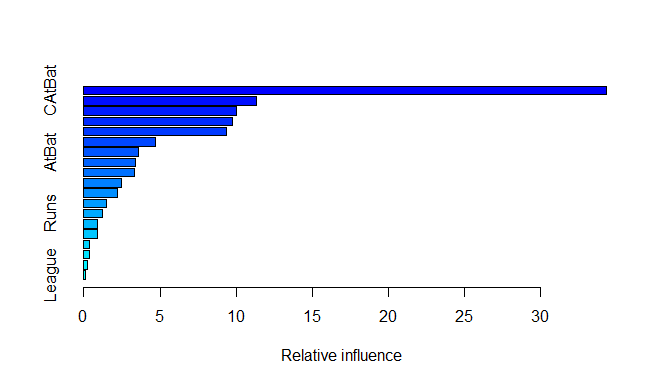
**Problem 1: Survey**

Completed survey on canvas.

**Problem 2: Boosting**

1. Refer to Rscript.
2. Refer to Rscript.
3. For a gradient boosting model, the best-performing model has 300 trees, an interaction depth of 3, a shrinkage value of 0.01 and a minimum number of observations in a terminal node set to 5
4. A screenshot of a computer

   Description automatically generatedThe predictor CAtBat indicates to be the most important predictor in the boosted model.



1. The test MSE for the boosted model: 0.2854964
2. Using 5 k-fold cross validation, the optimal m was 4. The test MSE with our optimal m was 0.2135547.

A screenshot of a computer program

Description automatically generated

**Problem 3: Bias of Trees**

1. The first predicted values for the first 5 iterations of the loop are:

14.59419, 17.08518, 5.480257, 14.40228, 10.36534

1. Squared bias is 155.9172. The variance is 17.42247
2. The first predicted values for the first 5 iterations of the loop are:

12.4918 11.93545 11.21848 16.48158 13.94433

1. Squared bias is 102.58. The variance is 3.71253
2. The change in the order of magnitude in squared bias between the two methods is -0.1818311
3. The change in the order of magnitude in variance between the two methods is -0.6714397
4. The squared bias for the random forest is lower, suggesting that the random forest tends to provide predictions that are closer to the true underlying values, resulting in lower bias. The variance for the random forest is significantly lower, indicating that the random forest helps reduce the variability in predictions compared to a single decision tree. Both the change in the order of magnitude for squared bias and variance are negative, suggesting that the random forest method provides a reduction in both bias and variance compared to a single decision tree. So, the random forest appears to offer an improvement over the decision tree.

