Day 04 - Exercise 02 - Clustering

Victor Calderon

16 August, 2018

Preamble

The exercises here are taken from the `clustering.R` file in the `scripts/day_04` directory.

In order to be able to run the code in here, you need to run the following:

```
## Installing packages
install.packages('ISLR')

## Loading packages
library('ISLR')
```

These packages are needed in order to run some of the commands.

We can now **load** the necessary libraries:

```
library('ISLR')
```

K-means section

Exercise 1

> Use the 'NCI60' data from the 'ISLR' package. This is a microarray data set with expression measurements on 6830 genes and 64 cancer cell lines. Treat genes as the input variable and perform PCA. Create a scree plot to examine the proportion of variability explained. How many PCs are needed to explain 80%, 90%, and 95% of the variability in the expression data (note that if p > n, there are at most p PCs)? Perform k-means clustering using the PC scores. Use only those PCs needed to achieve 80%, 90%, and 95% of the variability. Use k=4 to make four clusters. Explore how this PCA dimensionality reduction affects how the four clusters of cell lines group by cancer type.

First, you need to load the dataset `NCI60`.

```
library('ISLR')
data('NCI60')
```

Hierarchical clustering

Exercise 2

> Use hierarchical clustering to repeat the last part of the k-means lab question Use only those PCs needed to achieve 80%, 90%, and 95% of the variability. Use hclust and cutree to make 4 clusters. Explore how this PCA dimensionality reduction affects how the four clusters of cell lines group by cancer type.
