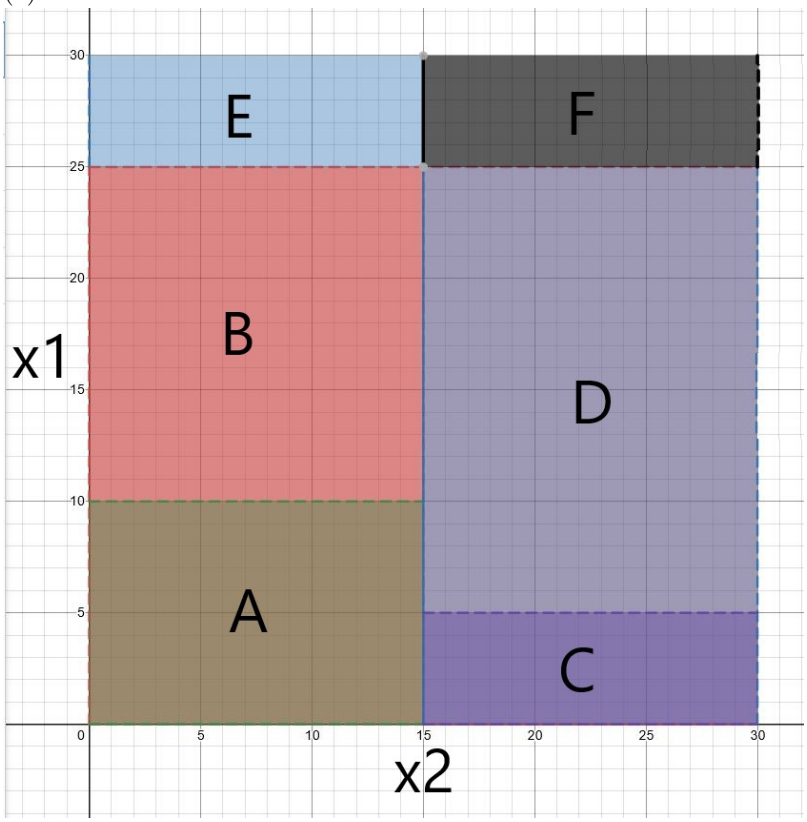
1 Exercises Decision Trees and Ensembles

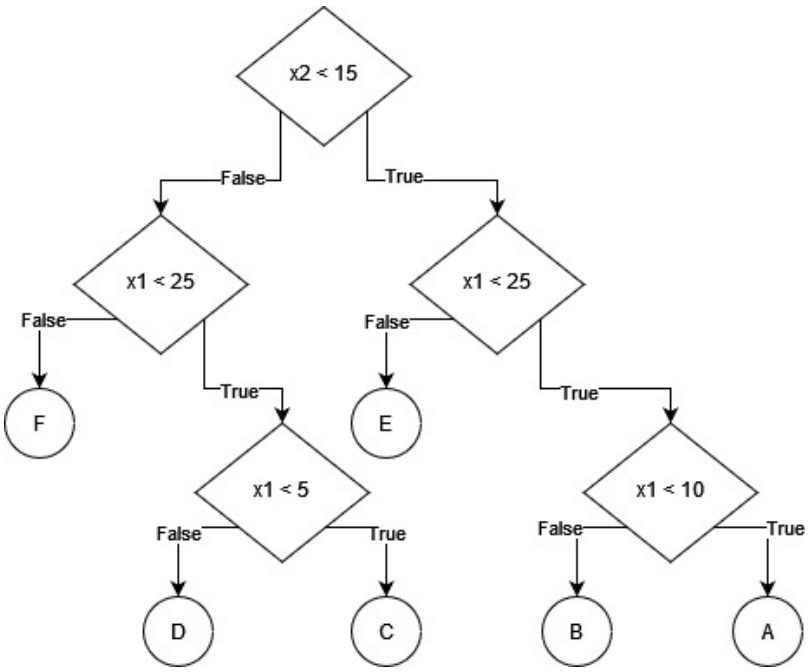
Problem 1
Q1 Drawing Decision Tree Predictions

