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?

|  |
| --- |
| Advantage: is able to handle many features without the model training too slowly  Disadvantage: could be prone to overfitting if the tree is too deep and becomes too complex |

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

|  |
| --- |
| Disadvantage: it can be very time consuming once we have many features  Advantage: simple and intuitive to understand. There’s also no training involved |

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

|  |
| --- |
| K-nearest Neighbors: to find the label of an unseen data point, we look at the labeled data points around it, and select the label that appears the most. For decision trees, we make our decision for our unseen data point based on a series of binary questions. The way these questions are asked is to reduce the entropy or uncertainty a specific feature creates, so based on some given features, we can clearly tell what the label of that data point would be. I would use KNN if I don’t have many features in my data matrix, and would use decision trees if I did have many features |

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

|  |
| --- |
| Hyperparameters are the properties we are able to adjust when building our model. For KNN, a hyperparameter would be K, and for decision trees it would be the max depth of our tree |

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

|  |
| --- |
| Overfitting is when our model fits too well to the data we gave it. So much so that it does poorly on unseen data and is unable to generalize. To avoid overfitting, adjust hyperparameters and also add a validation set so when we adjust our model, we aren’t affecting the testing dataset. |

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

|  |
| --- |
| So we can test whether our model does well in learning the data but also applying its knowledge on unseen data. Training set to train the model so it can learn the patterns. Validation set is for when we want to adjust our model more than once and we want to see if it improves. Testing set to apply our model on data it hasn’t seen before |

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