Some slides for MATH 122A Final Project

Allen Minch, Group Member 2, Group Member 3

December 14, 2022

Motivation for exploration

- Maybe using one and only one model to predict y for a given test point is not such a fair prediction; maybe it would be more accurate to give some weight to the prediction of each model
- However, weighting schemes should probably not put too much weight on models trained with data quite far away, as predictions from such models may be less accurate
- Also interesting to try to cluster the test data; might even be more efficient, not having to compute as many distances

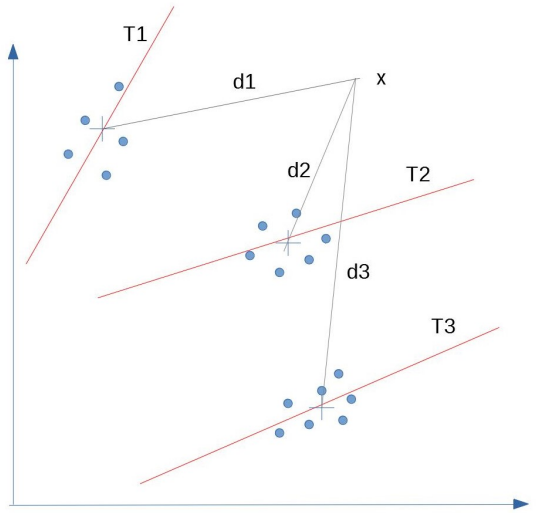
Six different approaches to making predictions in the test set

- Two options, A and B
  - Option A - Don't cluster the test data, and apply some rule to each individual test point to determine how to generate a prediction for y for each test point
  - Option B - Cluster the test data, and apply some rule to each cluster center of the test data to determine how to generate a prediction for y for each of the points in the relevant cluster of the test set; for simplicity, I implemented this approach with the same number of clusters in the test data as in the training data, but this is not essential

Six different approaches to making predictions in the test set

- Three different prediction rules, 1, 2, and 3
  - Rule 1 - the baseline: use exactly one model to make a prediction at a point, using the prediction of the model that was trained using the cluster with the closest cluster center to the point at which a prediction is being made
  - Rule 2 - Make use of all models to make a prediction at a given point as a weighted average of the predictions of the different models, where the weight given to the prediction of a certain model is inversely proportional to the distance from the point to the center of the cluster that was used to train the model.
  - Rule 3 - Make use of all models to make a prediction at a given point as a weighted average of the predictions of the different models, where the weight given to the prediction of a certain model is inversely proportional to the square of the distance from the point to the center of the cluster that was used to train the model.

Graphic illustrating the different prediction rules



Rule 1: $y(x) = T_2(x)$ since

$d_2 = \min\{d_1, d_2, d_3\}$

Rule 2:

$y(x) = w_1 T_1(x) + w_2 T_2(x) + w_3 T_3(x), w_1 d_1 = w_2 d_2 = w_3 d_3$

Rule 3:

