CS 4641: Unsupervised Learning and Dimensionality Reduction

Bradley Reardon

March 24, 2019

1 Introduction

TODO

TODO datasets section

2 Clustering

TODO

2.1 k-means Clustering

TODO

Car data set did not do well, testing with a subset of the whole set (since the set covers all possible instances) results in much more meaningful clusters with better scores. However, still not great at clustering this data set. For comparison, we use n_clusters=4 since that matches the number of classes in the data, and clustering performed best on a subset with this value. Training time 0.117s

TODO figures, figure refs Figure 1 on the following page

Cancer data set performed much better, with a best-average silhouette score of 0.577 with n_clusters=2, fit time 0.044s. Matches number of classes in the data set conveniently.

TODO figures, figure refs Figure 2 on page 3

2.2 Expectation Maximization

TODO

2.3 Discussion

TODO talk about how the clusters aren't very well-defined in the charts, too many features.. try to reduce dimensionality in the next section to make that easier to see

3 Dimensionality Reduction

TODO

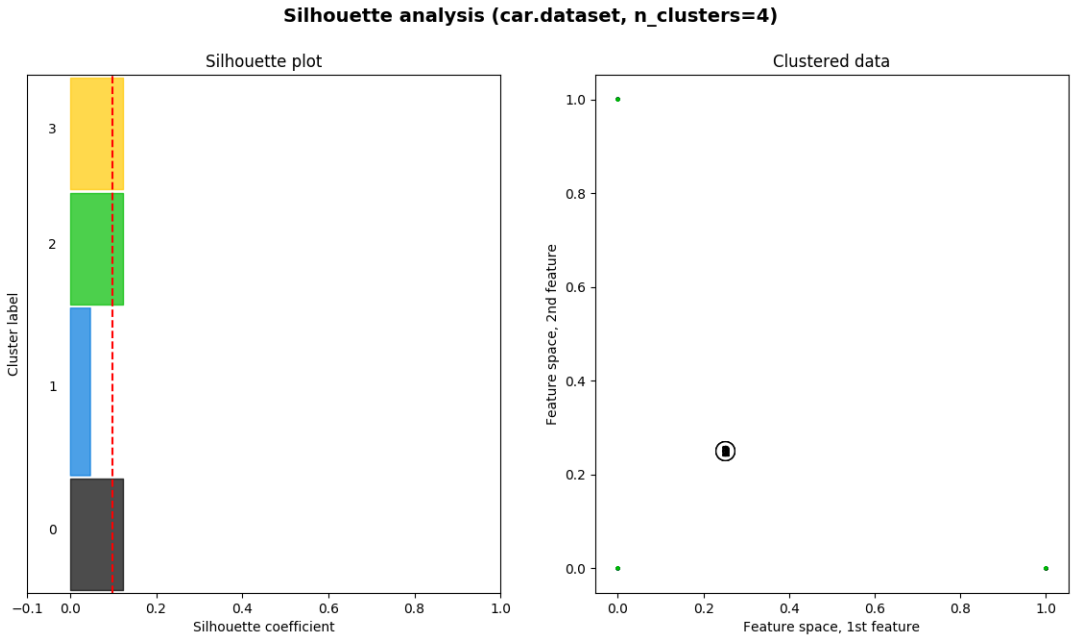

Figure 1: Silhouette plot and clustered data visualization on two features for the car data set.

3.1 PCA

TODO

3.2 ICA

TODO

3.3 Randomized Projections

TODO

3.4 TODO pick feature selection algo

TODO

4 Conclusion

TODO

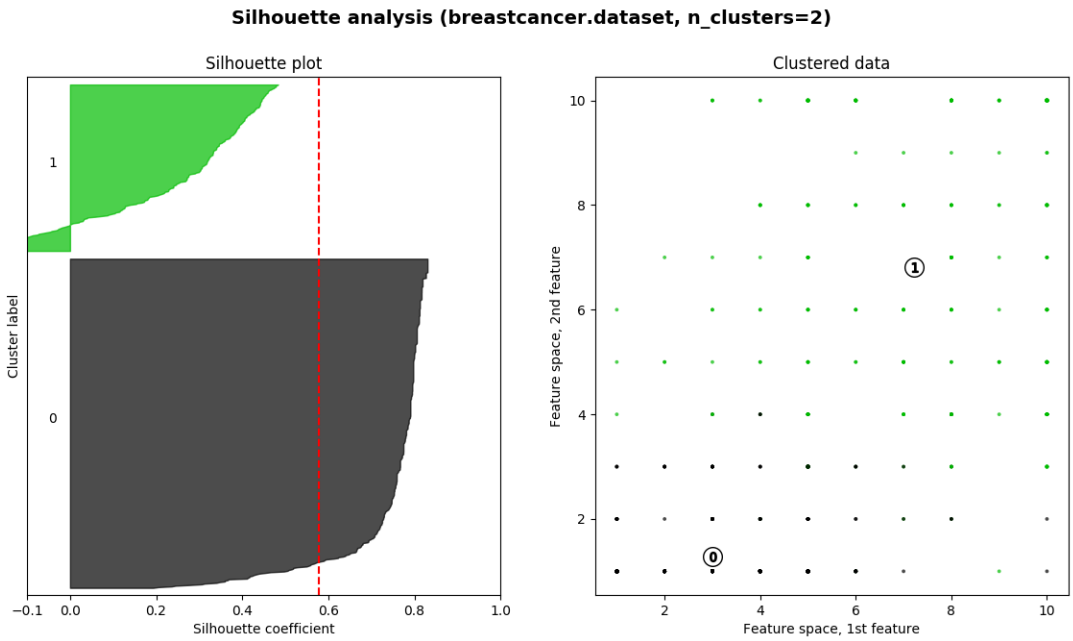

Figure 2: Silhouette plot and clustered data visualization on two features for the cancer data set.