Solution

1. (a) Decision Boundaries



2. (b) New Decision Tree



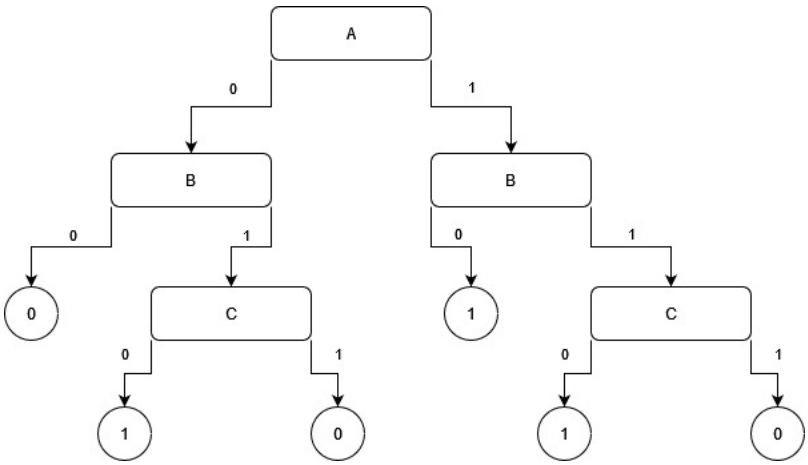
3. The redundancy may make it difficult to find an accurate and efficient decision tree. The more splits there are, the longer it will take to reach a decision. In addition, the redundancy may give rise to overly complex sub-trees, which may result in over-fitting.

Problem 2
Q2 Manually Learning A Decision Tree

Solution

1. $H(Y) = -(\frac{3}{6}log_2\frac{3}{6} + \frac{3}{6}log_2\frac{3}{6}) = 1$
2. $H(Y \mid A) = -\frac{3}{6}(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) * 2 = 0.918$
3. A Information Gain: $1 - 0.918 = 0.082$
4. $H(Y \mid B) = -\frac{3}{6}(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}) * 2 = 1$
5. B Information Gain: $1 - 1 = 0$
6. $H(Y \mid C) = -\frac{3}{6}(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}) * 2 = 0.918$
7. C Information Gain: $1 - 0.918 = 0.082$
8. Choose A as root decision node
9. $H(Y \mid A, B) = -\frac{1}{3}(\frac{1}{1}log_2\frac{1}{1}) - \frac{2}{3}(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}) = 0.667$
10. New B Information Gain: $1 - 0.667 = 0.333$
11. $H(Y \mid A, C) = -\frac{1}{3}(\frac{1}{1}log_2\frac{1}{1}) - \frac{2}{3}(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}) = 0.667$
12. New C Information Gain: $1 - 0.667 = 0.333$
13. Choose either B or C as next decision node

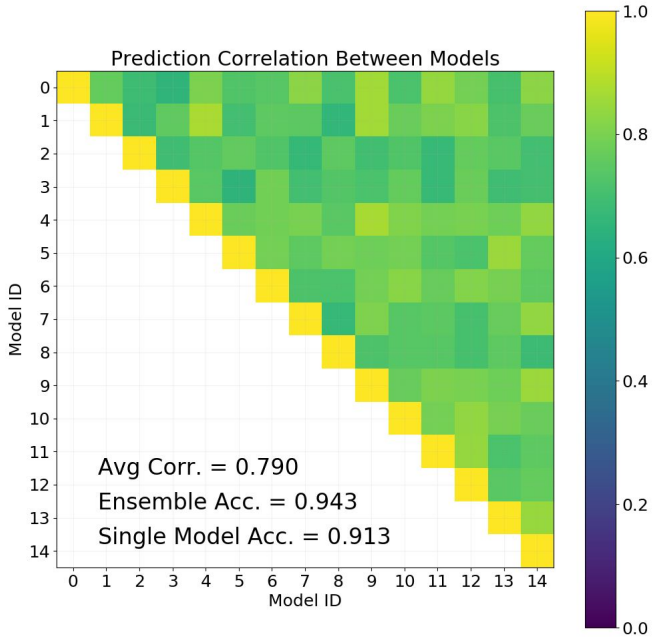
14.



