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 process where in we use multiple learning algorithms to obtain higher predictive performance, than any algorithm could achieve by itself.  It is an advantage to use this technique as it makes our ML model more accurate by capturing a more comprehensive understanding of our data. |

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

|  |
| --- |
| Bias is the model’s prediction error due to its assumptions. For example, a linear model might have high bias if the problem is inherently non-linear.  Variance represents the randomness in a model's predictions. A model with high variance might make very different predictions on similar data points depending on the specific training data it was exposed to.  The bias variance tradeoff is a concept that deals with making a model more complex (and thereby increasing its flexibility), by reducing bias (fitting the training data better), but risk increasing variance (overfitting the data). Conversely, a simpler model has the opposite drawbacks: underfitting of the data! |

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

|  |
| --- |
| Stacking is an ensemble method that involves using several different learning algorithms together (the algorithms can be different). By aggregating the models together, it improves the overall accuracy of the model.  Though it is more complex than the other two methods, it has the chances for highest accuracy.  Boosting is a method that focuses on reducing bias. An example of boosting is through using Gradient Boosted Decision Trees. it is functionally trained not on samples of training data, but on the residual (error) from the predicted data. Therefore, by focusing on the error, it minimizes bias.  Bagging is a method that focuses on reducing variance; an example implementation is through random forests, which are an ensemble of decision trees that are high in depth, trained on subsets of features and randomly sampled data. Different models see different data (it is randomly sampled training data!), and so the predictive error is minimized in this sense. |

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

|  |
| --- |
| Random forest combines decision trees into an ensemble (it takes many decision trees and uses all their outputs to improve its own accuracy and minimize variance). The models are trained independently (on data) with one small modification: For each split in each tree, a small subset features is randomly sampled, and all other features are ignored. This procedure reduces the variance of the final model as training trees on random features results in them being uncorrelated further. |

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

|  |
| --- |
| In GBDTs, the depth of a decision tree is very shallow; in random forest, it is very deep. In GBDTs, all the features are used to train the model; in random forest, we only use a subset of the features.  In GBDTs, the training data is the residuals, while in random forest, we randomly sample data to train the model. |

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

|  |
| --- |
| In supervised learning, when we train a model, we provide it with features and a label i.e., we tell the model that for a certain set of features (say age, bmi, etc.) we have a label (say predicted length of life).  In unsupervised learning however, when we train a model, we only provide it with features, and expect it to cluster our data based on any patterns it sees. The purpose of this is to uncover behaviors or events that deviate from normal patterns in data. |

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

|  |
| --- |
| If I had a bunch of colors and I wanted to classify them into separate labels, I would use unsupervised clustering to group them into those colors.  Then (if needed), I could assign specific colors to labels, and move on with my problem. |

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