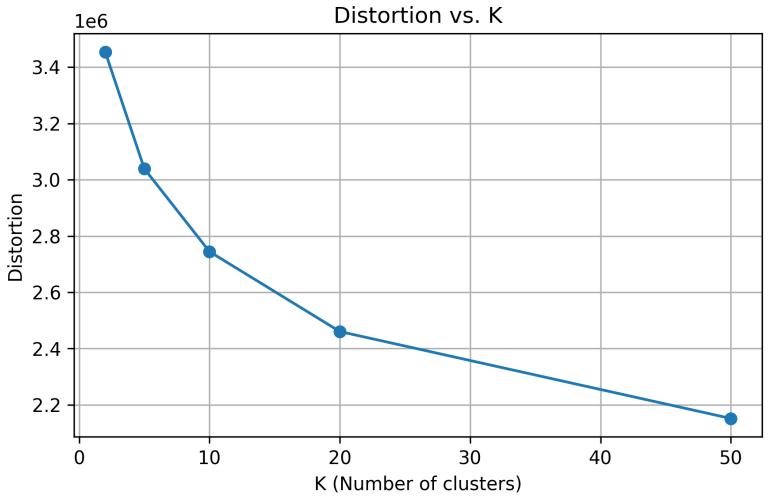
CMPT 423-820 MINI PROJECT 3 REPORT

Alireza Falamarzi (ylw576), Baptiste Rouquette (dly490), Princess Tayab (prt898), Talha Mansoor (kgy284)

1 Line Plot: Decrease in Distortion as the K Hyperparameter Increases



Distortion decreases as K increases, with diminishing returns.

2 Cluster Centroid Visualization

Figure 1: k = 50

Figure 2: k = 20

Figure 3: k = 10

Figure 4: k = 5

Figure 5: k = 2

How does the k parameter influence finding all 10 digit classes?

- Figure 1: All 10 digits appear multiple times.
- Figure 2: All digits appear at least once.
- Figure 3: Some digits are clearly missing such as 5 and 7 with some ambiguity between 3/8 and 4/9
- Figure 4: Majority of the digits are missing. 0,1, and 9 digits are the only digits that appear clearly and possibly 3.
- Figure 5: No digits appear clearly. Each cluster could be described as multiple numbers.

Does each cluster contain images of one or multiple digits?

- Figure 1: Most clusters contain only one image. Possibly leading to oversegmentation with some ambiguity between digits like 9 and 4. One cluster is extremely confused (multiple clusters for the same digit).
- Figure 2: Clusters become less defined than with k=50 and most clusters show one image digit. There are some still mixed multiple digits, particularly showing confusion between 4/9 (again) and 3/8.
- Figure 3: Most digits have a single main cluster, but the separation isn't as sharp as with k=20, and the 4/9 confusion persists.
- Figure 4: Clusters now group multiple digits together, but some structure emerges. For example, "1" (a simple straight line) is easier to distinguish, but most other digits are mixed.
- Figure 5: Severe mixing occurs where each cluster has multiple images resulting with no meaningful digit-specific clusters.

3 The Effectiveness of K-Means Clustering in Identifying 10 Digit Classes in the MNIST Dataset

K-Means clustering tries to group the digits in the MNIST dataset but has trouble separating them clearly because it only looks at pixel similarities, and digits can look very similar even if they belong to different classes. For example, digits like 1 and 7 or 3 and 5 have similar shapes or overlapping features, which confuses the algorithm. K-Means doesn't consider the actual digit labels and groups based on overall pixel structure, so it sometimes mixes up visually similar digits, like 4 and 9, or 3, 5 and 8, even though they are distinct. The number of clusters, or "k," really affects how well it works. With a high k (like k=50), the algorithm ends up splitting the same digit into multiple clusters, but it still struggles with digits that look similar. At k=10, which matches the number of digit classes, it's still not perfect as some digits are missing or mixed up because of overlapping features. Lower k values (like k=5 or k=2) end up grouping too many digits together, losing important differences. The main issue is that K-Means relies on measuring straight-line distance between pixels and doesn't take labels into account, so even if the distortion is reduced with a higher k, it doesn't necessarily improve digit separation. By testing clustering with different k values, we can see how unsupervised learning can reveal some structure in the digit images. However, K-Means isn't guaranteed to produce meaningful clusters without supervision, as shown by missing or combined digit representations.