$y(x) = w_1 T_1(x) + w_2 T_2(x) + w_3 T_3(x), w_1 d_1^2 = w_2 d_2^2 = w_3 d_3^2$

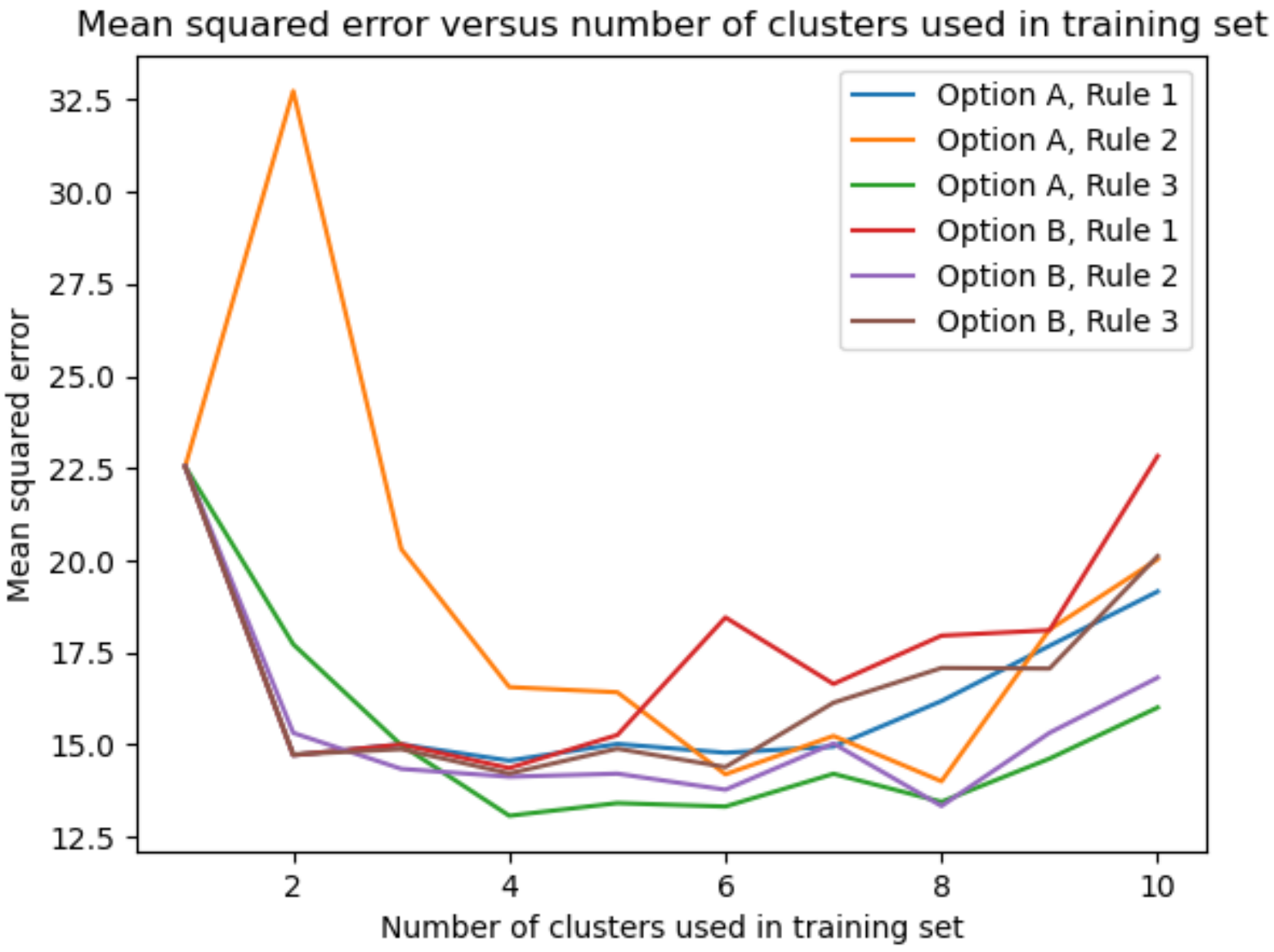
Note about the usage of rules 1, 2, and 3 in the context of each of Option A and Option B

- In Option A, the relevant distances for applying rules 1, 2, and 3 are distances from each individual test x to all of the cluster centers in the training data
- In Option B, the relevant distances for applying rules 1, 2, and 3 are the distances from the center of each cluster in the test data to the center of each cluster in the training data, and whatever a rule would specify to do to make a prediction at a cluster center in the test data, the same thing is done to make a prediction at every test point in that cluster

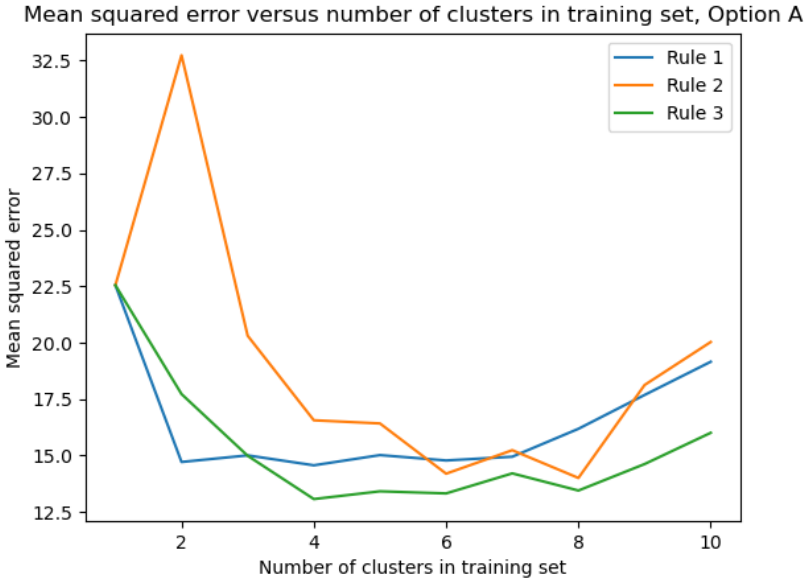
Visualization of Six Approaches

	Rule 1 - choose model with nearest training cluster center	Rule 2 - use weights inversely proportional to distance from training cluster center	Rule 3 - use weights inversely proportional to square of distance from training cluster center
Option A - don't cluster test data	Method 1	Method 2	Method 3
Option B - cluster test data	Method 4	Method 5	Method 6

