# UCLA Department of Statistics
# STATS M254 Homework 2

Instructor: Jingyi Jessica Li
Due date: Friday, Feb 25, 2022 at 11:59 pm on CCLE

Please staple your homework and write down your name and UID clearly.

In this homework, we will use the `NCI60` cancer cell line microarray data, which consist of 6,830 gene expression measurements on 64 cancer cell lines. You need to install the `ISLR` package in `R` by

```
> install.packages("ISLR")
```

Then you load the package into `R` using

```
> library(ISLR)
```

Then you have the object `NCI60` in your `R` workspace. The object is a list of two elements `NCI60$data` and `NCI60$labs`, where `NCI60$data` is a numeric matrix of 64 rows (i.e., cancer cell lines) and 6,830 columns (i.e., genes) and `NCI60$labs` is a 64-element character vector containing the cancer types of the cell lines. For example, you can explore the data using the following commands.

```
> dim(NCI60$data)
> head(NCI60$labs)
> table(NCI60$labs)
```

## Problem 1: Hierarchical Clustering (20 pts)

1. Perform hierarchical clustering on the 64 cancer cell lines using all the 6,830 genes. Use the

$$d_{ij} = \frac{1 - \text{Pearson correlation between cell lines } i \text{ and } j}{2}$$

   as the distance measure. Plot three dendrograms using complete, single and average linkage. You can use `R` functions `hclust()` and `dist()` to do the hierarchical clustering.

2. Standardize the data such that every gene will have mean zero and standard deviation one across the 64 samples. This standardization will make every gene on the same scale. Then perform the hierarchical clustering as in Problem 2.1. Plot the three resulting dendrograms.

3. Perform quantile normalization on the original data, so that all the 6,830 genes have the same empirical distribution in every cell line. Then perform the hierarchical clustering as in Problem 2.1. Plot the three resulting dendrograms.

4. Cut each of the six dendrograms generated in Problems 2.1, 2.2 and 2.3 into four clusters using `R` function `cutree()`, which takes the output of `hclust()` as input. Compare the nine sets of clustering results, which one looks most reasonable to you?

5. Based on the results in Problem 2.4, please explain whether you think standardization should be performed, whether you think quantile normalization is reasonable, and which linkage is the most suitable in this case.

## Problem 2: $K$-means Clustering (20 pts)

Perform $K$-means clustering on the the 64 cancer cell lines using all the 6,830 genes. Use the Euclidean distance as the distance measure. Set $K = 4$. Use 20 random initial starts in the `R` function `kmeans()` by setting the `nstart` argument. Set the random seed as 1 every time before you run the algorithm using

```
> set.seed(1)
```

so your results can be reproducible.

1. Perform $K$-means clustering on the original data.

2. Perform $K$-means clustering on the standardized data from Problem 2.2.

3. Perform $K$-means clustering on the quantile normalized data from Problem 2.3.

4. Compare the three sets of clustering results from Problems 3.1, 3.2 and 3.3. Also compare them to the most reasonable clustering result you obtained in Problem 2.4.

5. Based on the results in Problem 3.4, please explain whether you think standardization should be performed for $K$-means, whether you think quantile normalization should be used before $K$-means, and which clustering method, $K$-means with Euclidean distance or hierarchical clustering with correlation distance, performs better in this application.

## Problem 3: Choose $K$ in $K$-means Clustering (10 pts)

Choose the number of clusters $K$ for the $K$-means clustering on the 64 samples using the original data and the same settings (20 random starts and random seed 1) as in Problem 3.1. Use the following two approaches.

1. Use the `R` function `silhouette()` in the `cluster` package. Show the silhouette plot fors $K = 2, 3, 4, 5$ and 6. Which value will you choose for $K$ based on the plots?

2. Use the `R` function `clusGap()` for computing the gap statistic in the `cluster` package. Show the gap statistic plot fors $K = 2, 3, 4, 5$ and 6. Which value will you choose for $K$ based on the plot?

*Hint: Follow the examples in the help files of these `R` functions.*

## Problem 4: Tight Clustering Algorithm (10 pts)

Apply the Tight Clustering algorithm in the R package `tightClust` to find 4 tight clusters among the 64 samples. Interpret the results. Are the results reasonable and more meaningful than the $K$-means clustering results?

