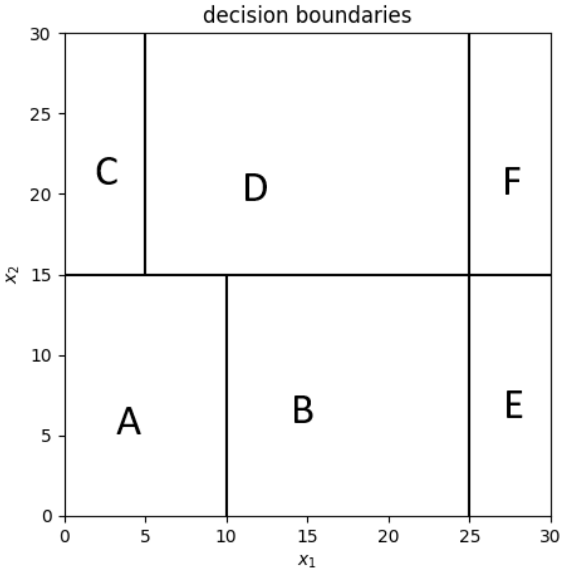
# CS434_HW4-3

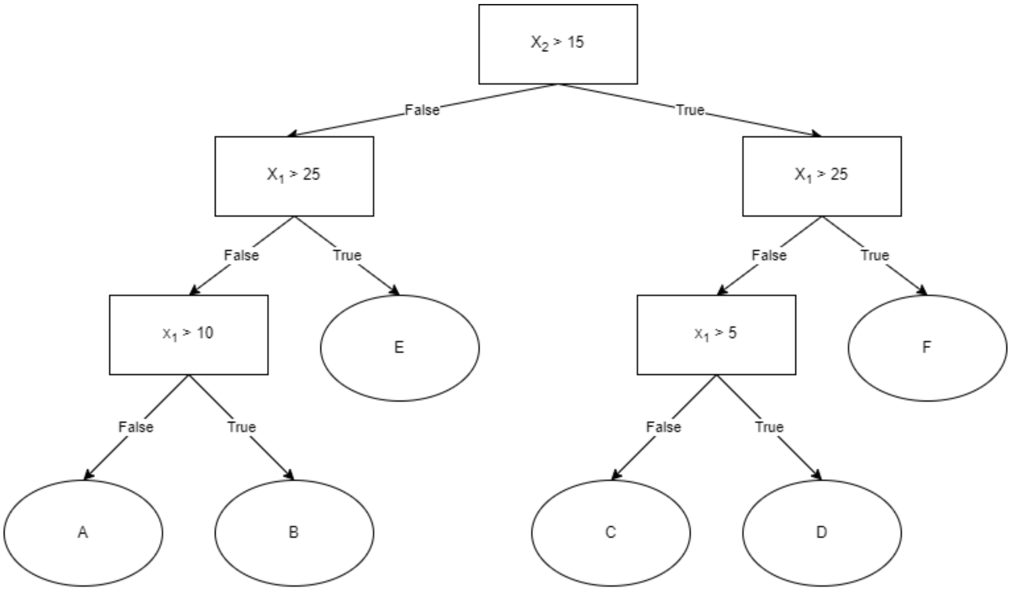
Lyon Kee

December 7, 2023

# 1 Exercises: Decision Trees and Ensembles [5pts]

## 1.1 Q1 Drawing Decision Tree Predictions [2pts]

a)  b)



c) There are a few factors that make it harder to obtain an accurate tree when the space of decision trees is synthetically redundant:

1. complexity: We need to compute more trees to find the accurate trees because there will be duplicates that cause redundancies, resulting in an added computational complexity, making the problem of finding the smallest tree NP-Hard.
2. overfitting: Because there are so many possibilities, there could be trees that are small and big, and these big trees are prone to overfitting while the small ones often underfit.

The factor that makes it easier to obtain a more accurate tree includes but is not limited to:

1. Ensembling: having the ability to represent trees in different structures and in different ways allows Ensemble methods to thrive, one of the more popular methods is known as Random Forest.

## 1.2 Q2 Manually Learning A Decision Tree [2pts]

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 |

Split on A:

