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 one advantage of decision trees is that you can trace the steps to get to a prediction visually, and another advantage is that they can be used for both regression and classification problems. The disadvantages of decision trees are that the tree depth can grow fast which can lead to overfitting and vice versa can happen which leads to underfitting. |

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

|  |
| --- |
| The disadvantages of k-nearest neighbors is that it can not do well with data that contains a large number of features (high-dimensional data), and that it takes time to test because each test point and training point must have its distance computed Thus an advantage of KNN is that it works well when data points are close together and can be used for regression and classification problems. |

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

|  |
| --- |
| One distinct difference between KNN and decision trees is that KNN is better used for low-dimensional data while decision trees seem more versatile with any amount of data. In addition KNN has slower time training, so I would choose decision trees for situations when data is higher than low-dimensional and KNN when data is low-dimensional. |

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

|  |
| --- |
| Hyperparameters are properties in a model that control its behavior and complexity and affect underfitting and overfitting. Some hyperparameters for decision trees are the tree’s depth, number of samples per leaf, and information gain. Some hyperparameters to k-near neighbors are the distance function and the number of neighboring examples. |

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

|  |
| --- |
| Overfitting is when a model is great at predicting based on the training data, but fails to predict correctly when given new data (low training error and poor generalization). You can avoid overfitting by splitting the full data set into tests, training, and validation. Like for the KNN activity, we split our data set in tests and training data. We also checked our model’s prediction with accuracy scores which were not bad, depending on the hyperparameters. |

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

|  |
| --- |
| We split data into different sets in order to properly evaluate our model’s performance and to prevent overfitting. |

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