## Problem 5: Justification of the $K$-means Algorithm (10 pts)

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ denote the expression levels of $n$ genes in $p$ samples, with $x_{ij}$ indicating the expression of gene $i$ in sample $j$. Let $C_1, \ldots, C_K$ denote the $K$ non-overlapping clusters, each containing a subset of $\{1, \ldots, n\}$, with $\cup_{k=1}^{K} C_k = \{1, \ldots, n\}$. Let $|C_k|$ denote the size of cluster $k$ and $m_k = (m_{k1}, \ldots, m_{kp})'$ be the center of cluster $k$. The objective function to minimize is

$$f(C_1, \ldots, C_K, m_1, \ldots, m_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - m_{kj})^2 = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2.$$

1. In the first step of the $(t+1)$-th iteration of the algorithm ($t = 0, 1, \ldots$), given the clusters from the $t$-th iteration $C_1^{(t)}, \ldots, C_K^{(t)}$. Show that updating the cluster centers as

$$m_{kj}^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^{(t)}} x_{ij}, \ j = 1, \ldots, p$$

satisfies that

$$f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right) \leq f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t)}, \ldots, m_K^{(t)}\right).$$

2. In the second step of the $(t+1)$-th iteration of the algorithm, given the cluster centers from the first step $m_1^{(t+1)}, \ldots, m_K^{(t+1)}$. Show that if we update the cluster membership of gene $i$ as

$$c(i)^{(t+1)} = \underset{k \in \{1, \ldots, K\}}{\arg\min} \sum_{j=1}^{p} \left(x_{ij} - m_{kj}^{(t+1)}\right)^2,$$

the resulting updated clusters

$$C_k^{(t+1)} = \{i : c(i)^{(t+1)} = k\}, \ k = 1, \ldots, K$$

satisfy that

$$f\left(C_1^{(t+1)}, \ldots, C_K^{(t+1)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right) \leq f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right).$$

## Problem 6: Practice of the Hierarchical Clustering (10 pts)

Suppose that we have four observations, for which we have a dissimilarity matrix

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.9 \\ 0.7 & 0.8 & 0.9 & \end{bmatrix}.$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

1. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchical clustering on these four observations using complete linkage. Be sure to indicate on the plot the height at which each merging occurs, as well as the observations corresponding to each leaf in the dendrogram. Cut the dendrogram to obtain two clusters. Which observations are in each cluster?

2. Repeat 1, this time using single linkage.

## Problem 7: EM Algorithm for the Gaussian Mixture Model (20 pts)

In the following Gaussian Mixture Model

$$X_i|Z_i = 0 \sim N(\mu_0, \sigma_1^2);$$

$$X_i|Z_i = 1 \sim N(\mu_1, \sigma_2^2);$$

$$Z_i \sim Bernoulli(\gamma),\ i = 1, \ldots, n,$$

where $X_i$'s are observable random variables, and $Z_i$'s are hidden random variables.

Given observed data points $x_1, \ldots, x_n$, derive the EM algorithm for estimating $\mu_0$, $\mu_1$, $\sigma_1^2$, $\sigma_2^2$ and $\gamma$ in the following steps .

1. Write down the complete log-likelihood $\ell(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$ in terms of $x_1, \ldots, x_n$ and $Z_1, \ldots, Z_n$ .

2. In the E-step of the $(t+1)$-th iteration $(t = 0, 1, 2, \ldots)$, derive the conditional expectation of $Z_i$ given $x_i$ and the current parameter estimates $\left(\widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right)$:

$$\tau_i^{(t+1)} = E\left[Z_i | x_i, \widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right] .$$

3. In the M-step of the $(t+1)$-th iteration, derive the updated parameter estimates based on $x_1, \ldots, x_n$ and $\tau_1^{(t+1)}, \ldots, \tau_n^{(t+1)}$.

$$\left(\widehat{\mu}_0^{(t+1)}, \widehat{\mu}_1^{(t+1)}, (\widehat{\sigma}_1^{(t+1)})^2, (\widehat{\sigma}_2^{(t+1)})^2, \widehat{\gamma}^{(t+1)}\right) .$$

4. Using $\mu_0 = 0$, $\mu_1 = 2$, $\sigma^2 = 1$, and $\gamma = 0.5$, simulate $(x_1, z_1), \ldots, (x_n, z_n)$ for $n = 500$. Use the full data $(x_1, z_1), \ldots, (x_n, z_n)$ to calculate the maximum likelihood estimates of $(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$.

5. Implement the EM algorithm you derived in 2 and 3 on $(x_1, \ldots, x_n)$ to estimate $(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$. Use initial parameter estimates $\widehat{\mu}_0^{(0)} = 0.5$, $\widehat{\mu}_1^{(0)} = 1.5$, $(\widehat{\sigma}_1^{(0)})^2 = 0.5$, $(\widehat{\sigma}_2^{(0)})^2 = 0.5$, and $\widehat{\gamma}^{(0)} = 0.3$. Display your estimates in the first 10 iterations. After how many iterations, are your EM estimates within $\pm 10^{-4}$ of the maximum likelihood estimates you calculated in 4 for every parameter?

## Extra Problem (30 pts)

Following our in-class discussion on the DESeq2 method, please read and evaluate the statistical method in MAGeCK, a bioinformatic tool developed for CRISPR/Cas9 data analysis:

`https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0554-4#Sec10`

Please write down your detailed comments and thoughts regarding each step of this method, including but not limited to: model formulation, parameter estimation, and hypothesis test. If you agree or don't agree with a particular step, please give your reasoning. You are encouraged to propose and write down alternative approaches for each step.