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 |

Split on B:

| A | B | C | Y |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

Split on C:

| A | B | C | Y |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |

$$H(Y) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$H(Y|A) = -\frac{3}{6}(\frac{1}{3}log2\frac{1}{3} + \frac{2}{3}log2\frac{2}{3}) - \frac{3}{6}(\frac{2}{3}log2\frac{2}{3} + \frac{1}{3}log2\frac{1}{3}) = 0.918$$

$$H(Y|B) = -\frac{2}{6}(\frac{1}{2}log2\frac{1}{2} + \frac{1}{2}log2\frac{1}{2}) - \frac{4}{6}(\frac{2}{4}log2\frac{2}{4} + \frac{2}{4}log2\frac{2}{4}) = 1$$

$$H(Y|C) = -\frac{3}{6}(\frac{2}{3}log2\frac{2}{3} + \frac{1}{3}log2\frac{1}{3}) - \frac{3}{6}(\frac{1}{3}log2\frac{1}{3} + \frac{2}{3}log2\frac{2}{3}) = 0.918$$

$$IG(A) = H(Y) - H(Y|A) = 0.082$$

$$IG(B) = H(Y) - H(Y|B) = 0$$

$$IG(C) = H(Y) - H(Y|C) = 0.082$$

New split on C sets:

| A | B | C | Y |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |

Split $A|C = 0$:

| A | B | C | Y |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 1 | 1 | 0 | 1 |

split $A|C = 1$:

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |

| A | B | C | Y |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |

Split $B|C = 0$:

| A | B | C | Y |
|---|---|---|---|
| 0 | 0 | 0 | 0 |

| A | B | C | Y |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 |

split $B|C = 1$:

| A | B | C | Y |
|---|---|---|---|
| 1 | 0 | 1 | 1 |

| A | B | C | Y |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |

For C = 0:

$$\begin{aligned}
H(Y) &= -\frac{2}{3}log2\frac{2}{3} - \frac{1}{3}log2\frac{1}{3} = 0.918 \\
H(Y|A) &= -\frac{2}{3}(\frac{1}{2}log2\frac{1}{2} + \frac{1}{2}log2\frac{1}{2}) - \frac{1}{3}(\frac{1}{1}log2\frac{1}{1}) = 0.667 \\
H(Y|B) &= -\frac{1}{3}(\frac{1}{1}log2\frac{1}{1}) - \frac{2}{3}(\frac{2}{2}log2\frac{2}{2}) = 0 \\
IG(A) &= H(Y) - H(Y|A) = 0.252 \\
IG(B) &= H(Y) - H(Y|B) = 0.918
\end{aligned}$$

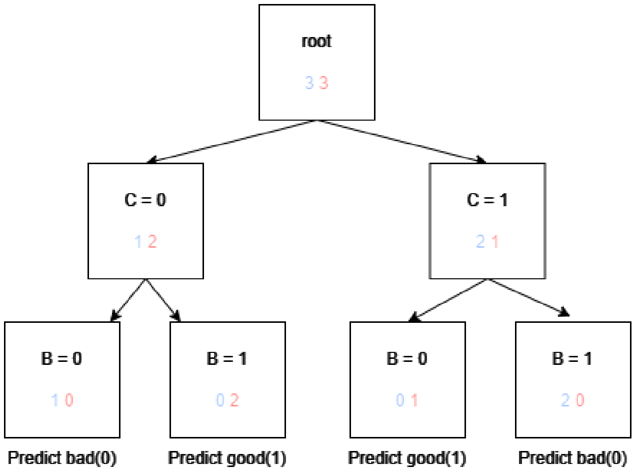
For C = 1:

$$\begin{aligned}
H(Y) &= -\frac{2}{3}log2\frac{2}{3} - \frac{1}{3}log2\frac{1}{3} = 0.918 \\
H(Y|A) &= -\frac{1}{3}(\frac{1}{1}log2\frac{1}{1}) - \frac{2}{3}(\frac{1}{2}log2\frac{1}{2} + \frac{1}{2}log2\frac{1}{2}) = 0.667 \\
H(Y|B) &= -\frac{1}{3}(\frac{1}{1}log2\frac{1}{1}) - \frac{2}{3}(\frac{2}{2}log2\frac{2}{2}) = 0 \\
IG(A) &= H(Y) - H(Y|A) = 0.252 \\
IG(B) &= H(Y) - H(Y|B) = 0.918
\end{aligned}$$

At this point, every node has all matching records with the same output value. By selecting the highest information gain for every layer we arrived at a split at C followed by a split at B.

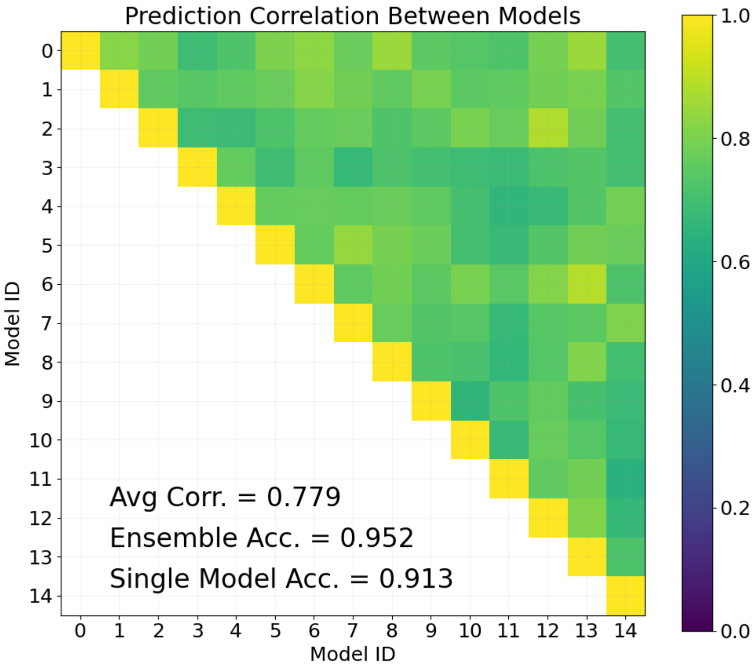
| A | B | C | Y | $Y_{pred}$ |
|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |

