Project: K-means Clustering **Lab: Beer**

Background: In genomics, it is common to use mRNA expression data to find clusters of similarly expressed genes to help understand their function. One very robust clustering algorithm is known as the k-means clustering algorithm. This algorithm takes as input the number of clusters to find, *k*, and the input data. The input data can be multidimensional data, but let's first consider gene expression data in three tissues, so each gene has three coordinates describing its normalized expression level in each tissue: x,y, and z. The kmeans algorithm is defined as follows:

1. Initialization: Set the *k* means to *k* randomly selected data points, *j*: $m_x^{(i)} = x_j$; $m_y^{(i)} = y_j$; $m_z^{(i)} = z_j$, for $i = 1..k$.

2. Assignment: Each data point, *j*, is assigned to the nearest mean, i.e. that $m^{(i)}$, $i = 1..k$ which minimizes: $\sqrt{\left(m_x^{(i)} - x_j\right)^2 + \left(m_y^{(i)} - y_j\right)^2 + \left(m_z^{(i)} - z_j\right)^2}$. So there is an indicator variable $r_j^{(i)}$ which is 1 when *i* is the nearest mean: $r_j^{(i)} = \begin{cases} 1 & \text{if } i \text{ is nearest mean to data point } j \\ 0 & \text{otherwise} \end{cases}$.

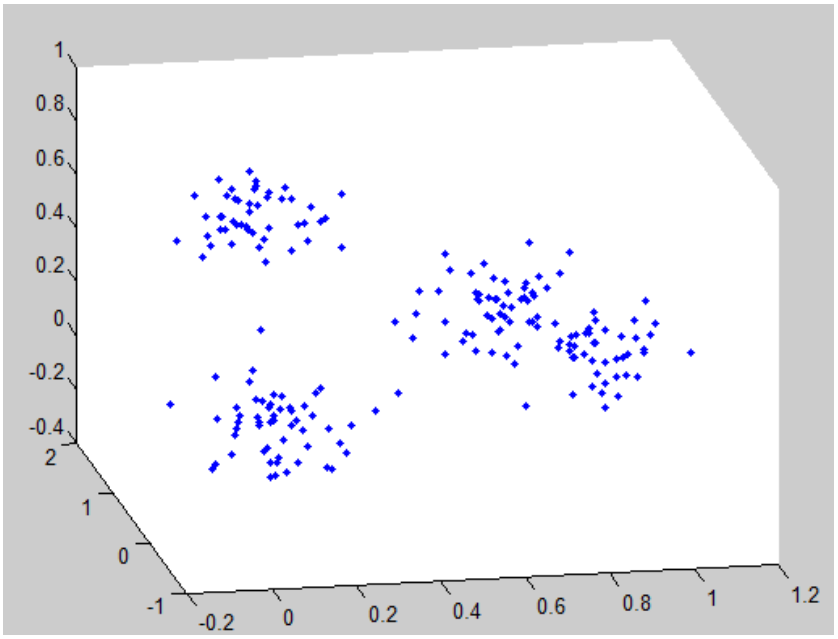
3. Update: Each of the *k* means are updated to be the mean of the data points that are assigned to it.

$$m_x^{(i)} = \frac{\sum_j r_j^{(i)} x_j}{\sum_j r_j^{(i)}};\ m_y^{(i)} = \frac{\sum_j r_j^{(i)} y_j}{\sum_j r_j^{(i)}};\ m_z^{(i)} = \frac{\sum_j r_j^{(i)} z_j}{\sum_j r_j^{(i)}}.$$

4. Repeat: Repeat the Assignment and Update steps until the assignments do not change.

Project:

1. Perform k-means cluster on the attached data, clusters.dat, which gives the x, y and z coordinates for a set of 200 data points representing gene expression in different tissues. Perform k-means clustering with k=2,3,4,5,6 and describe which k you think is most appropriate. What happens when k is not optimal? Plot the clusters with different colors representing cluster assignment to demonstrate.



2. Now use the attached human expression data, hnov_rnaseq4, and generate k=50 clusters using k-means. Use the DAVID functional analysis tool http://david.abcc.ncifcrf.gov/tools.jsp to find a few of your gene clusters with known common functions. Upload the gene lists using the ENSEMBL GENE ID.