

CS 4641: Unsupervised Learning and Dimensionality Reduction

Bradley Reardon

March 24, 2019

1 Introduction

In this assignment, we are tasked with implementing two clustering algorithms and four dimensionality reduction algorithms, and seeing how they perform when applied both separately and together on two data sets. In addition, we'll take a focus on comparing these results to those of previous assignments when run on the same data sets. For comparison, all tests run for the purpose of this report will run each algorithm over 500 iterations, with a fixed seed to ensure reproducibility across tests.

1.0.1 Car data set

The Car Evaluation Database was created in June, 1997 by Marko Bohanec. It contains 1728 instances and 6 attributes. The purpose of this database is to attempt to classify a car as an unacceptable, acceptable, good, or very good choice based on factors including cost of ownership, comfort, and safety. Full details about the data set can be found at the source link below. Note that the instances of this data set completely cover the attribute space, making it an interesting problem for testing overfitting.

Source: <https://archive.ics.uci.edu/ml/datasets/car+evaluation>

1.0.2 Breast Cancer Wisconsin data set

The Breast Cancer Wisconsin data set was donated to the UCI Machine Learning Repository in 1992, and contains data from one doctor's clinical cases, gathered from January 1989 to November 1991. In total, there are 699 instances signifying information about breast tumors such as clump thickness, uniformity in shape and size, and other screening factors. Data points are identified by two classes – benign or malignant. The features of the data points are encoded as 9 continuous attributes rating the screening factor from 1 to 10.

Source: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>

2 Clustering

Clustering algorithms are a method for unsupervised learning which attempt to place a number of instances into clusters based on the closest mean of input attributes to an existing prototype. Because clustering would occur in either 6 or 9 dimensions, the main focus of analysis will be scoring the algorithms for each data set, though visualizations of these clusters do still show interesting information.

2.1 k -means Clustering

2.1.1 Parameter selection

The main parameter for the k -means clustering algorithm is a value k , being the number of clusters that the algorithm will divide the data into. For both data sets, k values of 2, 3, 4, 5, 6 were tested for best performance. Each run of the

algorithm used 20 initializations to ensure that the best possible cluster labelling is chosen for each data set. Smart initialization of clusters using the “k-means++” method ensure initial prototypes outperform random chance.

The car data set did not perform very well in any of the k -means tests, with a maximum silhouette score of 0.119 being reported at a value of $k = 3$. However, the data was known to be separated into 4 output classes, and the completeness of the instances over the attribute space led me to believe that a subset of the data may perform better.

Running the algorithm again with only 500 of the 1728 total instances resulted in better silhouette scores overall, with the best performance occurring at $k = 4$. Figure 1 shows the silhouette plot and cluster visualization for this run.

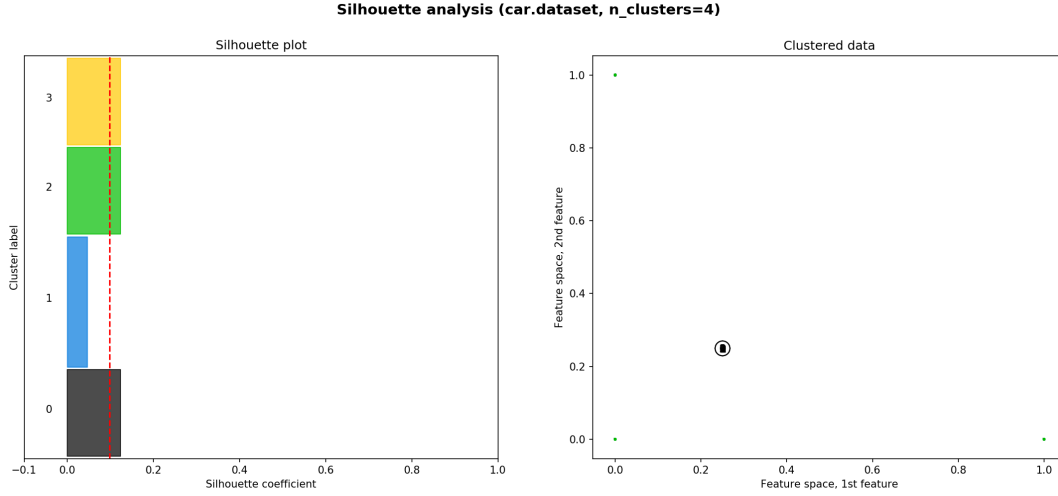


Figure 1: Silhouette plot and clustered data visualization on two features for the car data set.

