Pinching the Best k: Finding Optimal k Clusters in k-Means using CRAB Metric

Jasmine Cabrera, Allen Choi, Dr. Kelly Bodwin, Department of Statistics
California Polytechnic State University, San Luis Obispo

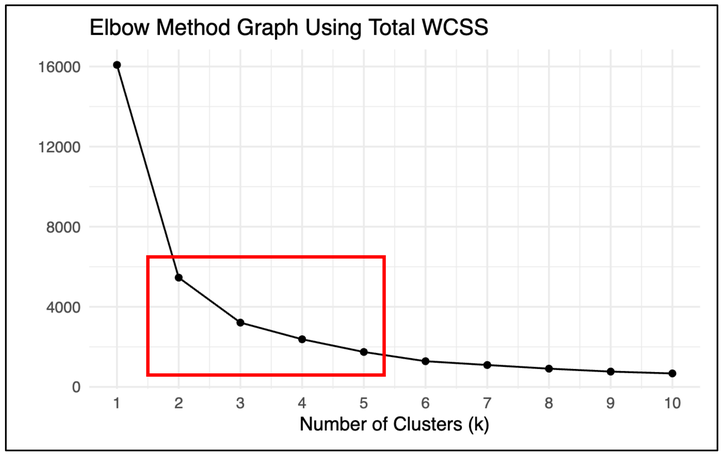
Cal Poly
Bailey College *of* Science & Mathematics

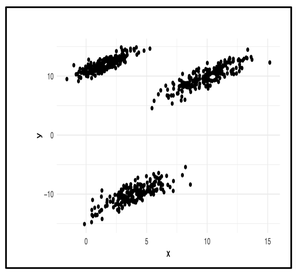
Motivation

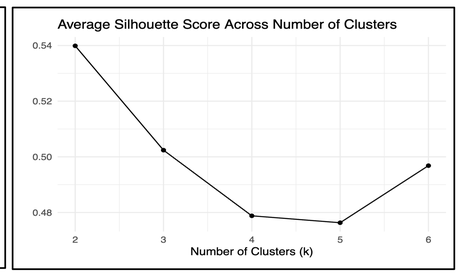
K-Means: A clustering algorithm for unsupervised learning

Problem: With no true value to compare to, it is difficult to determine the most optimal number of clusters

Current Methods for Evaluation:
Currently, some of the most important methods for evaluating k-Means are the Elbow Method and Silhouette Score. However, they are biased on separation and can be ambiguous.

