Project 6 - Clustering

Blair Gemmer

CSCI 548 - Pattern Recognition

Spring 2013

Smiley Face Data

K-Means

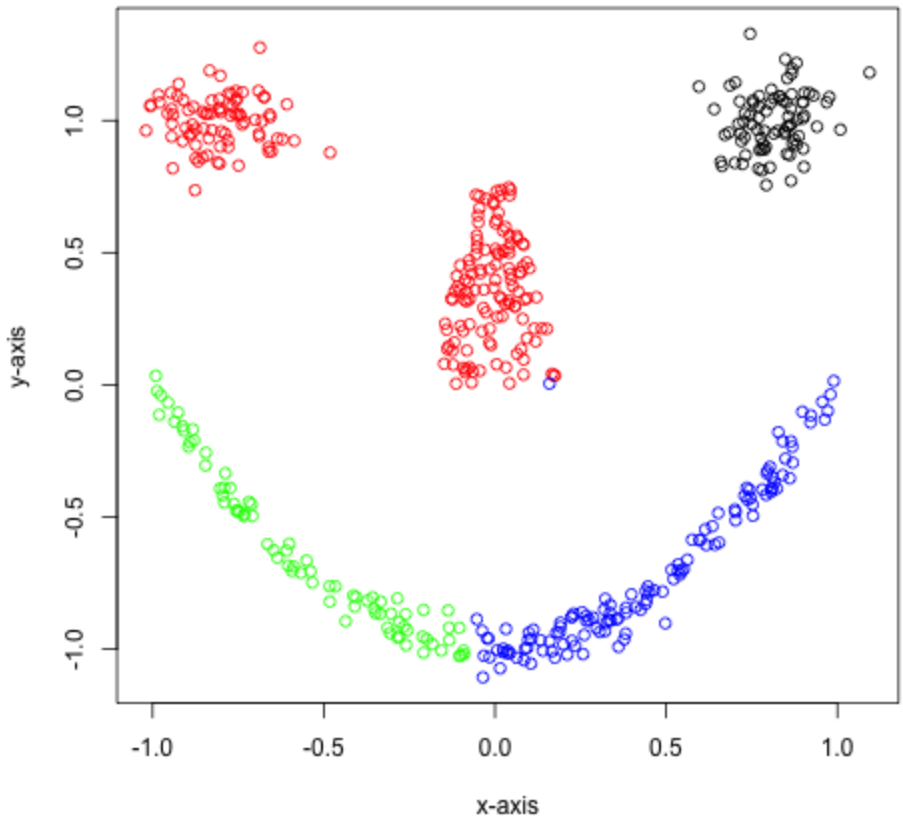

Figure 1. K-Means clustering on the Smiley Face Dataset. Colors represent different clusters. K=4 clusters.