The breast cancer data set was clustered better, with a best-performing $k = 2$, matching the number of output classes. This resulted in a silhouette score of 0.577, which is much higher than that of the car data set. Figure 2 shows the same visualizations for the cancer data set, which also reveals a very clear separation between the two clusters on the feature spaces for the first two features.

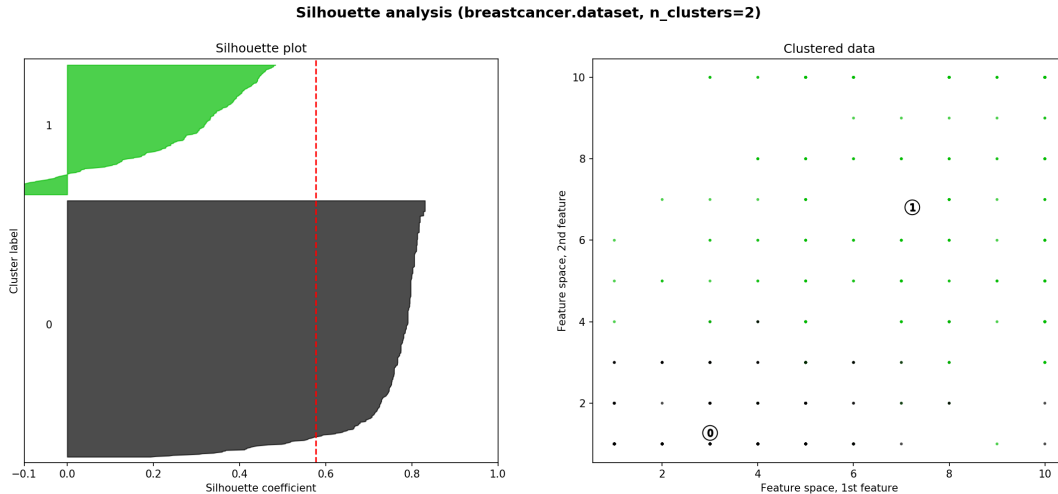


Figure 2: Silhouette plot and clustered data visualization on two features for the cancer data set.

2.1.2 Performance

Using the ideal k values of 4 and 2 for the cancer and car data sets respectively, learning curves were generated to evaluate the performance over the two data sets. These curves can be found in Figure 3.

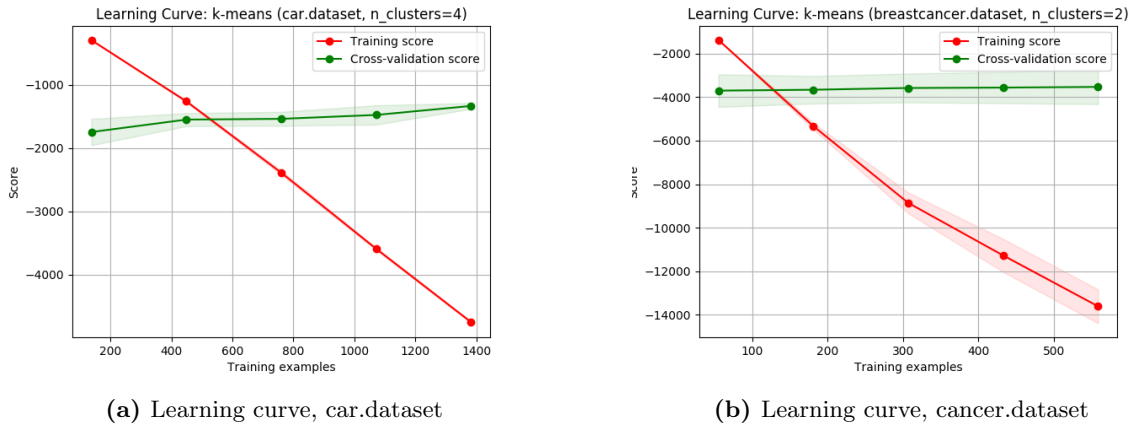


Figure 3: Learning curves for both data sets, using optimal parameters.

Both data sets seem to converge into clusters reasonably quickly, as shown by the low variance in the cross validation scores. Interestingly though, the variation in the car data set’s cross validation score narrows as the proportion of training examples used approaches 100%. The fit times of each test were recorded at 0.102s for the car data, and 0.044s for the cancer data. These running times will be compared to other algorithms and methods in a later section.

