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 the process of combining multiple models predicting a given task into a single prediction. When applied correctly, this technique enables one reduce the impact of idiosyncratic errors from one single model. The effect of this is generally better generalization performance. |

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

|  |
| --- |
| Bias is the degree to which a certain algorithm is able fit the true relationships or patterns within data. Variance is the degree to which a certain algorithms’s predictions fluctuate through variations of the training data. The bias-variance tradeoff is a way to decompose a model’s error, where bias and variance on the main components. These two forces usually compete with each other, such that when one increases a model’s bias one reduces the variance, and vice versa. |

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

|  |  |
| --- | --- |
| In bagging one builds multiple models by taking N bootstrap samples of the same data and usually applying the same algorithm to each. In boosting one creates multiple models by iteratively fitting a new model to the prior models’ combined prediction residual. In stacking one just simply aggregates multiple models of any type into a single prediction. |  |

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

|  |
| --- |
| A Random Forest is essentially the bagging procedure with Decision Trees used as the base algorithm. One key addition to traditional bagging is that in each tree that is built, a random subset of features are selected. |

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

|  |
| --- |
| Both methods use Decision Trees as the base classifier. The difference is that RF uses bagging with random feature selection and GBDT uses boosting (both described above). Additionally, for the individual trees, one usually fits larger and overfit trees in RF and shallower and underfit trees in GBDT. |

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

|  |
| --- |
| The biggest difference is that supervised learning uses training data that contains labeled examples and unsupervised learning does not. It learns patterns from the data without having a label. |

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

|  |
| --- |
| Clustering attempts to group subsets of data (clusters) that are collectively similar to one another based on the similarity of their feature values. One area in which clustering may be used is in image processing, in which a clustering algorithm will group images based on characteristics of each image, such as color or shape. Image data clustering is used to organize or analyze a large collection of images, such as medical images. Another area in which clustering may be used is in retail marketing. Clustering can group customers based on purchasing patterns and can help retailers discover distinct groups of customers. |

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