Samuel Moijueh
BTRY 4381 - Applied Bioinformatics Programming
Midterm 2013

K-means clustering algorithm, Pearson correlation coefficient, and the Elbow Method Using Perl

The K-means clustering algorithm was implemented in Perl in order to cluster 1000 genes based on their respective gene expression profile. The gene expression profiles were normalized over the cell cycle in HeLa cell line were obtained from Whitfield et al. The gene expression profile of each of the 1000 genes consisted of 114 different time points/conditions. The Pearson correlation coefficient was used as a distance metric. Each gene was assigned to the cluster with the closest center, or in other words to the center with the highest calculated Pearson correlation coefficient.

---> See the 'ssm87_BTRY4381_Midterm.pl' for more details. Please see the 'READ_ME.txt' file.

Problem 1: Implementation of k-means clustering [50 points]

In problem 1 of the midterm, the genes were partitioned into 10 clusters with the first 10 genes in the 'gene_expression.txt' file used as the intial centres. The K-means clustering algorithm was implemented iteratively to partition the 1000 genes into 10 mutually exclusive and exhaustive clusters in which each observation belongs to the cluster with the nearest mean.

The results of this k-means clustering is shown below. The 1000 genes converged into 10 mutually exclusive and exhaustive clusters in **31 iterations** of the K-means clustering algorithm.

See 'ssm87_BTRY4381_Midterm_problem1.txt' for more details.

```
Gene Center | Number of Points in Cluster | Within Cluster Sum of Squares
C8          137                           12265.0328002137
C2          127                           11147.224304689
C5          126                           11687.7148280105
C7          103                           8878.63356689513
C10         99                            9329.1077992224
C6          93                            8790.3866840686
C9          92                            8550.88422036132
C1          76                            7058.38768753915
C4          75                            7417.42898892368
C3          72                            6883.99418751809

DONE! Data points have convergenced in 31 Iterations!
```

Figure 1. Table describing the number of genes in each cluster (in ascending order) and the within cluster sum of squares of each cluster.

Problem 2: Find the optimal number of clusters by the elbow method [50 points]

In problem 2 of the midterm, the genes were partitioned into 20, 40, 60, 80, 100, 150 and 200 clusters using the first k genes in the input file as initial centers. The K-means clustering algorithm was then implemented to form k mutually exclusive clusters. The objective of this problem was to determine the optimal number of clusters by the elbow method.

The k-means clustering algorithm aims to minimize the within cluster sum of squares (WSS). The percentage of variance explained by clustering is ratio of between cluster sum of squares (BSS) and the total sum of squares (TSS). These quantities were calculated after each k clusters had converged as determined by the K-means clustering algorithm. A line plot, 'Percent of variance explained vs the Number of Clusters', was created to determine the optimal number of clusters by the elbow method.

See 'ssm87_BTRY4381_Midterm_problem2.txt' to see the file used to create the line plot.

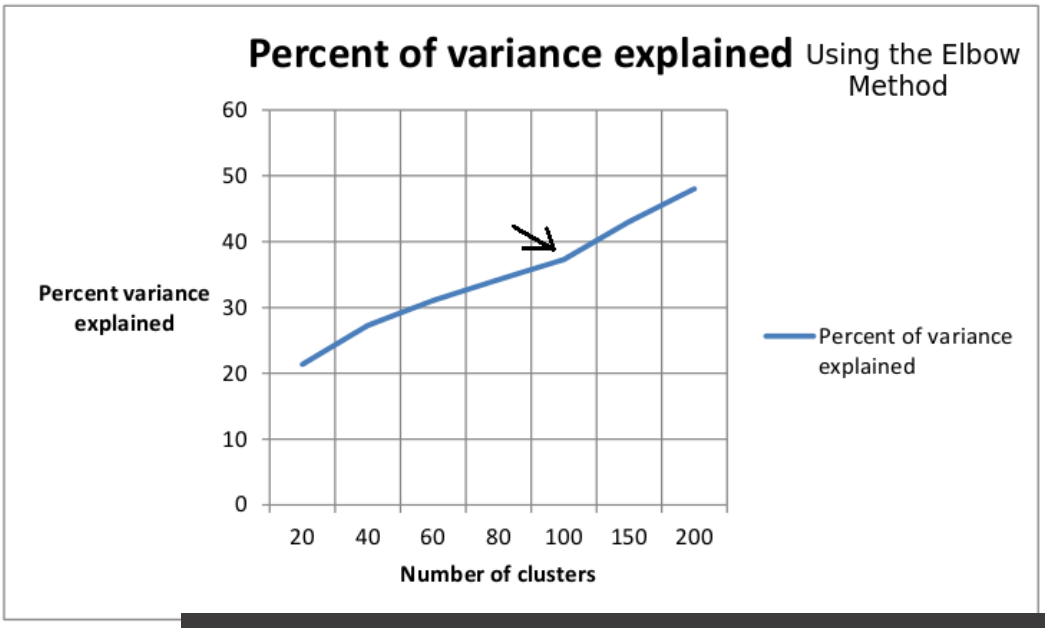

Figure 2 and 3. The line graph indicates that k = 100 is the optimal number of clusters.

Number of k Clusters	Percentage of variance explained
20	21.350638681717
40	27.2694807553267
60	31.0699406891405
80	34.2139994630639
100	37.2693066312966
150	43.0208955912858
200	48.0364330965085

See 'ssm87_BTRY4381_Midterm_problem2.txt' for more information.