COMP3105 Assignment 3 Report

Ryan Lo (101117765)

Q1d)

Training Acc:

n	Model 1	Model 2
16	0.95625	0.99375
32	0.91875	0.978135
64	0.865625	0.9609375
128	0.8734375	0.95

Test Acc:

n	Model 1	Model 2
16	0.7214	0.889
32	0.8026	0.8867
64	0.8316	0.8879
128	0.8364	0.9169

Training Accuracies:
Model 1:
Accuracy generally decreases as the dataset size (n) increases.
This behaviour could indicate that Model 1 may struggle to capture more complex patterns as the dataset size grows.
Model 2:
Accuracy remains high and even increases as the dataset size increases.
Model 2 appears to handle larger datasets well, suggesting it might be capable of capturing more intricate patterns.
Test Accuracies:
Model 1:
Test accuracy is lower compared to training accuracy, especially with smaller datasets.
This behaviour could suggest overfitting, where Model 1 is memorizing the training data but struggles to generalize to new data.
Model 2:
Test accuracy remains high and consistent across different dataset sizes.
Model 2 seems to generalize well to unseen data, indicating that it might be more robust.

Q2d)

Training Acc:

Dim k	Model 1	Model 2
1	0.85992187	0.64234375
2	0.85726563	0.63092

Test Acc:

Dim k	Model 1	Model 2
1	0.84488	0.63092
2	0.83802	0.91706

Q2e)

Training Accuracy:
Model 1 vs. Model 2:

For Model 1, the training accuracy decreases as the dimension (k) increases. From 1 to 2, the accuracy drops.
For Model 2, the training accuracy also decreases as the dimension (k) increases, but the drop is more significant.

This behaviour suggests that including more dimensions (features) may not always lead to better training accuracy.
It's possible that the additional dimensions in the high-dimensional space might be introducing noise or overfitting to the training data, especially for Model 2.

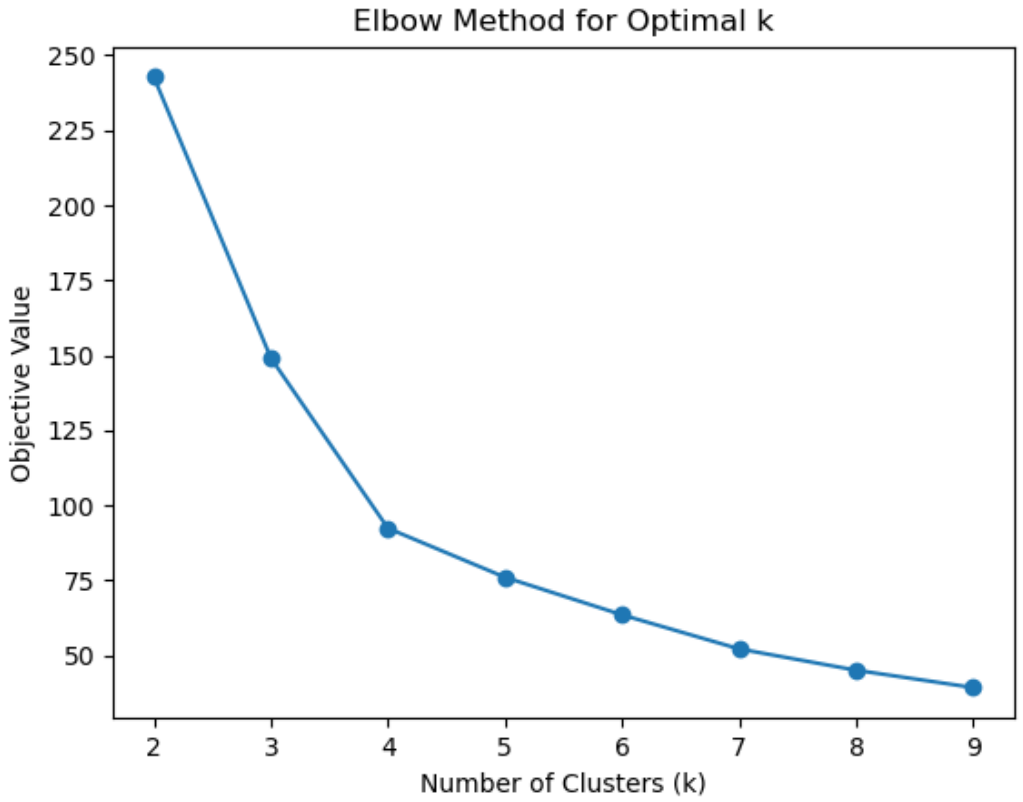
Test Accuracy:
Model 1 vs. Model 2:

For Model 1, the test accuracy decreases as the dimension (k) increases.
For Model 2, the test accuracy increases as the dimension (k) increases, and there's a significant improvement.

Model 1 might be suffering from overfitting, as evidenced by the drop in test accuracy when adding more dimensions.
Model 2, on the other hand, seems to benefit from additional dimensions, possibly capturing more complex patterns in the data.

Q3e)

k	2	3	4	5	6	7	8	9
obj_val	249.3538854374628 5	149.9199561975015	77.20749723483183	63.64030470456819 4	55.18093703246737	49.14782831039119	43.27658815916855 4	38.62690630640711



Plotting the objective_val on a graph and looking at the values, It seems that there is a noticeable decrease in the objective value up to k=4, and the rate of decrease slows down after that. Therefore, k=4 appears to be a reasonable choice based on the elbow method. This choice balances the trade-off between model complexity and the ability to explain the variability in the data.