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?

|  |
| --- |
| The advantages of decision trees are that it can be used to solve both regression and classification problems. It also doesn’t require scaling of the data. To visualize our decision trees, it’s also easy to follow along by its nodes and directed arrows.  The disadvantages of decision trees are that it can be complicated especially if we have too many features or too much data. Decision trees also generally need overfitting of data which isn’t ideal as it leads to high variance. |

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

|  |
| --- |
| The advantages of KNN is that there’s no training period. It makes data available during prediction time. So we just start at the point we’re predicting and look for the K nearest points using a distance formula.  The disadvantages of KNN is that there’s the curse of dimensionality, in which the higher the dimension, the more features, and this leads to the examples becoming less similar to a point where finding the closest neighbors to predict the test data is no longer feasible. |

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

|  |
| --- |
| The difference between KNN and decision trees is that in KNN, it uses neighbors of a certain data point value and the distance formula and make a prediction by looking at the class labels of the neighbors. Meanwhile in decision trees you need to provide a numeric or nominal vector input to predict a class. I would use KNN when I have lower dimensions, and I would use decision trees for non-linear data sets. |

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

|  |
| --- |
| Hyperparameters essentially declares the mechanics of the model such as its complexity, and how the model is trained such as how fast it learns. Some hyperparameters in KNN is size of neighborhood, and in decision trees it’s the depth of the tree. |

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

|  |
| --- |
| Overfitting can be defined as when the model performs well on training data but generalizes poorly on new incoming data. This typically occurs when the model is too complex. For example, if we have a model that we want to recognize trees, and we feed the model to recognize characters by inserting palm, oak, and spruce tree, and a car to indicate it’s not a tree, it can be too complex and the model will end up learning other characteristics (such as lead shape) that will end up identifying actual tree objects as not a tree. To avoid overfitting, we can evaluate our model’s ability to generalize well to new data. We can split our original data into 3 partitions: training set, validation set, and test set. To determine how to partition our data set. We can either use loss functions or evaluation metrics. |

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

|  |
| --- |
| The purpose of splitting data into different sets is to avoid overfitting. It’s essential to evaluate our model’s ability to generalize well to new data, as it is the goal of machine learning to make predictions. |

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