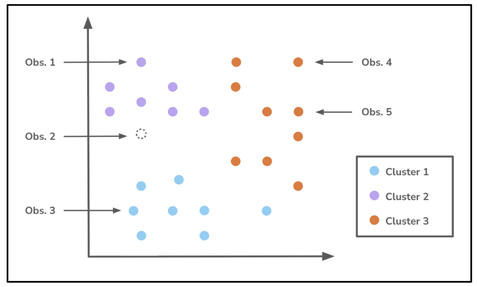


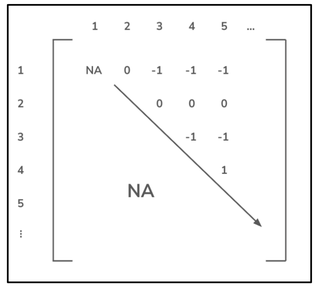


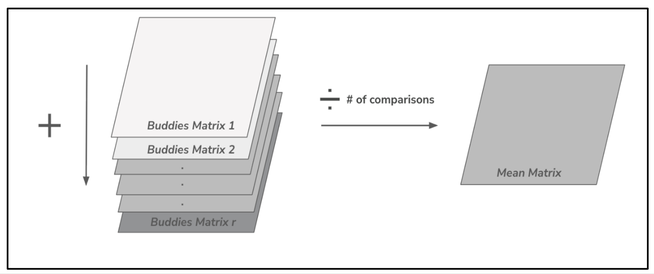


Clustering Rivals and Buddies (CRAB) Algorithm

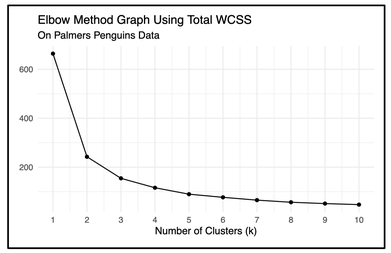
1. Select random subsamples of the data, each subsample a certain percentage of the full dataset.
2. On all subsamples of the data, obtain cluster assignments for every observation on a subsample by subsample basis.
3. For each pair of distinct points record whether they are buddies or enemies
4. Analyze the consistency of agreements or disagreements of the cluster assignments of the pairwise points
5. Repeat the above steps for all the different k values you are interested in looking through



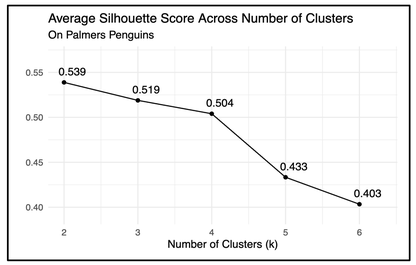




Comparison to Previous Methods (Elbow Method & Silhouette Score)

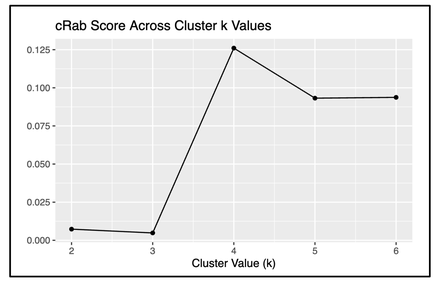


The Elbow method fails to reveal a clear inflection point or "elbow", leaving the choice of *k* to the user.



Although the Silhouette Score provides an objective answer at k = 2, it does not accurately capture the three distinct groups present in the data.

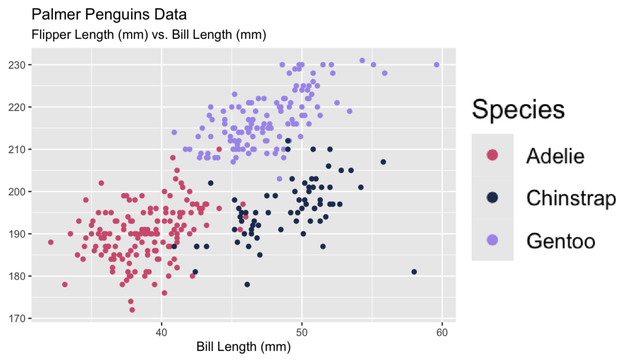
Using 100 resamples with 0.8 subsample proportion, the CRAB algorithm correctly identifies *k* = 3. A clear "pinching point" appears at *k* = 3, providing an objective criterion for selecting the correct number of clusters.



cRab Score Summary

k	Score	Time (seconds)	Best Cluster (k)
2	0.007319986	46.420	
3	0.004835052	44.765	BEST
4	0.126032053	42.935	
5	0.093206488	39.434	
6	0.093726209	36.254	

Results with Palmer Penguins Dataset



Discussion

CRAB is a viable alternative to other methods such as the Elbow Method and Silhouette Score for determining the optimal number of clusters. The CRAB algorithm is versatile, allowing it to theoretically work with any unsupervised clustering algorithm, although only the k-Means algorithm is supported at this time. Future work should involve implementing different clustering algorithms. Additionally, the algorithm should be made more time and memory efficient as the implementation currently scales quadratically in time and memory.

References

- Ben-Hur, A., Elisseeff, A., & Guyon, I. (2001). A stability based method for discovering structure in clustered data. *Biocomputing 2002*, 6-17. https://doi.org/10.1142/9789812799623_0002
- Jaeger, A., & Banks, D. (2023). Cluster analysis: A modern statistical review. *WIREs Computational Statistics*, *15*(3), e1597. https://doi.org/10.1002/wics.1597
- Kodinariya, T., & Makwana, P. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*, 90-95.
- Horst, et al., "Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises", The R Journal, 2022