Top of Form

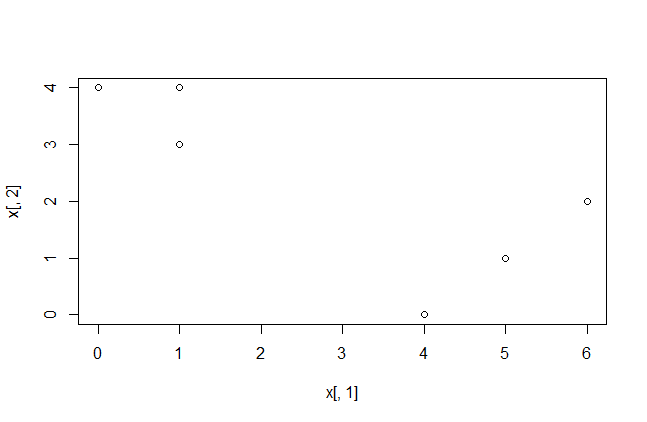
**Problem 4: Understanding K-Means**

1. Here is the proof:

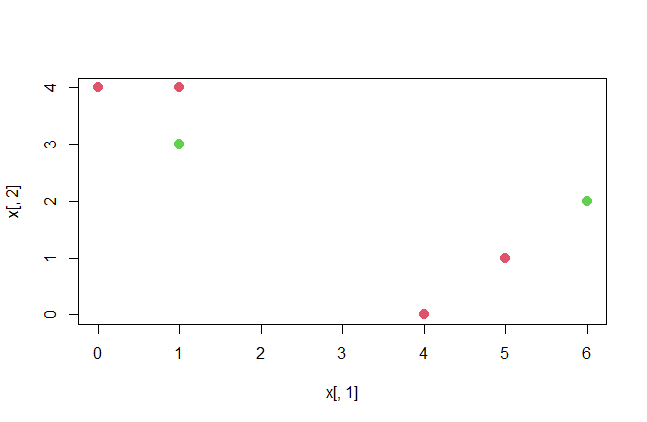
A notebook with math equations on it

Description automatically generated

1. In the first step of every iteration, it is minimizing the sum of deviations to the center in each cluster. So, by relocating the observations to the nearest center, the sum of the deviations will decrease. Basically, the identity shows that minimizing the sum of the squared Euclidean distance for each cluster is the same as minimizing the within-cluster variance for each cluster.
2. (i) Here is the plot:



