UNIT 4 ASSIGNMENT

Introduction to Linear Models

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the unit. You’ve practiced these concepts in the coding activities, exercises and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Unit 4 Written Portion

# Logistic Regression

Answer the questions below about linear models.

## Questions:

1. What is a linear model? What are the advantages and disadvantages of linear models?

|  |
| --- |
| A linear model is one that models the relationship between a dependent variable and one or more independent variables. It creates a linear relationship between the two. An advantage of the model is its simplicity because the relationship only relies on one exponent. It is also, therefore, easy to interpret. This model is computationally efficient and can handle large datasets. A disadvantage is its fixed linear form, as it assumes only a linear relationship between the variables, it will be inaccurate if the relationship is non-linear. The model is also sensitive to outliers as one value far from the set will highly offset the estimated coefficients. It also assumes that the observations are independent of one another, further restricting its use and applicability. |

1. What type of supervised learning problem is logistic regression best suited for? Give an example of a problem you would use a logistic regression model for. Explain what you are trying to predict.

|  |
| --- |
| Logistic regression is best suited for binary classification problems to predict the probability of an observation being in one of two classes. An example where a logistic regression model would be used is whether or not a bank gives a customer a loan. The bank has information on the customer such as their credit score, income, employment, and age. After considering all of these variables, a result will be given of whether or not to give a loan—this is a yes or no (binary/1 or 0) answer. |

1. Describe the training phase of a logistic regression model: explain the intuition behind using gradient descent algorithm and the use of loss functions.

|  |
| --- |
| Gradient descent is used to find the optimal values of the model’s coefficients. The goal is to adjust the coefficients as much as possible to minimize the loss function. The coefficients are first started with initial values and then the algorithm predicts probabilities for each observation in the training data using those coefficients. These probabilities are compared to the target values and the difference between them is used to alter and update the coefficients.  Loss functions measure the difference between the predicted probabilities and the actual target values. During the training phase, the gradient of the loss function is found and represents the direction and magnitude of the loss function. The gradient descent algorithm then takes steps using the negative gradient multiplied by the learning rate. Updating coefficients using the gradient descent algorithm and adjusting them using the direction of the largest change, logistic regression converges to values that minimize the loss function and create the best fit to the training data. |

1. Explain the purpose of using regularization when training a logistic regression model.

|  |
| --- |
| Regularization when training a logistic regression model is necessary to control the complexity of the model, prevent overfitting, improve generalization, make the data easily interpretable, and also more stable. Regularization adds a penalty to the loss function, encouraging the model to find a balance between minimizing the training error and reducing the model’s complexity. Penalizing large coefficients prevents them from taking extreme values and therefore controls the model’s complexity and makes it less sensitive to individual data points or excess noise in the training data. All of this leads to a more stable, interpretable, and more accurate model. |

1. Explain which linear model and accompanying loss function you would use for a classification problem and for a regression problem.

|  |
| --- |
| For a classification problem, I would use logistic regression which models the relationship between independent variables and their probability of belonging to a specific class. I would use the logistic loss function to measure the difference between the predicted probabilities and the actual class labels. It penalizes larger errors more heavily, allowing for the model to assign higher probabilities to the correct class and lower ones to the incorrect class.  For a regression problem, a linear regression model should be used. It models the relationship between the independent variables and the constant dependent variable. Then, it predicts the value of the dependent variable based on a linear combination of the independent variables. A mean squared error function (MSE) should be used in this case. It calculates the average squared difference between predicted values and the actual values of the dependent variable. The linear regression minimizes the overall squared error between the predicted values and the true values. |

*To submit this assignment, please refer to the instructions in the course*.