 represents overfitting and Figure 2 represents underfitting

B. Figure 1 represents underrfitting and Figure 2 represents overfitting

C. Both Figure 1 and Figure 2 represents underfitting

D. Both Figure 1 and Figure 2 represents overfitting

Ans: B

Explanation: Definition of overfitting and underfitting.

9. In principal component analysis, the projected lower dimensional space corresponds to –

A. subset of the original co-ordinate axis

B. eigenvectors of the data covariance matrix

C. eigenvectors of the data distance matrix

D. orthogonal vectors to the original co-ordinate axis

Ans: B

Explanation: We must first subtract the mean of each variable from the dataset to center the data around the origin. Then, **we compute the covariance matrix of the data and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix.** Then we must normalize each of the orthogonal eigenvectors to become unit vectors. Once this is done, each of the mutually orthogonal, unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This choice of basis will transform our covariance matrix into a diagonalised form with the diagonal elements representing the variance of each axis.

10. A time series prediction problem is often best solved using?

A. Multivariate regression

B. Autoregression

C. Logistic regression

D. Sinusoidal regression

Ans : B

Explanation: Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.