Do the following tasks in your group

Try to do the following questions individually with help from the group if needed.

Q1. Logistic Regression

Instance	X	Y
1	2.4	1
2	124.2	1
3	-23.9	0
4	-401.5	0
5	53.7	0

If m is 5, b is 1 and alpha is 0.01, please compute

1. The Log loss
2. The new value of m
3. The new value of b

Solution

1. Given the values: m = 5, b = 1, alpha = 0.01

   Log loss is defined by below equation:
   Log Loss = -[Y * log(P(Y=1|X)) + (1 - Y) * log(1 - P(Y=1|X))]
   P(Y=1|X) = 1 / (1 + e^(-(m*X + b)))
   OR

$$-log[L(\theta)] = -\sum_{1}^{n} y * log[\sigma(\theta^T x^i)] + (1-y) * log(1-\sigma(\theta^T x^i)$$

For X=2.4, Y=1:

y(2.4) = 1 / (1 + e^-(5 * 2.4 + 1)) = 0.9896

Log Loss = -(1/5) * [1 * log(0.9896) + (1 - 1) * log(1 - 0.9896)] = 0.0102

For X=124.2, Y=1:

y(124.2) = 1 / (1 + e^-(5 * 124.2 + 1)) = 1.0

Log Loss = -(1/5) * [1 * log(1.0) + (1 - 1) * log(1 - 1.0)] = 0.0

For X=-23.9, Y=0:

y(-23.9) = 1 / (1 + e^-(5 * -23.9 + 1)) = 0.0

Log Loss = -(1/5) * [0 * log(0.0) + (1 - 0) * log(1 - 0.0)] = 0.0

For X=-401.5, Y=0:

y(-401.5) = 1 / (1 + e^-(5 * -401.5 + 1)) =0.0

Log Loss = -(1/5) * [0 * log(0.0) + (1 - 0) * log(1 - 0.0)]= 0.0

For X=53.7, Y=0:

y(53.7) = 1 / (1 + e^-(5 * 53.7 + 1)) = 1.0

Log Loss = -(1/5) * [0 * log(1.0) + (1 - 0) * log(1 - 1.0)] = 0.0

2. The new value of m:

m_{new} = m_{old} - alpha*gradient

OR

m_{new} = m_{old} - alpha * (sigmoid (mx) - y)*xj

Given the values: m = 5, b = 1, alpha = 0.01

For X=2.4, Y=1:

Gradient1: (0.9896 - 1) * 2.4 = -0.0224

For X=124.2, Y=1:

Gradient2: (1.0 - 1) * 124.2 = 0.0

For X=-23.9, Y=0:

Gradient3: (0.0 - 0) * (-23.9) = 0.0

For X=-401.5, Y=0:

Gradient4: (0.0 - 0) * (-401.5) = 0.0

For X=53.7, Y=0:

Gradient5: (0.0 - 0) * 53.7 = 0.0

Sum of above 5 gradient terms = -0.0224 + 0.0 + 0.0 + 0.0 + 0.0 = -0.0224

m_{new} = m_{old} - alpha * (1/n) * Sum of gradient terms

m_{new} = 5 - 0.01 * (1/5) * (-0.0224) = 4.983

3. New Value of 'b':

b_{new} = b - alpha * (1/n) * Σ[P(Y=1|X_i) - Y_i]

For X=2.4, Y=1:

b1: 0.9896 - 1 = -0.0104

For X=124.2, Y=1:

b2: 1.0 - 1 = 0.0

For X=-23.9, Y=0:

b3: 0.0 - 0 = 0.0

For X=-401.5, Y=0:

b4: 0.0 - 0 = 0.0

For X=53.7, Y=0:

b5: 1.0 - 0 = 1.0

Sum of update terms = -0.0104 + 0.0 + 0.0 + 0.0 + 1.0 = 0.9896

b_{new} = 1 - 0.01 * (1/5) * 0.9896 = 1.001

Q2. For the below data, compute

1. The odds of scoring more than 2 goals
2. The log odds of scoring less than 0 goals
3. The log odds of scoring 3 goals
4. The odds of scoring exactly 4 goals

Goals	Matches
0	272

1	84
2	57
3	33
4	18

Solution

1. Odds of scoring more than 2 goals:
   Odds = (Number of goals > 2) / (Total Matches)
   Calculate the odds for this case:
   Odds = (57 + 33 + 18) / (272 + 84 + 57 + 33 + 18)
   Odds = 108 / 464 = 0.2328
2. Log Odds of scoring less than 0 goals:

   The logarithm of zero or a negative number is not defined.

