UNIT 6 ASSIGNMENT

Special Topics: Ensemble Models and Unsupervised Learning

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought 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.*

Week 6 Written Portion

# Implementing Ensemble Models and Unsupervised Clustering

Answer the questions below about ensemble models and unsupervised learning.

1. Explain ensemble modeling. What is the advantage of using this technique?

|  |
| --- |
| Ensemble modeling is a way to approach building a model to have low bias and low variance by combining different models so that we can average the errors from individual models. The advantage of using this technique is that it works for both classification and regression problems and we end up with the best model for the ML problem. |

1. Explain what bias and variance are, along with the bias-variance tradeoff.

|  |
| --- |
| Bias is the difference between the average prediction of our model and the correct prediction, and variance is the measure of how consistent the model’s predictions are if it is trained on different training dataset. Bias-variance tradeoff means to find the right balance of bias and variance for the best model. |

1. Explain the differences among the ensemble methods bagging, boosting, and stacking.

|  |
| --- |
| Unlike bagging and boosting, stacking does not have a specific method attached; it is a combination of supervising learning algorithms. Boosting is the most mathematically rigorous of the three methods where we are building models over multiple iterations in order to reduce the prediction error. Boosting is common for gradient-boosted decision trees while bagging is for random forest (decision trees). Bagging is short for bootstrapping aggregating, a method where we generate multiple models from the same data and then average the model predictions. |

1. Explain the random forest algorithm and how it relates to decision trees and bagging.

|  |
| --- |
| The random forest algorithm can be used for both classification and regression problems and it is a set of decision trees because it generalizes data better than a single decision tree. We implement the random forest algorithm with the ensemble method called bagging. Bagging helps us use a subset of features as bootstrap data and we build multiple decision trees from the bootstrap data. |

1. What’s the difference between gradient boosting decision trees and random forest?

|  |
| --- |
| Gradient boosting decision trees consist of shallow individual decision trees, it trains on residuals or losses of the previous tree, and uses all features of a dataset. Instead of shallow trees, random forests use deep trees. Random forests also differs from GBDT in that it trains on randomly sampled data, and use a subset of features. |

1. What’s the difference between supervised and unsupervised learning?

|  |
| --- |
| Supervised learning is a class of ML problems where labeled data is available and that helps us when we use algorithms to see what is associated with labels and the data values. On the other hand, unsupervised learning is a class of ML problems where labeled data is unavailable and so we use algorithms to discover patterns within the dataset. |

1. Give an example of an ML problem where you would use unsupervised clustering.

|  |
| --- |
| For an ML problem where I wanted to be able to detect images of mystical beasts, then I would use unsupervised clustering. The algorithm would have to learn characteristics like multiple tails, more than 4 legs, colorful fur, and multiple heads are traits of mystical beasts. Then the algorithm would also have to cluster mystical beasts from images of actual animals. |

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