Smiley Face Data
Hierarchical Clustering

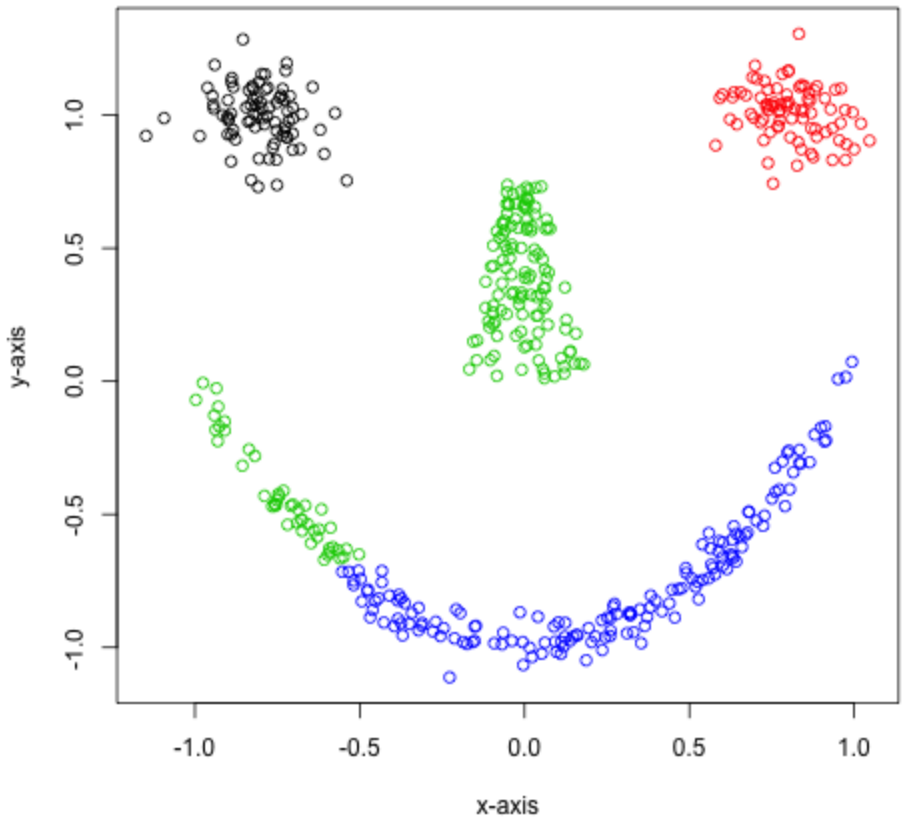

Figure 2. Hierarchical clustering of the Smiley Face Dataset. Colors represent different clusters. K=4 clusters.

Smiley Face Data

Hierarchical Clustering Using Warded Method

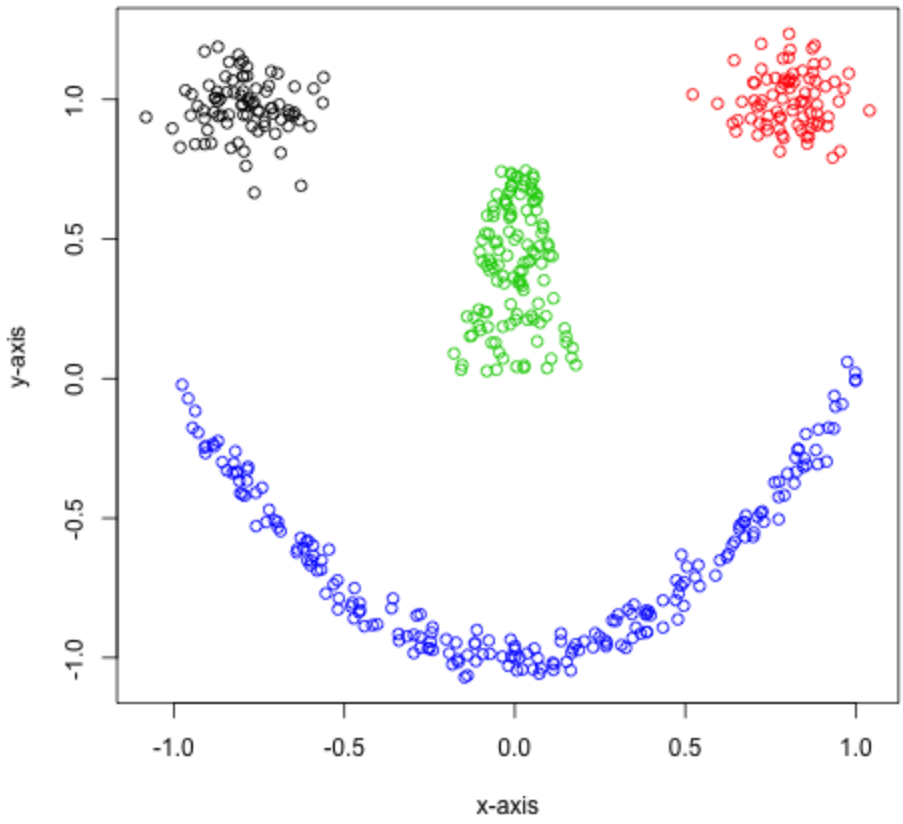

Figure 3. Hierarchical clustering of the Smiley Face Dataset using the ward method to cluster. Colors represent different clusters. K=4 clusters.

