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 a statistical model that assumes a linear relationship between the input variables (also known as independent variables or features) and the output variable (also known as the dependent variable or target). It is represented by a linear equation in which the output variable is a linear combination of the input variables, possibly with some added noise.  Advantages of linear models:   * Simplicity: Linear models are straightforward and easy to understand. The linear relationship allows for intuitive interpretation of the coefficients, as they represent the change in the output variable for a unit change in the corresponding input variable. * Computational efficiency: Linear models are computationally efficient and can handle large datasets with many features. * Inference: Linear models provide statistical inference, allowing you to assess the significance and confidence intervals of the coefficients. This information helps in understanding the relationships between variables. * Feature importance: The coefficients of linear models can indicate the relative importance of different features in influencing the output variable.   Disadvantages of linear models:   * Limited flexibility: Linear models assume a linear relationship between the variables, which may not be suitable for complex or nonlinear relationships. They may fail to capture more intricate patterns in the data. * Assumptions: Linear models have certain assumptions, such as linearity, independence of errors, and homoscedasticity (constant variance of errors). Violation of these assumptions can lead to inaccurate results. * Feature engineering: Linear models heavily rely on feature engineering to capture nonlinear relationships. Transforming or creating new features might be necessary to improve their performance. * Outliers: Linear models can be sensitive to outliers, which can disproportionately influence the estimated coefficients and affect the model's overall performance. |

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 in supervised learning. In binary classification, the goal is to predict one of two possible classes or outcomes based on input variables.  Here's an example:  Problem: Predicting Email Spam  Suppose you have a dataset of emails, and your task is to build a model that can classify whether an email is spam or not based on its content and other relevant features.  In this case, you have a binary outcome: either an email is spam (class 1) or it is not spam (class 0). The input variables or features could include the presence or absence of certain keywords, the length of the email, the frequency of capital letters, etc.  By training a logistic regression model on this dataset, you aim to create a model that can analyze the email content and features and predict the probability of an email being spam. The output of the logistic regression model would be a probability value between 0 and 1, representing the likelihood of an email being spam. You can then define a threshold (e.g., 0.5) to classify emails as spam or not spam based on the predicted probabilities.  Logistic regression is well-suited for this problem because it can model the relationship between the input features and the probability of an email being spam. It provides a probabilistic framework for classification and can handle binary outcomes efficiently. |

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

|  |
| --- |
| The training phase of a logistic regression model involves estimating the parameters (coefficients) that define the relationship between the input variables and the probability of the binary outcome. This estimation is achieved through an optimization process using the gradient descent algorithm and a suitable loss function.  Here's an overview of the training phase:   * Initialization: Initially, the model's parameters, including the coefficients, are initialized with some values (often randomly or with zeros). * Calculating the Loss: The loss function measures the discrepancy between the predicted probabilities and the actual binary outcomes in the training data. In logistic regression, a common loss function is the log-loss or cross-entropy loss. It quantifies the difference between the predicted probabilities and the true class labels. * Gradient Descent: The gradient descent algorithm is used to update the parameters iteratively in order to minimize the loss function. The intuition behind gradient descent is to iteratively adjust the parameters in the direction of steepest descent of the loss function, effectively "descending" the loss surface towards a minimum. * Compute Gradient: The gradient of the loss function with respect to each parameter is computed. This gradient represents the direction and magnitude of the steepest ascent of the loss function. * Update Parameters: The parameters are updated by subtracting a fraction of the gradient from the current parameter values. This fraction is determined by the learning rate, which controls the size of the parameter updates in each iteration. * Iterative Update: Steps 2 and 3 are repeated iteratively until convergence or a predetermined number of iterations. Convergence is typically achieved when the change in the loss or parameter values falls below a certain threshold.   The intuition behind using gradient descent in logistic regression is to find the optimal set of parameters that minimizes the loss function and maximizes the likelihood of the observed data. By iteratively updating the parameters based on the gradients, the algorithm adjusts the coefficients to better fit the training data, thereby improving the model's ability to predict the binary outcome.  The choice of the loss function, such as the log-loss or cross-entropy loss, is motivated by the probabilistic nature of logistic regression. These loss functions are designed to measure the discrepancy between the predicted probabilities and the true class labels. By minimizing the loss, the model learns to assign higher probabilities to the correct class and lower probabilities to the incorrect class. |
|  |

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

|  |
| --- |
| The purpose of using regularization when training a logistic regression model is to prevent overfitting and improve the model's generalization performance on unseen data. Regularization helps to control the complexity of the model by adding a penalty term to the loss function during training. |

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, one suitable linear model is logistic regression. Logistic regression models the relationship between the input variables and the probability of a binary outcome, making it well-suited for binary classification tasks.  Accompanying Loss Function:  The accompanying loss function commonly used with logistic regression is the log-loss or cross-entropy loss. This loss function measures the discrepancy between the predicted probabilities and the true class labels. It penalizes models that assign high probabilities to the incorrect class and encourages models to assign higher probabilities to the correct class.  For a Regression Problem:  For a regression problem, a suitable linear model is linear regression. Linear regression models the relationship between the input variables and a continuous numerical output variable, making it appropriate for predicting quantitative values.  Accompanying Loss Function:  The accompanying loss function used with linear regression is the mean squared error (MSE) loss or the sum of squared errors (SSE) loss. This loss function calculates the average squared difference between the predicted values and the true continuous target values. It penalizes larger prediction errors more severely, and the goal is to minimize this loss function by adjusting the model's parameters. |

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