(ii) The cluster labels are: 1 2 1 1 2 1



(iii) The centroid for the red cluster is

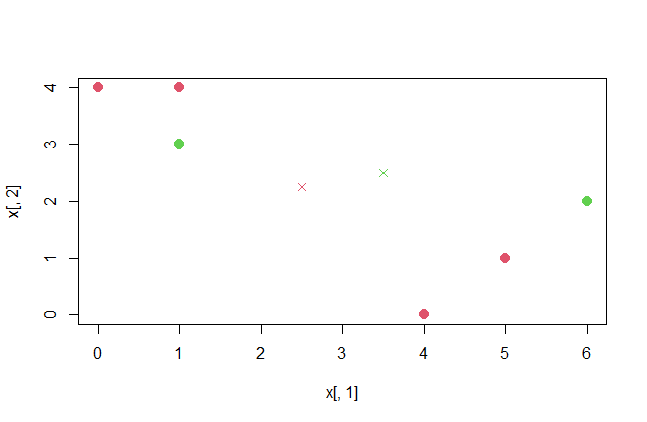
x̄11 = 1/4(0 + 1 + 4 + 5) = 2.5

x̄12 = 1/4(4 + 4 + 0 + 1) = 2.25

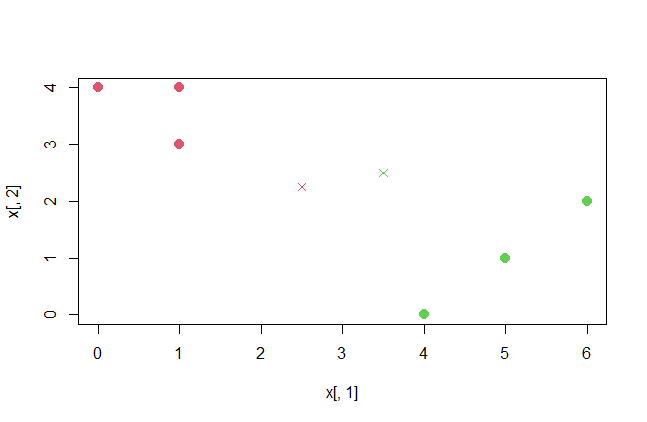
The centroid for the green cluster is

x̄21 = 1/2(1 + 6) = 3.5

x̄22 = 1/2(3 + 2) = 2.5



(iv) The updated cluster labels are: 1 1 1 2 2 2



(v) The centroid for the red cluster is

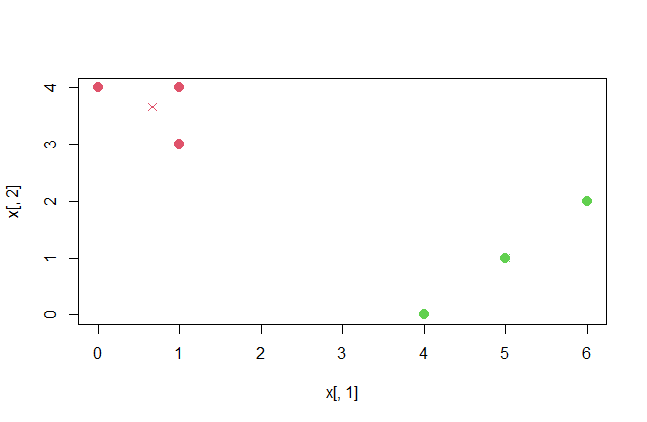
x̄11 = 1/3(0 + 1 + 1) = 0.6666667

x̄12 = 1/3(3 + 4 + 4) = 3.6666667

The centroid for the green cluster is

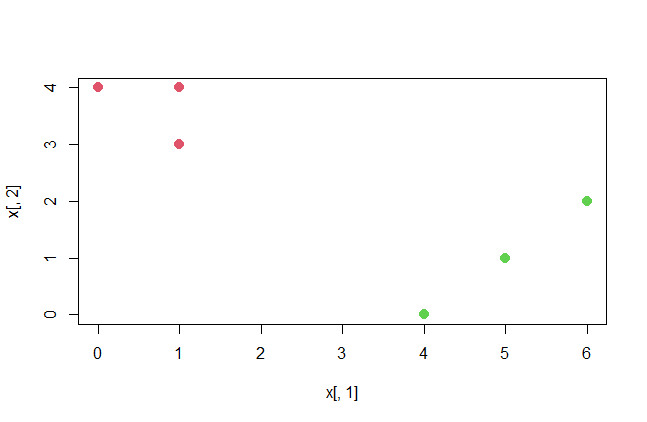
x̄21 = 1/3(4 + 5 + 6) = 5

x̄22 = 1/3(0 + 1 + 2) = 1



If the observation is assigned to the centroid to which it’s closest, nothing changes, and the algorithm is terminated at this step.

(vi) Here is the plot:



**Problem 5: Dendrogram**

1. Using complete linkage, start with {1},{2},{3},{4}

The most similar observations are 1 and 2 which makes {1,2} at 0.3

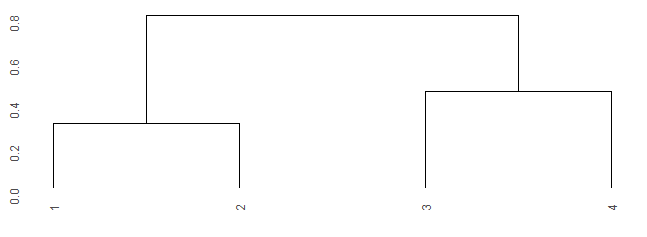
The next iteration is:

{1,2} to {3} is 0.5

{1,2} to {4} is 0.8

{3} {4} is 0.45 so fuse {3,4} at 0.45

Fuse {1,2} to {3,4} at 0.8



1. Using single linkage, start with {1}, {2}, {3}, {4}

The most similar observations are 1 and 2 which makes {1,2} at 0.3

Iteration 2:

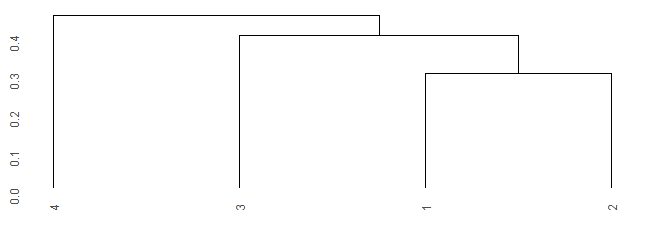
{1,2} to {3} is 0.4

{1,2} to {4} is 0.7

{3}{4} is 0.45

Fuse {1,2} to {3}

Fuse {1,2,3} to {4} at 0.45



1. {1,2} and {3,4}
2. {1,2,3} and {4}