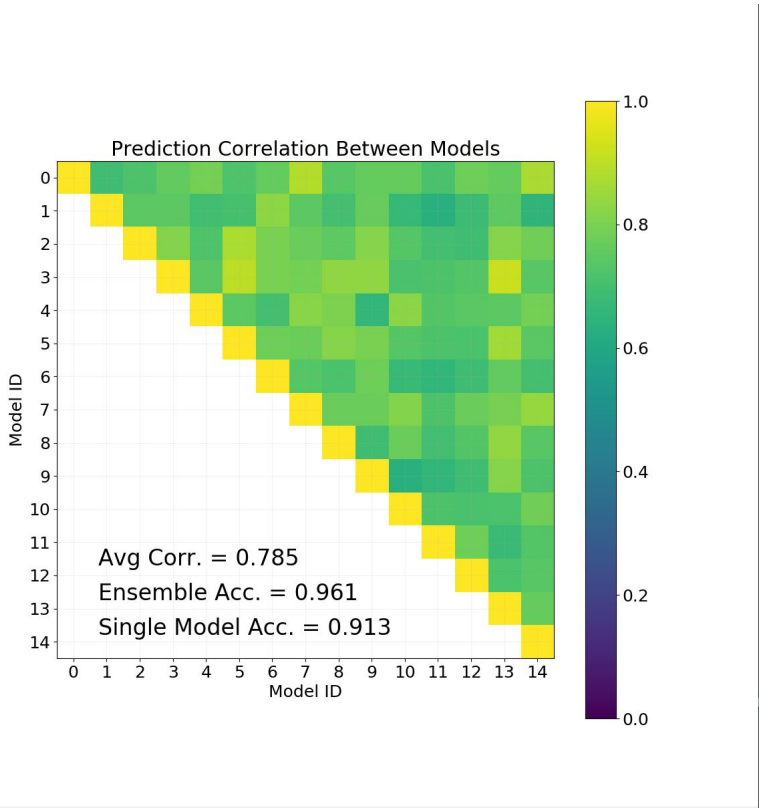
15. Since all predicted y's match the y's from the training dataset, the training accuracy is 100%

Problem 3
Q3 Measuring Correlation in Random Forests

Solution



1. After implementing bagging, it appears the prediction correlation between each of the models lowered significantly. Lowering each model's prediction correlation also seems to have further improved the ensemble's average accuracy by a significant margin.



2. For the DecisionTreeClassifier's max_features argument, I set it to 15 or half of the original number of features. The result of introducing randomness to the chosen features for best splits caused the average correlation between models to drop significantly. Due to the decreased correlation between models, the ensemble's accuracy appears to have increased significantly as well.

