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?

|  |
| --- |
| Advantages: very simple, non-parametric method, do not require data scaling, less sensitive to outliers  Disadvantage: prone to overfitting with noisy data, bias-variance tradeoff |

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

|  |
| --- |
| Advantage: very simple, non-parametric method, used for both classification and regression  Disadvantage: need to normalize data and handle outliers, expensive for a large data set, curse of dimensionality |

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

|  |
| --- |
| KNN: instance-based learning, memory intensive, no training phase, faster(need to calculate all data points), sensitive to outliers/noisy data  Decision tree: model-based learning, memory efficient, training phase, slower(traverse the tree), less sensitive to outliers/noisy data  I will choose KNN when a data set and dimensions are not too large and the prediction needs to be made based on similarity. On the other hand, I will choose Decision tree when a data set is large, handle categorical data, and interpretability is important. |

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

|  |
| --- |
| Hyperparameters are the setting of an algorithm that can be adjusted to optimize performance. For KNN, it has K(neighbor count) and distance function. For Decision tree, it has the number of depth, split, and leaves. |

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

|  |
| --- |
| Overfitting is that the prediction is too specific to the training data only, so that the prediction for the new data is poor. To avoid overfitting in KNN, we need to find appropriate value K. Low K value prone to be overfitting and sensitive(specific) to noisy data. For Decision tree, we need to find appropriate number of depth, split, and leaves. Too large trees(high depth) prone to be overfitting and higher complexity. |

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

|  |
| --- |
| To ensure that the model is trained, validated, and tested properly for a finite data set. Since we do not have infinite data set, we need to split data and repeat training and validating phase to improve prediction. Test data set is used only once to complete an evaluation of the model. |

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