Unsupervised Learning – Part 4: Analyzing the k-Means Algorithm

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Discuss the weaknesses of the k-means algorithm

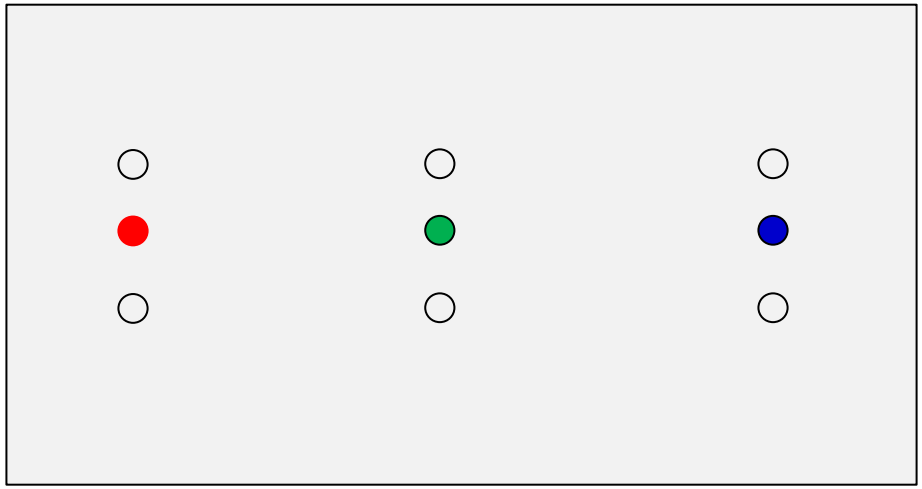
Objective

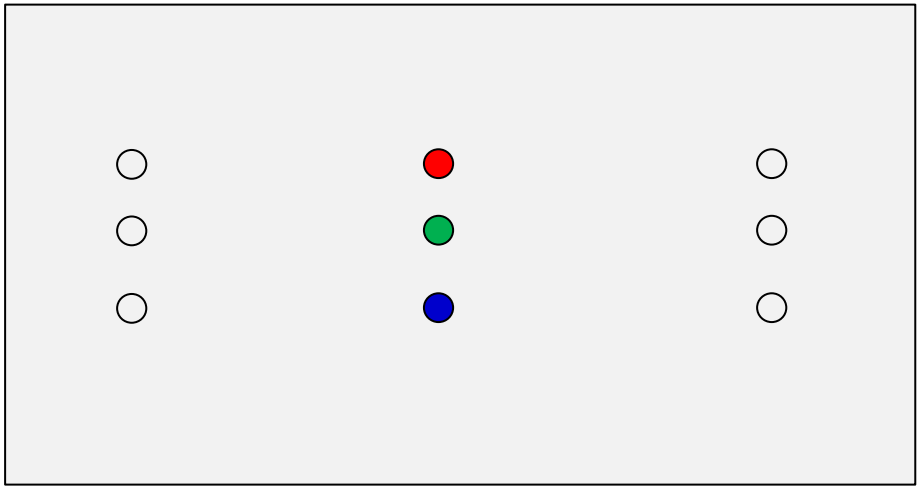
Discuss a few common techniques for potential improvement

Properties of the k-Means Algorithm

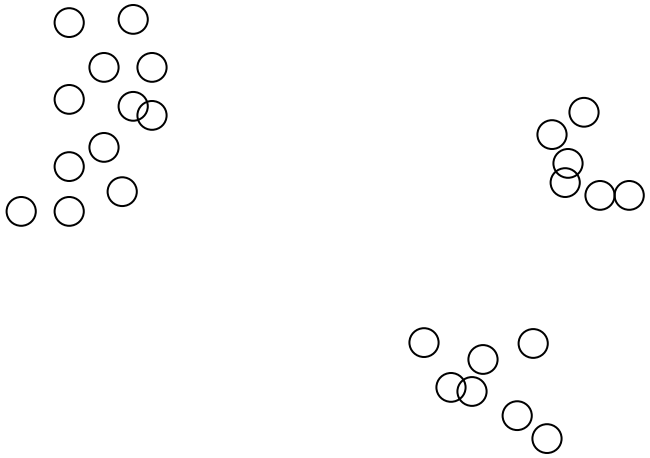
- The algorithm will converge when the cluster centers no longer change.
- But the results may not be an optimal solution.

