Lecturer: Jan Baumbach Term: Winter 2015/2016

Class: Introduction to Bioinformatics

**Exercise sheet – Bioinformatics Clustering**

1. What is the difference between supervised and unsupervised learning? Define clustering and name its advantages and disadvantages.

2. K-means is usually executed several times. Can you imagine why? How would you deal with the (potentially) different results?

3. The Silhouette Value of an object *i* in cluster *c* relates the average distance between object *i* and all other objects of *c* to the average distance between object *i* and all objects from the closest other cluster. Now think of a case, in which a large class (in the gold standard) has a single outlier, such that it is split into while clustering into one big cluster and one one-element-cluster (singleton).

(a) How does this affect the Silhouette Value?

(b) Think of a way to tune the Silhouette Value such that it becomes less sensitive to outliers?

4. Download and unpack the ZIP file from URL 1 and the Java implementation of Transitivity Clustering from URL 2 below.

Here's a description of the ZIP file's content:

java_sources.zip: Some java sources as template. Please fill the blank routines with own code (see tasks below).

ALB_ALT_AML.1000genes.res: A gene expression file for bone marrow cancer patients. Data structure: Each line starts with the patient ID. It is followed by a list of expression values for a set of genes. These values are TAB-separated.

ALB_ALT_AML.1000genes.pairwise_sims: Pairwise similarities of the patients computed with a normalized pairwise Spearman Correlation Coefficient. This function is symmetric. Each line consists of two patient IDs and one correlation value – separated by a TAB.

ALB_ALT_AML.txt: Gold standard (bone marrow cancer type of each patient). Each line consists of the patient ID and (again TAB-separated) a class identifier (0, 1 or 2). The class identifier reflects the cancer subtype.

Your tasks:

(a) Complete the implementation of k-Means

(b) Complete the implementations of F-Score and Silhouette Value

We aim to use clustering methods to unravel the cancer subtypes from the patients' gene expression data. Therefor:

(c) Use your k-Means implementation to cluster the gene expression levels files from URL2 into k=2,3,4 clusters.

(d) Use Transitivity Clustering to cluster the pairwise similarities file from URL3 with varying thresholds T = 0.22,0.255,0.4.

(e) Store the clustering outputs of the tools in six files. Use the following line and TAB separated format. One line for each cluster, headed by the cluster ID and followed by all patients in this cluster separated by a TAB. Illustration:

**ClusterID1 _TAB_ patientID1 _TAB_ patientID2**
**…**
**ClusterIDx _TAB_ patientIDa _TAB_ patientIDb _TAB_ patientIDc**

Hint: As clusterIDs, you may simply use an incrementing number.

(f) Use your implementations of F-Score and Silhouette Value to assess the validities of your six clustering results by comparing them against the gold standard file from URL 4.

(g) Which clustering is the best? Do the two indices, F-score and Silhouette Value, agree?

**Please send all your Java implementations (k-Means, F-Score, Silhouette Value) and the 6 clustering files in the above**

**mentioned format via email to your TAs. Also email the names of all group members.**

URL1:
http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws15-16/DM847/exercises/bioinformatics_intro_class_clustering.zip

URL2:
http://transclust.compbio.sdu.dk/downloads/binaries/TransClust.jar