2 Implementation: k-Means Clustering

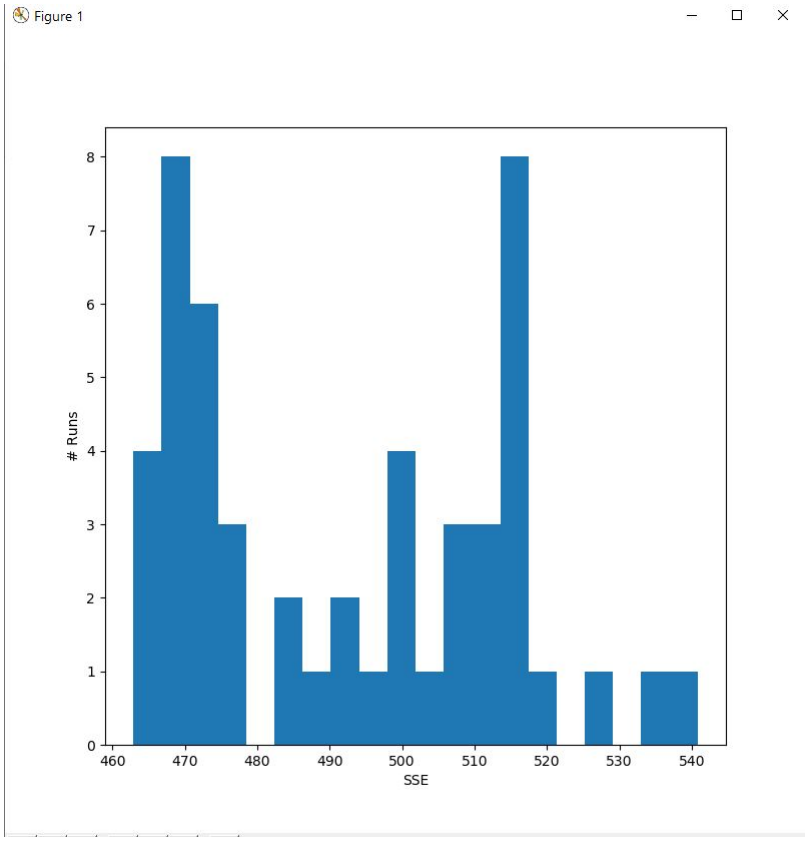
Problem 4
Q4 Implement k-Means

Solution

1. Completed

Problem 5
Q5 Randomness in Clustering

Solution

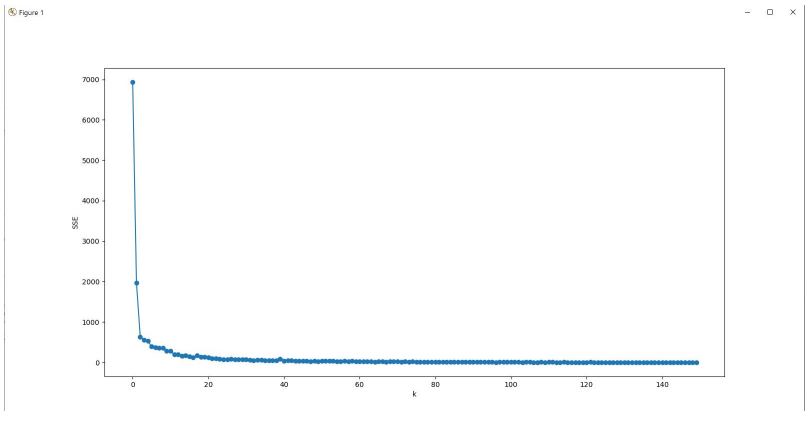
1. 

2. It appears that the SSEs are unevenly spread out. Most of the SSEs hover around 470 and 515, forming two peaks. The rest are positioned between the two peaks or to the left of the second peak. There does not seem to be a pattern to the distribution which may solidify k-Means's sensitivity to outliers.

3. Given k-Mean's sensitivity to how centroids are initialized and outliers, we can manage this issue by employing better centroid initialization methods, such as making the cluster centers far from each other or by taking the median for each centroid update step instead of the mean.

Problem 6
Q6 Error Vs. K

Solution

1. 

2. According to the graph, after roughly k = 20, the rate of descent for the SSE began plateauing. This implies that at some point, increasing k will yield negligible returns on SSE. Moreover, choosing k

based on SSE does not make sense because even when the SSE eventually plateaus, k can keep going on infinitely, giving us an infinite amount of k's to choose from.

Problem 7
Q7 Clustering Images

Solution

1. All of the images appear to be related to the outdoors. The images are roughly grouped into tall buildings, highways, and forests. Although the images in each cluster are mostly consistent with similar features, some clusters appear to possess images with non-correlated features. This implies that k = 10 may be a bit higher than needed since there are clusters that do not give us meaningful insight.

2. k = 3 Cluster Samples

3. The final SSEs for k = 10 runs were lower than the final SSEs for k = 3 runs, however k = 3 clusterings consistently yielded more meaningful patterns. This shows that SSE is not a good indicator of clustering quality.

Problem 8
Q8 Evaluating Clustering as Classification

Solution

1. Cluster 1: Highway, Cluster 2: Tall Building, Cluster 3: Forest
2. Cluster 1 Purity: 5/50, Cluster 2 Purity: 1/50, Cluster 3 Purity: 8/50

3 Debriefing

Problem 1
Approximately how many hours did you spend on this assignment?

Solution

1. About 7 hours.

Problem 2
Would you rate it as easy, moderate, or difficult?

Solution

1. Moderate-easy.

Problem 3
Did you work on it mostly alone or did you discuss the problems with others?

Solution

1. Worked on it alone.

Problem 4
How deeply do you feel you understand the material it covers (0%–100%)?

Solution

1. I felt I understood about 90 percent of the material covered.

Problem 4
Any other comments?

Solution

1. None.