Iris Dataset

Basic Sequential Algorithmic Scheme (BSAS)

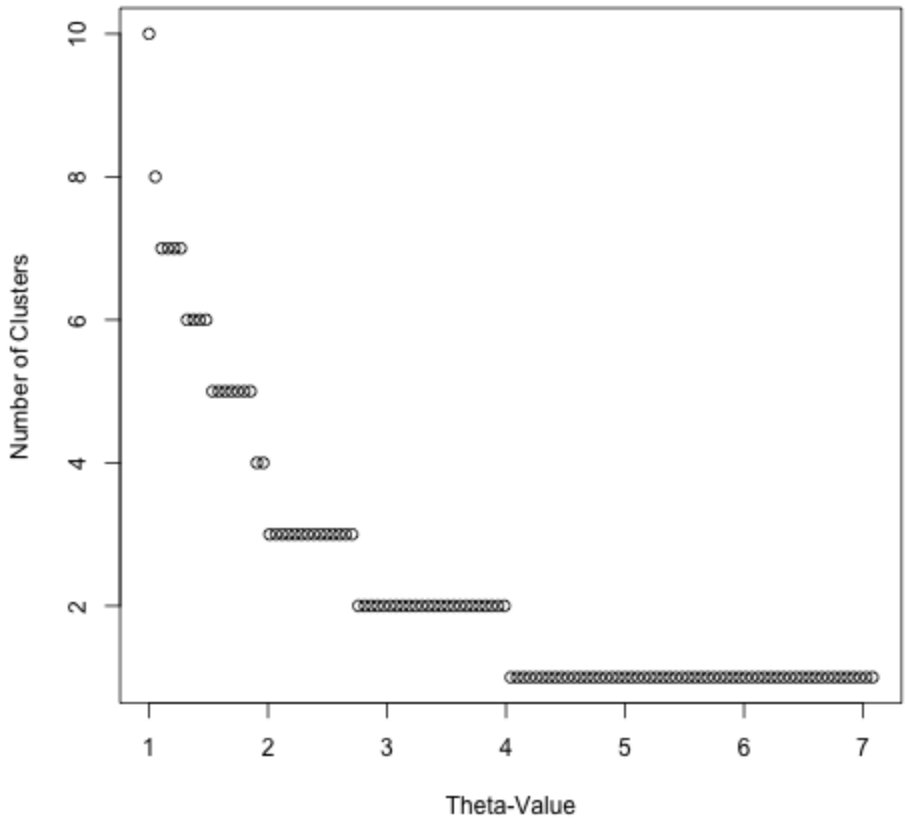

Figure 4. Number of Clusters vs. Different Theta Values for the Iris Dataset. Found each point by running BSAS on the individual theta values. Each point represents the number of clusters, based on the specified theta-value. Started at 1 instead of the minimum value, which was 0 (which would give us 150 clusters, 1 per row of the dataset). n=>100 points.

