Unit 3: Correlation and Regression

All four programmes were run in Jupyter Notebook and the changes in data points impacting correlation and regression were observed.

1. covariance_pearson_correlation.ipynb

Data point1 created with random seed generator which is independent variable, Datapoint2 is dependent on datapoint1 and correlated, hence change in datapoint1 impacts value of datapoint2

2. linear_regression.ipynb

For given x and y values, the key values such as slope, intercept were calculated and a linear relationship function between x and y established. With this function, for any value of x, y can be predicted

3. multiple_linear_regression.ipynb

For multiple (2) independent variables ‘weight’ and ‘volume’ and dependent variable ‘co2 emission’ linear regression model was created and co2 values were predicted from the model for any given value of weight and volume. The regression coefficient is a factor that helps to predict the unknown value by describing a relationship between dependent and independent variable.

4. polynomial_regression.ipynb

When there exists a polynomial regression (r squared) between the independent (x) and dependent (y) variable. R squared is 0 when no relationship and is 1 when 100% relationship. Building a polynomial regression model and calculating the r squared value, we could predict the value y value for given x value.

Unit 5: Jaccard Coefficient calculations

The table shows the pathological test results for three individuals

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	A
Mary	F	Y	N	P	A	P	N
Jim	M	Y	P	N	N	N	A

Calculate Jaccard coefficient for the following pairs

- (Jack, Mary)
- (Jack, Jim)
- (Jim, Mary)

Jaccard coefficient is used to calculate dis/similarity between asymmetric binary data.

Jaccard Coefficient = (f01+f10)/(f01+f10+f11)

For the above data, let Y=1, N=0 and P=1 and A=0

Jack 1 0 1 0 0 0

Mary 1 0 1 0 1 0

Jim 1 1 0 0 0 0

Jack and Mary, Jaccard coefficient = 1+0/1+0+2=0.33

Jack and Jim, Jaccard coefficient = 1+1/1+1+1=0.67

Jim and Mary, Jaccard coefficient = 2+1/2+1+1=0.75

Unit 7: Perceptron Activities

I listened to the lecturecast on ANN and run all the three python activities in Jupyter Notebook simple_perceptron.ipynb, perceptron_AND_operator.ipynb, multi-layer Perceptron.ipynb The function of perceptron, binary classification with OR, AND, XOR operator (Sigmoid function), Multi-layer perceptron with hidden layers, use of functions such as activation function, weight updating, error functions, gradient

decent, back propagation for error reduction, delta rule, calculating bias and applications of ANN were studied.

2. Google Colab: online browser tool allows working with TensorFlow, and to develop and train neural networks. I attended few tutorials.

Unit 8: Gradient Cost Function

1. I read the article by Mayo (2017) and studied the key points about updating Weights and definitions. Weights are associated with neuron connections. The error represents the difference between actual and predicted values and this is used at neurons to make weight adjustments, and the error is fed backward through the network after calculation called backpropagation of error. Gradient descent is used to determine optimal weights by acting as a guide when searching for a optimal value of the cost function and at the same time determine the error. Stochastic gradient descent is a randomisation of data sampling on which a single selection is used for error backpropagation (and weight updates)

   Reference:

   Mayo, M. (2017) Explained: Updating Weights with Gradient Descent & Backpropagation

2. The tutorial - gradient_descent_cost_function.ipynb. was run and observed the change in cost for change in iteration number and learning_rate.

S.No	Iteration number	Learning rate	cost
1	100	0.08	0.004120600119124239
2	50	0.08	0.06528102333197575
3	200	0.08	1.711074266712627e-05
4	100	0.07	0.008189416102046723
5	100	0.09	2640508310116.111

Selection of optimal value of learning rate and iteration number is important for finding the global minimum. Selecting lesser value of learning rate decreases the step size to find the global minimum but requires more number of iterations, while higher value of learning rate misses the global minimum.

Unit 9: CNN Model Activity

1. My thoughts on the ethical and social implications of CNN technology.

The article ‘Biased and wrong? Facial recognition tech in the dock’ by Matthew Wall is about the use of AI technology for facial recognition of identifying terrorists and criminals. While this technology is being used for good cause for the society, it is also evident that some groups of people are vulnerable to this technology, which is unethical. The training data has to represent all population proportionally to avoid bias until then this should not be deployed. Also the algorithms created for the model should be evaluated for its accuracy before implementation. This also necessitates the need for regulatory authority to scrutinize the basis used for building the facial

recognizing model and in turn there is a national need to set up standards for AI technology deployment in the society.

2. The python notebook Object Recognition.ipynb was run and different sections of the algorithm were reviewed. The input image for prediction was changed by changing the value of the variable - plt.imshow(x_test[x] - from 1 to16 and checked if the model predicts correctly.

Image no.	Object name	Predicted correctly?
1,2,15	ship	Yes
3, 10	airplane	Yes
4,5,7	frog	Yes
6, 9	automobile	Yes
8	cat	Yes
11,14	truck	Yes
12,16	dog	Yes
13	horse	Yes

Unit 11: Model Performance Measurement

The code model_Performance_Measurement.ipnyb was run in Jupyter notebook and the values AUC and R2 error were observed.

The receiver operating characteristic (ROC) is a graph plotted against true positive rate and false positive rate for different threshold values. The area under this ROC curve is called ROC AUC. A higher value of ROC AUC is expected for a good prediction meaning there is less false positive and more true positive prediction.

r2 score gives the correlation between the observed value and the predicted value in themodel. 100% is perfectly correlated with no variance and low values is low level of correlation..