UNIT 3 ASSIGNMENT

Understanding the Mechanics of   
ML Algorithms

## 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; let’s now formulate your programming into well-reasoned 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 3 Written Portion

# Building and Evaluating a Model

Answer the questions below about building and evaluating your models using algorithms such as decision trees and k-nearest neighbors.

## Questions:

1. What are the advantages and disadvantages of decision trees?

|  |
| --- |
| One advantage of decision trees is how easy they are to interpret. They make non-linear relationships easy to see and show all possible relations in order. Disadvantages of decision trees are that they are prone to overfitting, are sensitive to small changes in the data, and are therefore unstable. Because of this, predictions can vary greatly. In addition, they are biased towards features with larger numbers of categories which results in biased decisions. |

1. What are the advantages and disadvantages of k-nearest neighbors?

|  |
| --- |
| Some advantages of k-nearest neighbors are that KNN is simple in the fact that similar instances will have similar class labels. Also, it does not require a training phase and does not make assumptions about the underlying data distribution. Some disadvantages are that as the number of training instances increases, so does its computational cost. Also, since it relies on measurements of distance to find the nearest neighbor, if features have different scales the ones with larger values can overshadow the influence of smaller features. Similarly, as the number of features increases, the distance between the instances becomes less meaningful. KNN is powerful for small datasets with low dimensions but can face challenges with large datasets, high-dimensional spaces, and imbalanced data distributions. |

1. Explain the difference between k-nearest neighbors and decision trees. When would you decide to use one over the other?

|  |
| --- |
| Both k-nearest neighbors and decision trees are non-parametric and can be used for regression and classification problems. Some differences are that decision trees are better for capturing complex relationships and handling diverse data types but can be computationally expensive. Decision trees work better for structured data and can create interpretable models, however, they struggle with complex relationships. Because of this, the decision to use one over the other relies on how big the data set is, and whether the data is diverse or structured. |

1. What are hyperparameters? List some hyperparameters in k-nearest neighbors and decision trees.

|  |
| --- |
| Hyperparameters are parameters that aren’t learned from the data from the training process but instead are set before the training starts. Some hyperparameters in k-nearest neighbors are n\_neighbors, algorithm, and metric. Some hyperparameters in decision trees are max\_depth, min\_samples\_split, and min\_samples\_leaf. |

1. What is overfitting? How can you avoid overfitting? Give examples using a model discussed   
   so far.

|  |
| --- |
| Overfitting is when the model learns the training data too well and starts to memorize data or reduce the noise in it. Because of this, the model struggles with new data because it became too specialized in the original data. To avoid this cross-validation, regularization, feature selection, or early stopping can be used. For example, in the Unit 3 assignment, different values of max\_depth, the hyperparameter, were used to find the optimal depth to balance the complexity of the model and its ability to generalize. Picking an appropriate max\_depth, therefore, allowed us to avoid overfitting and allow for the model to handle new data well. |

1. What is the purpose of splitting data into different sets?

|  |
| --- |
| Splitting data into different sets allows you to evaluate the models performance on all data and prevent overfitting. The training set is used to train the machine learning model, the validation set is used to finetune the model’s hyperparameters and evaluate its performance. Then, the test set is used to check the models final performance and estimate its ability to generalize the full data. |

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