Test 2 CSC 3304 Machine Learning

Question 1

1. Linear regression is a basic method of predictive analysis, and is widely used. These estimates of regression are used to describe the relation between a dependent variable and one or more independent variables. For example, predicting the revenue of a business for the following years.
2. Simple linear regression has one dependent and one independent variable while multiple linear regression has one dependent and one or more independent variables. For example, when calculating rent for rooms, in single linear regression model, we predict the rent based on the number of days rented. While in multiple regression model, we predict the rent based on the number of days and the size of the room.
3. Boosting is a machine learning algorithm that can convert a weak learning algortithm into a strong learning algorithm. Boosting is important in machine learning because it improves the accuracy of the outcome and reduce the error rate in the outcome. In solving face detection problems, AdaBoost can be used by applying many weak classifiers at a different scale on detecting the face. These classifiers will identify the eyes of the face. One weak classifier is not enough. Therefore, many classifiers are used to compute the weighted average of the weak classifiers, creating a better and more accurate outcome.
4. Example of the problem: Predicting the rate of cancer cases in Malaysia

   Process of machine learning algorithm:

   Multiple linear regression method. The dependent variable will be the cancer rate and the independent variables will be the states, gender, age and many more.

   Choosing the best algorithm:

   By looking at the datasets on the rate of cancer in Malaysia, it can be seen that the data has a relationship with other data which opens the path for executing a simple or multiple linear regression model for prediction. Other than that, regression models are the basic and most common method for prediction purposes in machine learning.

Evaluate performance of algorithm:

By comparing the predicted cancer rates with the actual cancer rates for a particular year, the accuracy of the results can be calculated from the comparison. If the accuracy percentage is high, the performance of the experiment is a success.

If the algorithm does not show good results, what is the next step:

The next step would be trying to execute a single linear regression model based on the relationship between cancer rates and states. This method is chosen because the probability of a more accurate outcome would possible if the variables are reduced and become simpler.

Question 2

x	**y**	**x^2**	**xy**
609	241	370881	146769
629	222	395641	139638
620	233	384400	144460
564	207	318096	116748
645	247	416025	159315
493	189	243049	93177
606	226	367236	136956
660	240	435600	158400
630	215	396900	135450
672	231	451584	155232
6128	**2251**	**3779412**	**1386145**

i. m = (1386145 – (6128 * 2251)) / ((10 * 3779412) – (6128^2))

= -51.3287

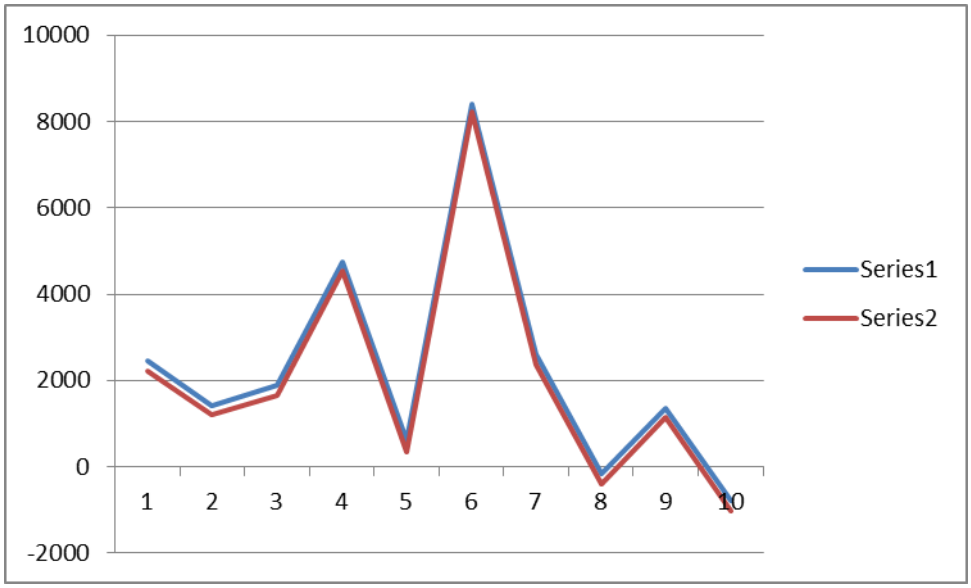
b = 2251 – ((-51.3287 * 6128) / 10)

= 33705.2274

x	y	y = -51.3287x + 33705.2274	error	Error^2
609	241	2446.0491	2205.0491	4862241.5334
629	222	1419.4751	1197.4751	1433946.6151
620	233	1881.4334	1648.4334	2717332.6742
564	207	4755.8406	4548.8406	20691950.8042
645	247	598.2159	351.2159	123352.6084
493	189	8400.1783	8211.1783	67423449.0744
606	226	2600.0352	2374.0352	5636043.1308
660	240	-171.7146	-411.7146	169508.9119
630	215	1368.1464	1153.1464	13229746.6198
672	231	-787.659	-1018.659	1037666.1583
6128	**2251**	**22510.0004**	**20259.0004**	**117325238.1**

ii. 117325238.1

iii.



Question 3

Positively labelled data: (3, 3), (3, -3), (-3, -3), (-3, 3)

Negatively labelled data: (1,1), (1, -1), (2, 1), (1, -2)

i. (10, 10), (7, 1), (1, 1), (1, 7)

(1,1), (1, -1), (2, 1), (1, -2)

S1 = (1, 1) S2 = (1, 1)

$3a1 + 3a2 = -1$

$3a1 + 3a2 = +1$

$a2 = -1 – 3a1$

$3a1 + 3(-1 – 3a1) = +1$

$3a1 – 3 – 9a1 = +1$

$-6a1 – 3 = +1$

$a1 = - 2/3$

$a2 = -1 – 3(-2/3)$

$a2 = 1$