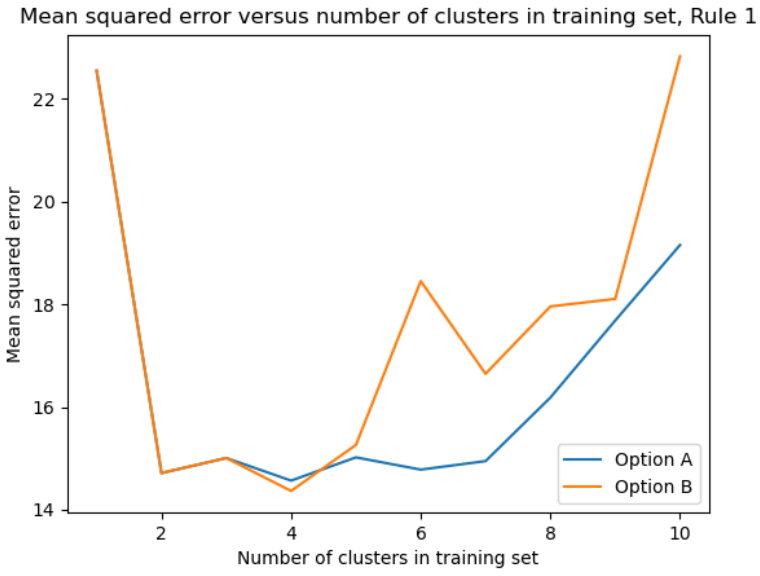
All six methods compared

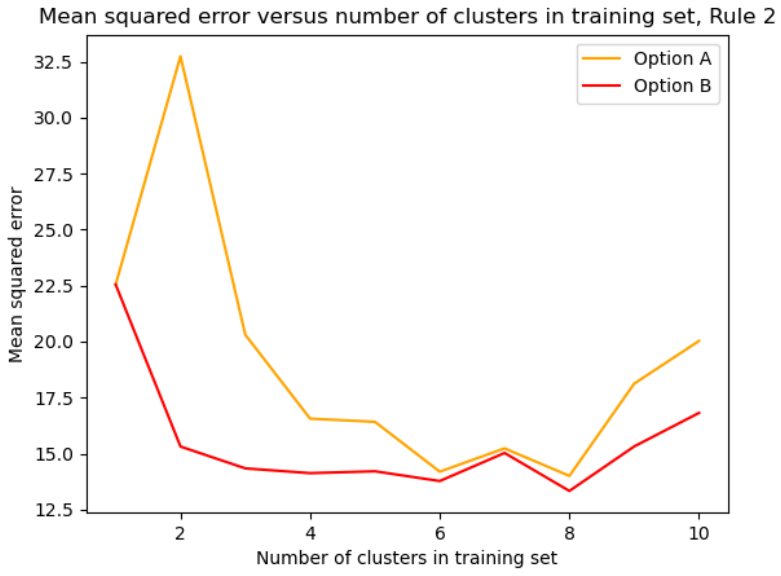


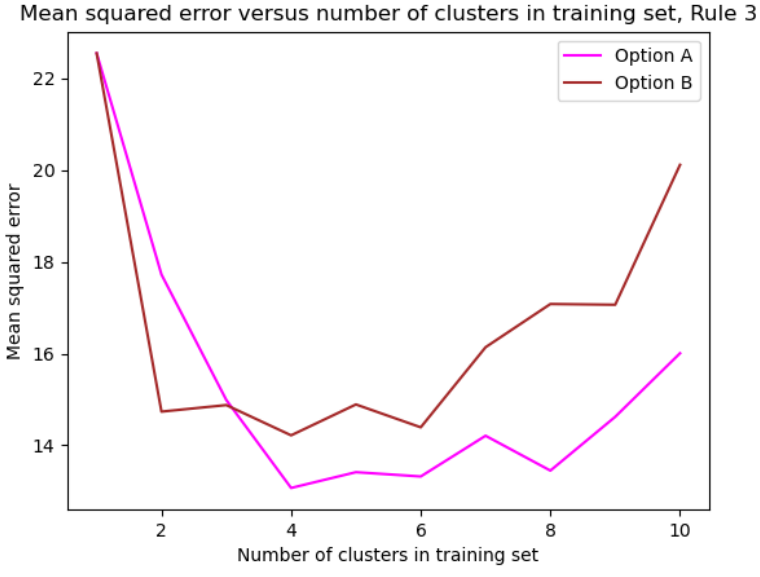
Comparison of Rule 1, 2, or 3 for Option A and Option B



Comparison of Option A and B for Rule 1, Rule 2, and Rule 3







Immediate Observations

- Of all of the methods, Option A/Rule 3 and Option B/Rule 2 (Methods 3 and 5) seem to perform the best
- If Option A is used, Rule 3 seems to perform the best
- If Option B is used, Rule 2 seems to perform the best
- If Rule 1 is used, Option A seems to perform better
- If Rule 2 is used, Option B seems to perform better
- If Rule 3 is used, Option A seems to perform better (overall)
- Method 2 (Option A/Rule 2) sticks out as the only method that performs worse with two clusters than with 1 (not sure why this would be)
- The best number of clusters to use with the different methods seems to be subject to variation (for some, 4 clusters seems best, while for others, 8 seems best)

Other Possible Variations on, or Generalizations of, The Six Methods Here

- Make a prediction as some sort of weighted or unweighted average of the predictions of the models associated with the n nearest cluster centers (as opposed to using only 1 model or all of them)
- Use Option B without necessarily having the same number of clusters in the test set as in the training set