IS-733 CLASSWORK – Logistic Regression and SVM (03-10-2025) – UB01976

Task1a: Interpreting Logistic Regression model
Given a logistic regression model

$$\ln\left(\frac{p}{1-p}\right) = -3 + 0.8 \times \text{Hours\_Studied} + 1.5 \times \text{Review\_Session}$$

Answer the following questions:
(you may use the provided "logistic regression" notebook and AI assistant.)
a. Thomas studied for two hours and did not attend the review session. What is his (1) log odds, (2) odds, and (3) likelihood of passing the exam?

a Ans: - Thomas has studied for two hours and did not attend the review session. So, this gives us the following information;

Hours_Studied = **2**

Review_Session = **0 (Not attended)**

(1) Log odds = ln (p/1-p) = -3 + 0.8 X 2 + 1.5 X 0

Log odds = -3 + 1.6 + 0

Log odds = -1.4

(2) Odds = exp (Log odds) = exp (-1.4)

Odds ≈ 0.2466

(3) Likelihood of passing the exam = Odds / (1+Odds) = 0.2466 / (1+0.2466)

Likelihood of passing the exam ≈ 0.198

b. If Thomas goes to the review session, what is the updated 1) log_odds, (2) odds, and (3) likelihood of passing the exam?

b Ans: - Thomas went to the review session and studied for two hours. So, this gives us the following information;

Hours_Studied = **2**

Review_Session = **1 (Attended)**

(1) Log odds = ln (p/1-p) = -3 + 0.8 X 2 + 1.5 X 1

Log odds = -3 + 1.6 + 1.5

Log odds = 0.1

(2) Odds = exp (Log odds) = exp (0.1)

Odds ≈ 1.105

(3) Likelihood of passing the exam = Odds / (1+Odds) = 1.105 / (1+1.105)

Likelihood of passing the exam ≈ 0.525

c. If Thomas studied more or less hours, would the answer change?

c Ans: - Yes, the answer would have changed if Thomas studied more or less hours. As the hours_studied parameter is directly included in the formula, changing its value will alter the values of log_odds, odds, and livelihood, as a result the answers would have been impacted. Hence, Thomas studying more or less hours will change the answer.

d. How would you interpret the coefficient of review_session (1.5) from the above experiment?

d Ans: - The co-efficient of review_session (1.5), is the one which indicates the change in the log_odds of exam pass, assuming participation in review session. This means that attending the review_session will increase the log_odds of passing the exam by 1.5 times, considering hours_studies constant. On the odds scale, review session attendance increases the odds of passing by exp (1.5) **≈ 4.48,** remaining all being equal.

e. Using similar reasoning, how would you interpret the coefficient of hours_studied (0.8)

e Ans: - The co-efficient of hours_studied (0.8) suggests that each additional hour studied increases the log_odds of passing the exam by 0.8, holding the review_session constant. The odds of passing the exam are exp (0.8) **≈ 2.23** times higher for each additional hour studied, given the same review session attendance.

f. How would you interpret the intercept?

f Ans: - The intercept (-3) is the log_odds of passing the exam when hours_studied and review_session both are zero. Which means the log_odds for student who did not study or did not attend any review session. So, in this case the odds are exp (-3) **≈ 0.0498** and the probability of passing is 0.0498 / (1+0.0498) **≈ 0.047,** taking hours_studied and review_session as zero. So, by this we can conclude that the student passing the exam is really low or equivalent to zero.

g. For someone who studied 8 hours, would you recommend him/her to attend the review session?

g Ans: - Let us consider, 2 cases here: - i.e. First one is when the **student has studied for 8 hours and did not attend the review session** and then **studied for 8 hours and attended the review session**.

Case1: - Not Attended the review session
(1) Log odds = ln (p/1-p) = -3 + 0.8 X 8 + 1.5 X 0

Log odds = -3 + 6.4 + 0

Log odds = 3.4

(2) Odds = exp (Log odds) = exp (3.4)

Odds ≈ 29.964

(3) Likelihood of passing the exam = Odds / (1+Odds) = 29.964 / (1+29.964)

Likelihood of passing the exam ≈ 0.9677

Case2: - Attended the review session
(1) Log odds = ln (p/1-p) = -3 + 0.8 X 8 + 1.5 X 1

Log odds = -3 + 6.4 + 1.5

Log odds = 4.9

(2) Odds = exp (Log odds) = exp (4.9)

Odds ≈ 134.289

(3) Likelihood of passing the exam = Odds / (1+Odds) = 134.289 / (1+134.289)

Likelihood of passing the exam ≈ 0.9926

So, this means that the student who studied for 8 hours has likelihood of passing the exam of **0.9677 (High)** even though he/she did not attend the review_session. Hence, the student will be able to pass the exam. But if the student attends the review session, then there is slight increase in likelihood of passing from **0.9677 to 0.9926** which is marginal but can increase the likelihood of passing. So, it is recommended for the student to attend the review session even though he/she studies for 8 hours.

h. What type of students seems to benefit most from the review session?
h Ans: - Mainly students **who study for less amount of time or spend fewer hours** on studying will be most beneficial from the review session. Cause for the students who have spent lot of time on hours_studied, for them the marginal benefit from review session will gradually diminish. But for those who don't spend much time studying, for them a review session will be really helpful and improve their **likelihood of passing the exam**.

Task 1b: Build a logistic regression model

Using the dataset “student_data.csv,” write code to (1) create a visualization of the data, (2) fit a model using logistic regression, (3) output model coefficients and performance metrics such as accuracy and AUC and ROC

1b Ans: - Please find the solution code for the above program in the following GitHub link below:

https://github.com/UB01976/is7332025/blob/main/data-mining-project-repo/03102025_CW/Classwork_03102025.ipynb

Task 2: Understanding and Prevent Overfitting in the context of SVM

Write code to fit a Support Vector Machine model using (1) linear kernel and (2) RBF kernel. For the RBF kernel, use grid search to find the best gamma parameter using k-fold cross-validation.

2 Ans: - Please find the solution code for the above SVM model in the following GitHub link below:

https://github.com/UB01976/is7332025/blob/main/data-mining-project-repo/03102025_CW/Classwork_03102025.ipynb