Training accuracy = 100%



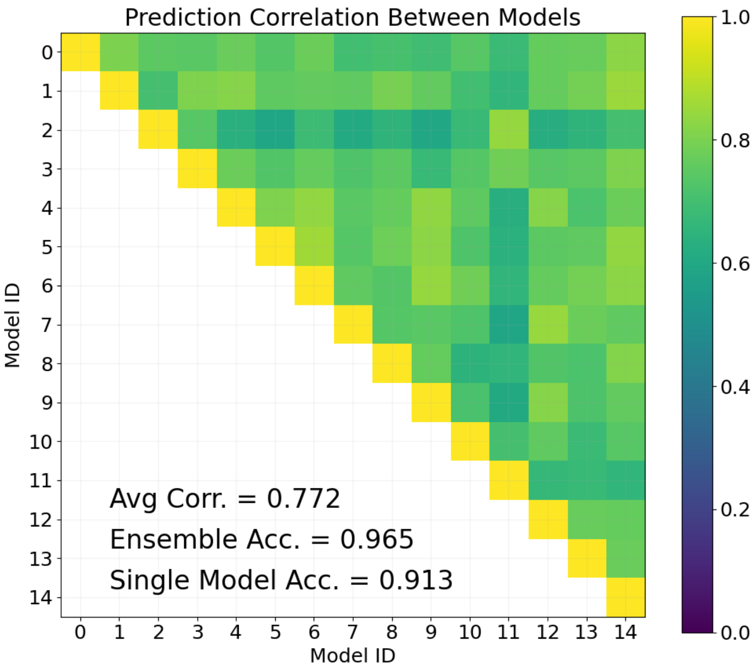
## 1.3 Q3 Measuring Correlation in Random Forests [1pts]

1.



This is the resulting plot after sampling train data points with replacement, we observe that the average correlation has decreased but there is an increase in ensemble accuracy. This is to be expected because an assemble with full correlation is not going to improve accuracy, and by having uncorrelated results, we will be able to improve the performance as there are more models that contribute to the final ensemble.

2.



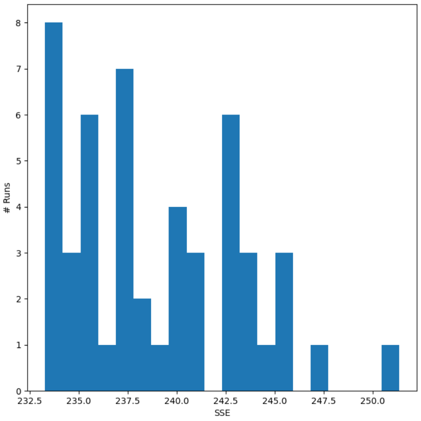
This is the resulting plot after changing the maximum features to 20. We observe that there will be more differences between each model and the increased model will mean that there will be less correlation, thus the decrease in average correlation. The specific example shows an increase in Ensemble accuracy however this is not to be expected as the decrease in features should yield less accurate data and also a simpler model. However, it is difficult to tell as the previous model could be overfitting, and reducing the max features could have made the model more accurate. Another reason for a better ensemble accuracy is because there is lesser correlation which allows us to have higher ensemble accuracy. A common max feature input is to have $\sqrt{d}$ where d is the number of features making it around 6 features would usually result in the highest ensemble accuracy but it is not used here because I believe we would require more models if we are to reduce the number of features which we are not doing right now.