Iris Dataset

Basic Sequential Algorithmic Scheme (BSAS) using PCA

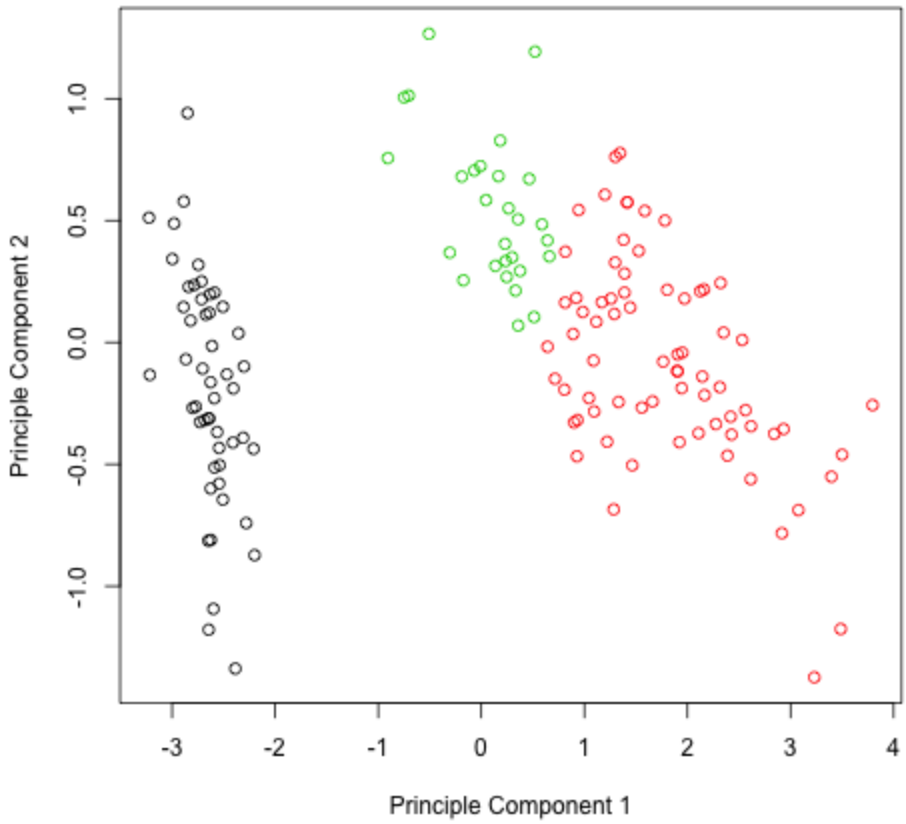

Figure 5. Principal Components Analysis run on BSAS of the Iris Data (theta = 7, k=3 clusters). This was run on principal components 1 and 2.

E. Coli Gene Expression Dataset

Heatmap

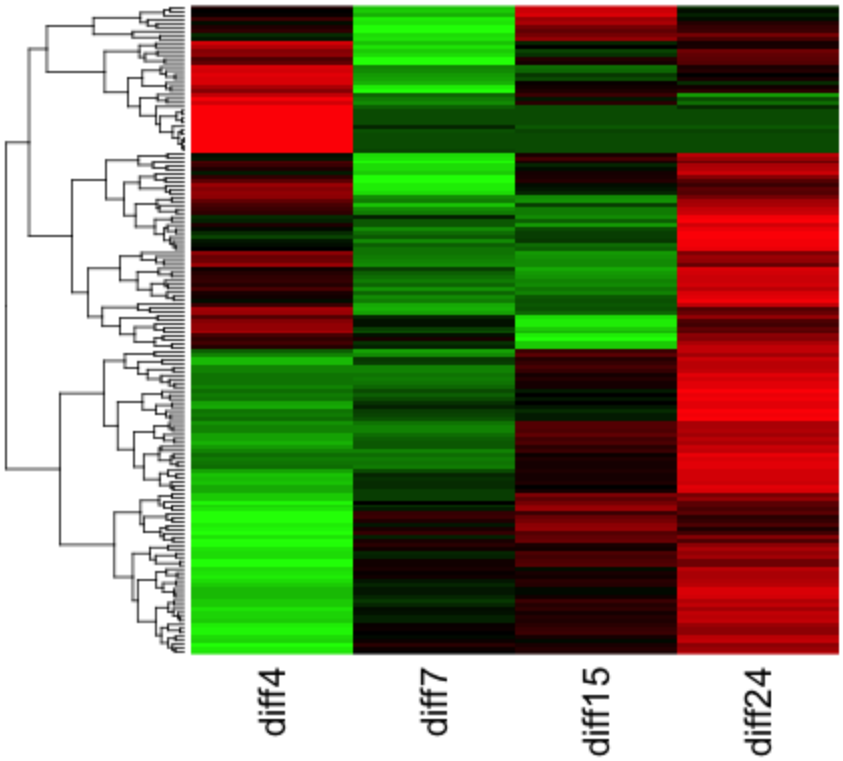

Figure 6. Red/Green heatmap of the E. Coli Gene Expression Dataset. The Y-Axis labels represent the hour when each sample was taken (4,7,15, and 24 hours). Each row represents the gene's differential expression profile (how much is expressed in biofilm vs. suspension).