3. Log Odds of scoring 3 goals:
   Odds = (Number of goals) / (Total Matches)
   Odds = 33 / 464 = 0.0716
   Calculate the log odds for scoring 3 goals:
   Log Odds = ln(0.0716) = -2.6336
4. Odds of scoring exactly 4 goals:

   Odds = (Number of goals = 4) / (Total Matches)

   Odds = 18 / 464 = 0.0388

Q3 Goodness of Fit

A. For each of the below confusion matrices compute below details
   a. Accuracy
   b. Precision
   c. Recall
   d. Sensitivity
   e. Specificity
   f. F1 score

g. F2 score
h. F0.5 score
i. Null error rate
j. Balanced accuracy
k. Positive prevalence
l. Negative predictive value

– less commonly used

m. Miss rate
n. Fall out
o. False discovery rate
p. False omission rate
q. Positive likelihood ratio
r. Type I error rate
s. Type II error rate
t. Diagnostic odds ratio

1.

		Observed	
		+ve	-ve
Predicted	-ve	750	2000
	+ve	250	100

Solution

1. Accuracy:
   Accuracy = (TP + TN) / (TP + TN + FP + FN)
   Accuracy = (250 + 2000) / (250 + 2000 + 100 + 750) = 2250 / 3100 = 0.7258

2. Precision:
   Precision = TP / (TP + FP)
   Precision = 250 / (250 + 100) = 250 / 350 = 0.7143

3. Recall (Sensitivity):
   Recall = TP / (TP + FN)
   Recall = 250 / (250 + 750) = 250 / 1000 = 0.25

4. Specificity:
   Specificity = TN / (TN + FP)
   Specificity = 2000 / (2000 + 100) = 2000 / 2100 = 0.9524

5. F1 Score:
   F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
   F1 Score = 2 * (0.7143 * 0.25) / (0.7143 + 0.25) = 0.36

6. F2 Score:
   F2 Score = 5 * (Precision * Recall) / (4 * Precision + Recall)
   F2 Score = 5 * (0.7143 * 0.25) / (4 * 0.7143 + 0.25) = 0.2957

7. F0.5 Score:
   F0.5 Score = 1.25 * (Precision * Recall) / (0.25 * Precision + Recall)
   F0.5 Score = 1.25 * (0.7143 * 0.25) / (0.25 * 0.7143 + 0.25) = 0.7143

8. Null Error Rate:
   Null Error Rate = (TN + FP) / (TP + TN + FP + FN)
   Null Error Rate = (2000 + 100) / (250 + 2000 + 100 + 750) = 2100 / 3100 =0.6774

9. Balanced Accuracy:
   Balanced Accuracy = (Sensitivity + Specificity) / 2
   Balanced Accuracy = (0.25 + 0.9524) / 2 = 0.6012

10. Positive Prevalence:
    Positive Prevalence = TP / (TP + FN)
    Positive Prevalence = 250 / (250 + 750) = 250 / 1000 = 0.25

11. Negative Predictive Value:
    Negative Predictive Value = TN / (TN + FN)
    Negative Predictive Value = 2000 / (2000 + 750) = 2000 / 2750 = 0.7273

12. Miss Rate:
    Miss Rate = FN / (TP + FN)
    Miss Rate = 750 / (250 + 750) = 750 / 1000 = 0.75

13. Fall Out:
    Fall Out = 1 – Specificity
    Fall Out = 1 - 0.9524 = 0.047

14. False Discovery Rate:
    False Discovery Rate = FP / (TP + FP)
    False Discovery Rate = 100 / (250 + 100) = 100 / 350 = 0.2857

15. False Omission Rate:
    False Omission Rate = FN / (TN + FN)
    False Omission Rate = 750 / (2000 + 750) = 750 / 2750 = 0.2727

16. Positive Likelihood Ratio:
    Positive Likelihood Ratio = Sensitivity / (1 - Specificity)
    Positive Likelihood Ratio = 0.25 / (1 - 0.9524) = 5.2632

17. Type I Error Rate:
    Type I Error Rate = 1 – Specificity
    Type I Error Rate = 1 - 0.9524 = 0.0476

18. Type II Error Rate:
    Type II Error Rate = 1 – Sensitivity
    Type II Error Rate = 1 - 0.25 = 0.75

19. Diagnostic Odds Ratio:
    Diagnostic Odds Ratio = (Sensitivity * Specificity) / (FN / TN)
    Diagnostic Odds Ratio = (0.25 * 0.9524) / (750 / 2000) = 0.2393

