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.

*Once all parts are complete, submit your assignment on the Unit 3 Assignment page online. Information about the grading rubric is available on any of the 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?

| Decision Trees are both simple and powerful. They are easy to train and interpret, and are flexible enough to capture many relationships in data. At the same time, this flexibility can lead to overfitting, so care needs to be taken when training them. |
| --- |

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

| KNNs are also simple and powerful. They can fit any arbitrary relationship between X and Y and technically don’t require training. This flexibility means they are also prone to overfitting. Additionally, KNNs are sensitive to data scale and the number of features so pre-processing is usually required. |
| --- |

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

| Decision Trees fit a graph based tree structure by iteratively segmenting data to reduce the entropy of the label. KNNs don’t “fit” the data at all. KNNs represent instance based learning where inference is done by finding nearby examples and computing the average label over them. Deciding between them is an empirical question. One should first determine if there are any computational or business constraints that limit the use of either. If not, then it is a matter of performance, which can be tested. |
| --- |

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

| Hyper-parameters are inputs to the model specification (or set up) that control how the algorithm will be configured during training. For instance, with KNN the hyper-parameter is “K”, or the number of neighbors used for inference. For Decision Trees, we have “max\_depth” and “min\_samples\_split”, which dictate how large the tree can grow and how many samples are required before a split can be done. |
| --- |

1. What is overfitting? How can you avoid overfitting?

| Overfitting is the phenomenon where a model treats noise in the data as if it were the true signal. The tell-tale sign of overfitting is when a model performs on training data but poorly on out-of-sample data. One can avoid it with multiple techniques: 1) using more data, 2) user fewer features, 3) using simpler models, 4) doing hyper-parameter optimization with a validation set. Ideally one would use all of these methods. |
| --- |

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

| The purpose is to do model selection and evaluation in an effort to avoid over fitting. We generally want to train and validate or evaluate models on different data (the splits). The non-training data represents “new” data that we can use to determine if our model will generalize well. |
| --- |

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