E. Coli Gene Expression Dataset

PCA

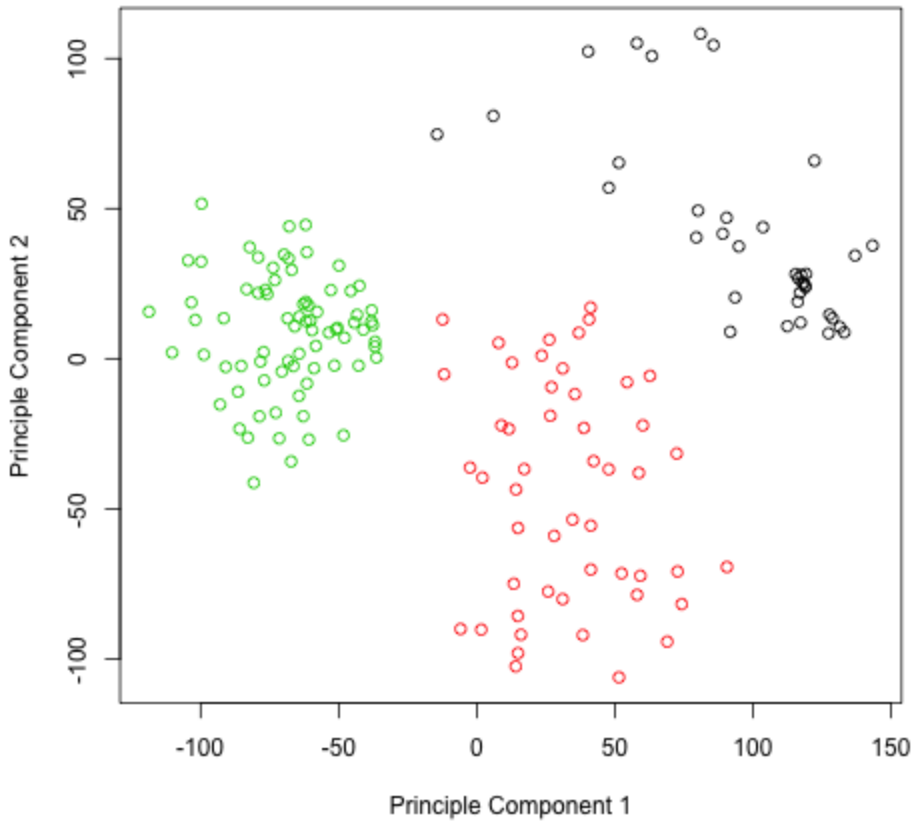

Figure 7. Principal Components Analysis run on E. Coli gene expression dataset. Colors represent different clusters. K=3 clusters.

E. Coli Gene Expression Dataset

Line Plot Time Series

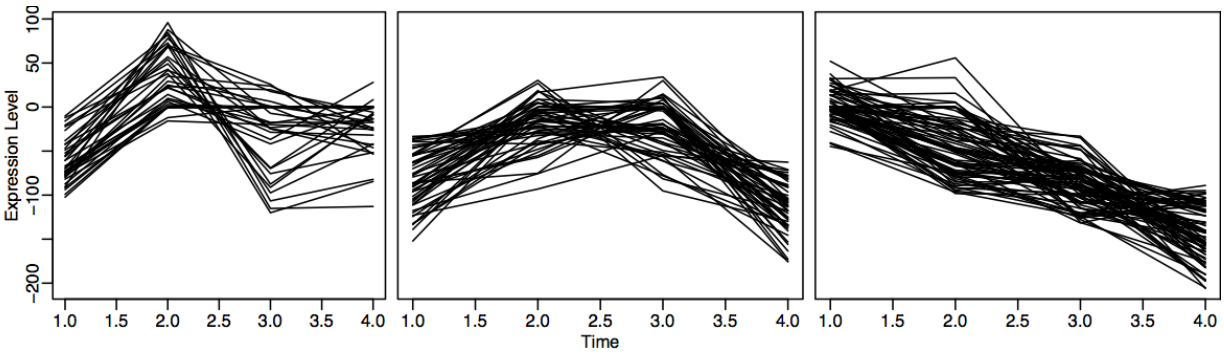

Figure 8. Line plot time series triptych of the E. Coli Gene Expression Dataset. The lines represent the expression level and the x-axis represents the amount of time.

Answers to Project Questions

1. Why does the clustering we see with the smiley face happen (the smile gets broken in half, and part of the nose is included with the right side of the mouth, and the left eye and most of the nose comprise the same cluster)? What about k-means causes this?
Since k-means chooses a random starting point to center the cluster on each iteration, it sometimes seems to pick up random points that aren't in the actual cluster. For instance, the eye cluster might accidentally center on the far right part of the nose or mouth.

2. Why did ward do better?
According to the man pages, "A number of different clustering methods are provided. *Ward's* minimum variance method aims at finding compact, spherical clusters." What I gather from this is that the ward method allows for a tighter clustering, so it finds the correct clusters most of the time (if not every time).

3. What was the number of clusters found in Iris Dataset? What was the theta required to achieve this number of clusters? Is it 3 clusters? If not, why not?
The highest number was 10, but the most reasonable number of clusters was either 2 or 3, based on the number of theta values that achieved those cluster numbers. If 3, it should be right on, but if 2, the reason is that there appear to be 2 very distinct clusters of data. If you color the data based on class, however, you see the 3 clusters easily. This is probably because the 2 furthest right clusters are very similar species of Iris, as far as the 4 dimensions go.