2. While predicting Benign tumors

		Predicted	
		Malignant Tumors	Benign Tumors
Observed	Malignant Tumors	100	200
	Benign Tumors	50	5000

Solution:

1. Accuracy:
   Accuracy = (TP + TN) / (TP + TN + FP + FN)
   Accuracy = (50 + 5000) / (50 + 5000 + 200 + 100) = 5050 / 5250 = 0.9619

2. Precision:
   Precision = TP / (TP + FP)
   Precision = 50 / (50 + 200) = 50 / 250 = 0.2

3. Recall (Sensitivity):
   Recall = TP / (TP + FN)
   Recall = 50 / (50 + 100) = 50 / 150 = 0.3333

4. Specificity:
   Specificity = TN / (TN + FP)
   Specificity = 5000 / (5000 + 200) = 5000 / 5200 = 0.9615

5. F1 Score:
   F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
   F1 Score = 2 * (0.2 * 0.3333) / (0.2 + 0.3333) = 0.25

6. F2 Score:
   F2 Score = 5 * (Precision * Recall) / (4 * Precision + Recall)
   F2 Score = 5 * (0.2 * 0.3333) / (4 * 0.2 + 0.3333) = 0.28

7. F0.5 Score:
   F0.5 Score = 1.25 * (Precision * Recall) / (0.25 * Precision + Recall)
   F0.5 Score = 1.25 * (0.2 * 0.3333) / (0.25 * 0.2 + 0.3333) = 0.225

8. Null Error Rate:
   Null Error Rate = (TN + FP) / (TP + TN + FP + FN)
   Null Error Rate = (5000 + 200) / (50 + 5000 + 200 + 100) = 5200 / 5350 = 0.9710

9. Balanced Accuracy:
   Balanced Accuracy = (Sensitivity + Specificity) / 2
   Balanced Accuracy = (0.3333 + 0.9615) / 2 = 0.6474

10. Positive Prevalence:
    Positive Prevalence = TP / (TP + FN)
    Positive Prevalence = 50 / (50 + 100) = 50 / 150 = 0.3333

11. Negative Predictive Value:
    Negative Predictive Value = TN / (TN + FN)
    Negative Predictive Value = 5000 / (5000 + 100) = 5000 / 5100 = 0.9804

12. Miss Rate:
    Miss Rate = FN / (TP + FN)
    Miss Rate = 100 / (50 + 100) = 100 / 150 = 0.6667

13. Fall Out:
    Fall Out = 1 – Specificity
    Fall Out = 1 - 0.9615 = 0.0385

14. False Discovery Rate:
    False Discovery Rate = FP / (TP + FP)
    False Discovery Rate = 200 / (50 + 200) = 200 / 250 = 0.8

15. False Omission Rate:
    False Omission Rate = FN / (TN + FN)

False Omission Rate = 100 / (5000 + 100) = 100 / 5100 = 0.0196

16. Positive Likelihood Ratio:
    Positive Likelihood Ratio = Sensitivity / (1 - Specificity)
    Positive Likelihood Ratio = 0.3333 / (1 - 0.9615) = 8.9999

17. Type I Error Rate:
    Type I Error Rate = 1 – Specificity
    Type I Error Rate = 1 - 0.9615 = 0.0385

18. Type II Error Rate:
    Type II Error Rate = 1 – Sensitivity
    Type II Error Rate = 1 - 0.3333 = 0.6667

19. Diagnostic Odds Ratio:
    Diagnostic Odds Ratio = (Sensitivity * Specificity) / (FN / TN)
    Diagnostic Odds Ratio = (0.3333 * 0.9615) / (100 / 5000) = 32.04

3. While detecting Negative Sentiment

		Observed	
		Negative	Non-negative
Predicted	Negative	626	574
	Non-negative	274	326

B. For the below data compute the SSR, MSR, RMSE and MAE for the given model M0

Model M0 has the hypothesis function
y' = 2.7x1 - 1.6x2 + 0.87

x1	x2	y'	y
53.7	18		59
28.5	17		56
21.5	12		41

-12	-4		-7
0.25	-1.5		0.5
	-12		-31
10.5			26
8.7	287		17

C. For the below data trying to identify Fraud, determine

a. TP
b. FP
c. TN
d. FN
e. Negative Prevalence

	Actual	Predicted
	Fraud	Fraud
	Non-fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Fraud	Non-Fraud
	Non-Fraud	Non-Fraud

D. For the below model perf determine the area under the RoC curve

TPr = eFPr+FPr3-3/2FPr-1/2