Though the cancer data set is clustered better than the car data set by k -means, this same drop in variance is not observed. This suggests to me that the completeness of the instances in the car data had some effect on that outcome. However, this will be looked into further when dimensionality reduction algorithms are applied prior to clustering.

2.2 Expectation Maximization

The expectation maximization algorithm was tested using gaussian mixture models, with negative log-likelihood as the performance metric.

2.2.1 Parameter selection

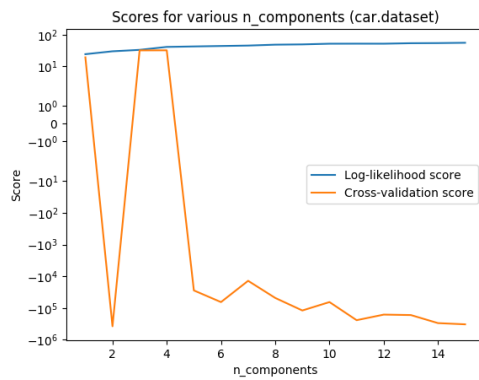
Expectation maximization, in this implementation, takes one main parameter, which is the number of “components”. Each algorithm was tested for best performance with between 1 and 15 components inclusive, as shown in Figure 4 on the next page.

In this case, the goal was to find the lowest number of components which resulted in the best scoring, as higher numbers of components increase running time. To satisfy this requirement, the car data set was chosen to have 4 components, while the cancer data set performed better with 6.

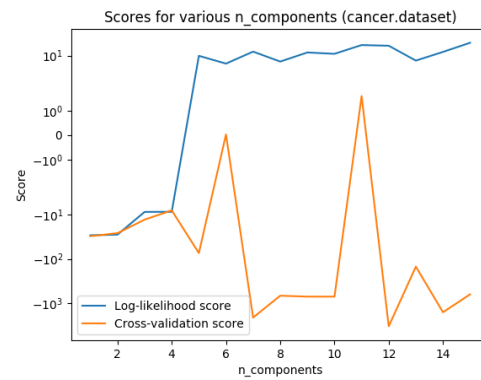
2.2.2 Performance

The learning curves for both data sets using expectation maximization are shown in Figure 5 on the following page.

The fit times for these classifiers were 0.025s for the car data, and 0.045s for the cancer data. Interestingly enough, both data sets have high variance in their cross validation scores at some point in training. This result will be discussed further after attempting dimensionality reduction.

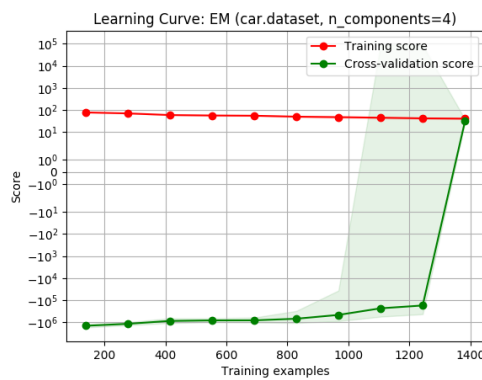


(a) Scores, car.dataset

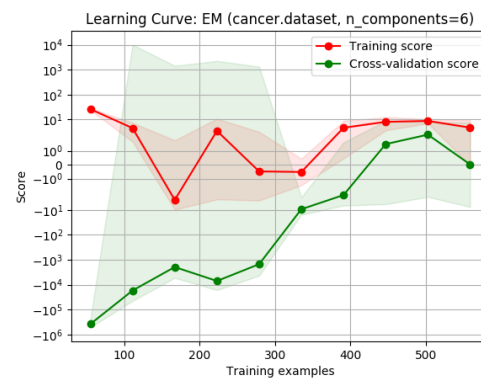


(b) Scores, cancer.dataset

Figure 4: Log-likelihood of classifiers for both data sets with varying $n_components$.



(a) Learning curve, car.dataset



(b) Learning curve, cancer.dataset

Figure 5: Learning curves for both data sets, using optimal parameters.

3 Dimensionality Reduction

TODO

3.1 PCA

TODO

3.2 ICA

TODO

3.3 Randomized Projections

TODO

3.4 TODO pick feature selection algo

TODO

4 Conclusion

TODO