Khan Inan

Excercise 8

BI-GY 7633 Transcriptomics

What are the different ways one can come across a set of genes?

**There are several different ways that genes can be discovered. The main ways are through disease risk and family/adoptive studies. As long as a gene results in the development of a trait or phenotype, then it is worth discovering to researchers, and they usually use screening/marker methods.**

How can you learn from grouping genes with the same expression pattern?

**Using clustering methods to group genes with the same expression pattern allows us to determine how various treatment conditions can have an impact on gene expression levels**

What is the advantage, to the scientific community, for using GO-terms?
How can we use GO-terms to better understand our list of genes? What statistical method(s) are applicable for this analysis?

**The main advantage of using GO-terms in the scientific community is to provide consistency and a unified "language" for the representation of molecules. We can use GO to better understand our genes by making use of the geneontology database that has already made a collection of numerous resources. Filtering based on significant threshold is the main statistical method that can be applied to this analysis**

What measurement is needed in order to perform clustering?

**Basic euclidean distance between data points is required in order to generate a cluster**

What feature of gene expression is captured by Euclidean Distance.

**The feature that is captured by euclidean distance is essentially the amount of divergence within various profiles of expression**

What feature of gene expression is captured by Pearson Correlation.

**Pearson correlation mainly captures linear relationships within two variables used in gene expression**

How do we convert a correlation value to a distance value?

**We can use the formula d = 1-r to convert correlation value to distance where r is the correlation between two variables that are being analyzed**

What are the steps to create hierarchical clustering?

**Step 1: you calculate the proximity matrix using distance**

**Step 2: you assign the data to clusters**

**Step 3: you merge the clusters base on similarity**

**Step 4: you update the matrix for distance**

**Step 5: you repeat the previous two steps until you only have a single cluster**

What are the different linkage methods and what are they used for?

**The most widely used are:**

**Single-linkage: calculated the shortest distance in a pair of data points in two clusters**

**Complete-linkage: The distance is measured between two of the farthest pairs of data points in two clusters**

**Average-linkage: The average distance is calculated between two data points in two separate clusters**

**Centroid-linkage: The distance between the centroids of two clusters of interest**

What are the steps for k-means clustering?

**Step 1: determine the number of clusters which is k**

**Step 2: select k points randomly**

**Step 3: make k number of clusters**

**Step 4: Calculate the centroid of each cluster**

**Step 5: determine the quality of each cluster**

**Step 6: repeat the 3 previous steps**

What does the K in K-means mean?

**K is the number of clusters that is being used in the clustering algorithm**

What are the caveats of K-means clustering and how do you overcome them?

**The main caveats of K-means is the need for manual input of the initial values and how this affects the end results. The variance can be alleviated by selecting a k value that properly fits the dataset that is being observed**

Explain the silhouette width equation.

**The silhouette width equation is used to determine the fit of individual data points within the general classification. Additionally, it can be used to determine the quality of clusters**

Is it possible to get clusters if you cluster random data?

**It is definitely possible because even random data can show possible correlation in the form of clusters. It essentially comes down to luck and whether the data shows grouping or does not**