➔ Sensitivity to initialization

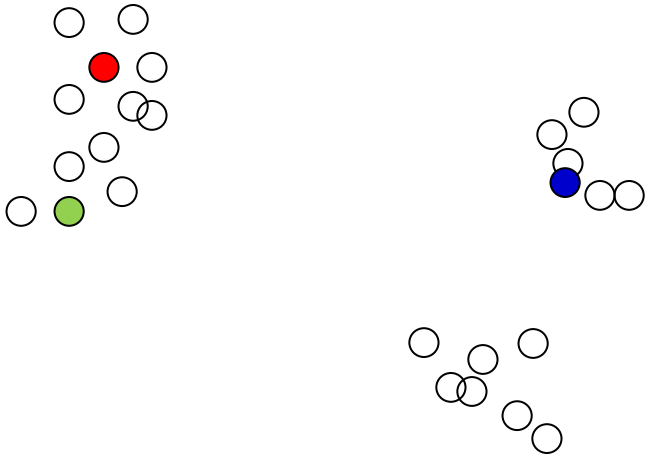




Another Example



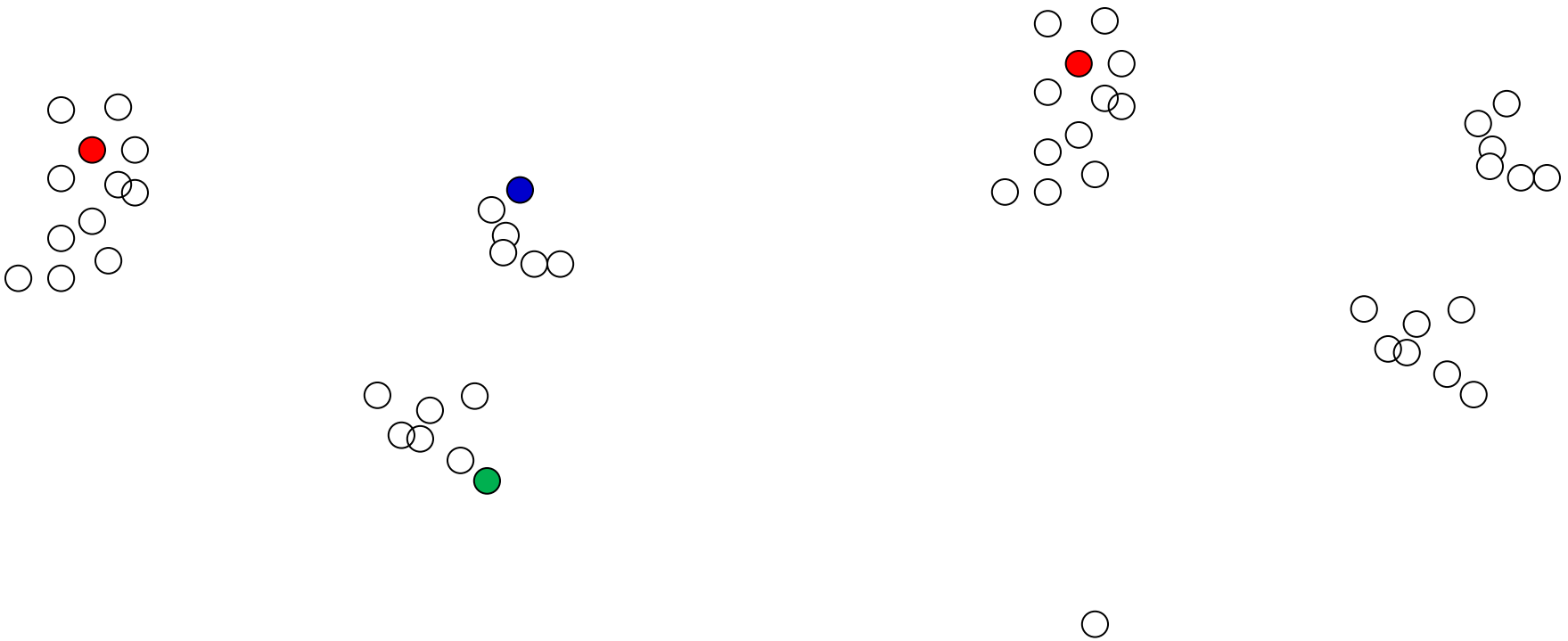
- ➔ The natural grouping seems to be so well defined.
- ➔ For k=3, what will be the clusters?



What can we do to improve?

A Few Common “Tricks”

- Multiple runs with different initial centers.
- Choosing the point furthest from the previous centers.
  - Drawback: might be sensitive to “outlies”.



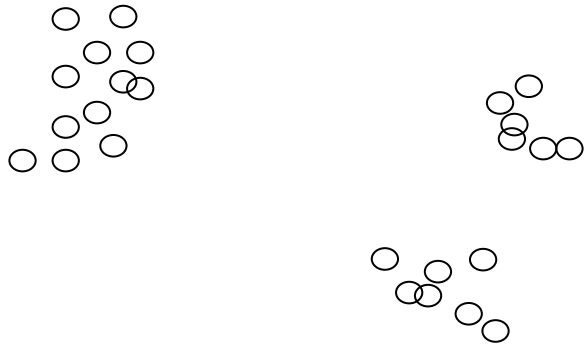
Other Variants of Basic k-Means

k-Means++:

- New centers are chosen with probabilities (as a function of distance to closest prior centers).
- Kind of between “random” and “furthest point” techniques.

Hierarchical approaches

- Agglomerative vs divisive.



The Question of Choosing k

Two trivial extremes

- If k=1, the error is the variance of the samples.
- If k=n, the error can become 0.

What is a proper 1<k<n for capturing the structure of the samples?

Some tricks

- Trick 1: Will the cost function drop dramatically at some point?
- Trick 2: Cross-validation (on, e.g., a classification task)