# 2 Implementation: k-Means Clustering [20pts]

## 2.1 Implementing k-Means Clustering

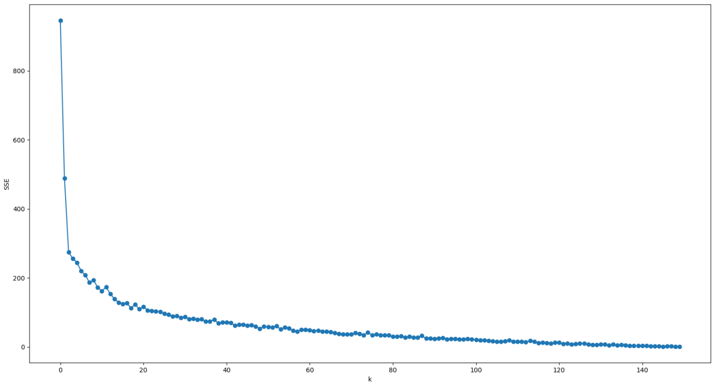
### 2.1.1 Q4 Implement k-Means [10pts]

### 2.1.2 Q5 Randomness in Clustering [2pt]



When applying k-means to a real dataset, we observe that the resulting SSE is random in nature and it weighs heavily on the initialization centroids. As such, we should run k-means multiple times to obtain a rough estimate of an SSE that is the lowest. We could also perform ensembling methods to improve performance. When we have more runs, we observe that most of them average around 235 so we just have to run it a few times to ensure that we did not start with a bad initial centroid placement, or we will yield an SSE that is off by quite a margin like in the figure shown above.

### 2.1.3 Q6 Error Vs. K [2pt]

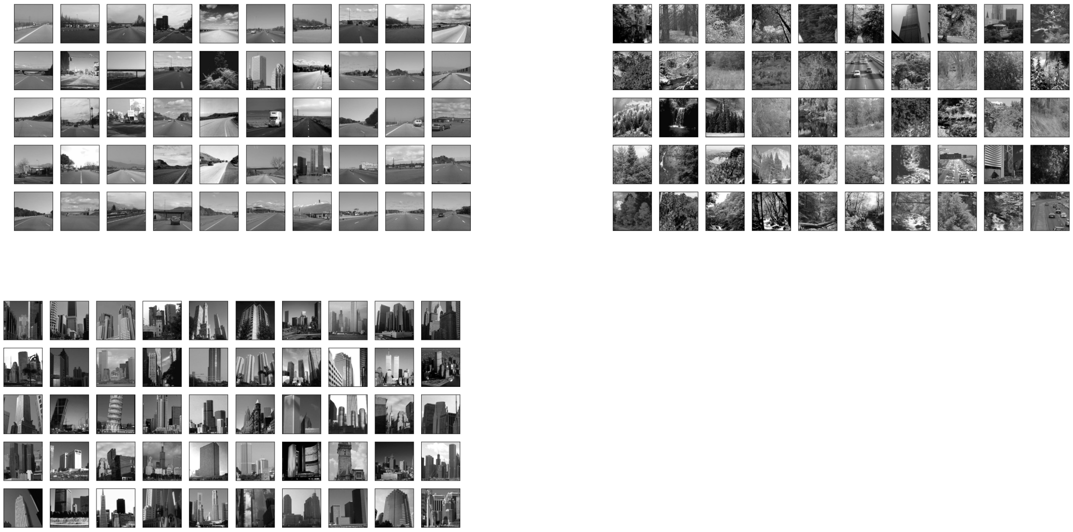


Choosing k based on SSE might not always be the best case, this is because SSE is sensitive to scales and a point far away will greatly affect the SSE. SSE also does not tell us the true properties of a cluster. In our case, we are only looking at isotropic Gaussian, and a dataset might be better suited to a different Gaussian that is misrepresented on our SSE. Looking for the drop in the graph with the slightest increase in k is also harsh due to the fact that it could be hard to identify the drop, another reason is that the ”elbow” is also quite dependent on the initializing centroids.

## 2.2 k-Means on an Unorganized Image Collection

### 2.2.1 Q7 Clustering Images.[4pt]

1. k=10 seems to be too high, the data set contains images in 4 categories: buildings, roads, forests, and an outlier of a panda. Judging by what is in the dataset, we can argue that our best k is around 3 and we should just ignore the outlier. If we can capture non-isometric Gaussian, we should be able to increase k to capture different types of buildings, different types of roads, and also forests.

2. 

3. the chosen k has an SSE of 1196 and k=3 has an SSE of 1292, as expected k is not a good indicator of clustering quality. It is subject to converge to small clusters of dense clusters instead of finding true underlying clusters, and as humans, these clusters provide no meaningful purpose as to a lower k with a higher SSE.

### 2.2.2 Q8 Evaluating Clustering as Classification [2pt]

- Forest, Roads, and Buildings.

- purity:
  buildings: 50/50
  roads: 47/50

forest: 44/50

# 3 Debriefing (required in your report)

## 3.1 Approximately how many hours did you spend on this assignment?

6 hours.

## 3.2 Would you rate it as easy, moderate, or difficult?

This assignment is pretty easy, I was able to complete this assignment without much trouble, its just time-consuming. Unlike others, I was not stuck or anything.

## 3.3 Did you work on it mostly alone or did you discuss the problems with others?

Alone, with no help other than class slides.

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

80%, I believe I have some issues with how sk learn implements its tree class and the DecisionTreeClassifier.

## 3.5 Any other comments?

No