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 model that is represented by a single equation, unlike KNN and decision trees. The advantages of linear models are lower complexity, faster to train and predict, good for small data sets with many features, and it’s easier to explain the model to stakeholders. The disadvantages of linear models are it’s not suitable for large data sets, and that they work only for data that has linearity, which is not always the case. This is why we need decision trees and KNN. |

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, where there are only two possible outcomes for the prediction/label– True or False – like medical predictions. For example, I would use a logistic regression model to predict if a Twitter account is a bot or not. |

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

|  |
| --- |
| During the training phase of a logistic regression model, we use a gradient descent algorithm. Why? Well, in general, the gradient descent function finds the minimum or maximum of a function, so a gradient descent algorithm used to minimize log loss function to find the optimal logistic regression model with low errors. We figure out if the model has low loss via loss functions like the log loss function. Therefore, during training, we obtain the best logistic regression model by computing the gradient descent and using that information to update the model parameters until the loss function is minimized. |

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 minimize loss and, also, avoid overfitting (not generalizing new data) by minimizing the model’s complexity. If we only concentrated on low training loss, then we have an overfit model, so we must use regularization. |

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 a logistic regression model and the common loss function, log loss , because logistic regression predicts the probability of a categorical label (the label), which is the definition of a classification problem.  For a regression problem, I would use a linear regression model and MSE (mean squared errors) loss function because we are trying to predict a continuous label. |

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