Assignment 4.
Due: through course web (Moodle) on 11/10/2016, before 11:55 PM
Note: See the course web page for the late turn-in policy.
Note: Collaboration policy: see “Basic Information” link on the course web page for more details.

Problem 1. [25 points]:
Each of a set of 20 genes was tested for differential expression between cancer and normal tissues, using a statistical test where the null hypothesis was that there is no differential expression. The p-values of these 20 genes were found to be as given below, in ascending order. Which genes would you predict as being differentially expressed if you want to achieve an FDR (false discovery rate) of 20% ? (**Show your calculations for full credit**.)

1E-7
1E-6
1E-5
1E-4
1E-3
0.01
0.02
0.03
0.05
0.07
0.11
0.12
0.14
0.17
0.47
0.48
0.51
0.53
0.78
0.97

Problem 2. [25 points]:

The purpose of this problem is to familiarize you with online procedures for statistical testing with gene sets, a task whose theoretical basis was discussed in class. Exploring an online interface and finding out where and how to get your answer is the main task here, which is why the procedure is not described in more detail.

Your task is to see what insights can be found for two gene sets, by performing statistical tests of enrichment between each gene set and other pre-determined gene sets. There are several online tools that enable the discovery of associations between gene sets. You will use the popular tool called "DAVID" (http://david.abcc.ncifcrf.gov/) in this assignment.

Considering the background of all human (Homo sapiens) genes, use DAVID's online Functional Annotation Tool to discover which of all possible sets ("GOTERM_BP_ALL") of genes sharing a Gene Ontology Biological Process annotation is most significantly associated with a particular query set.

Be sure to use the newest version (6.8) of DAVID. (it is the default website at the above URL.)

For each dataset, **report the name of the association term ("Term"), the number of overlapping genes, the percent overlap, and the p-value.**

Datasets/subproblems:
MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) is a repository for several experimentally and computationally derived gene sets.

(a) "POOLA_INVASIVE_BREAST_CANCER_UP" is a set of 287 genes that were shown to be up-regulated in patients with breast cancer (ADHC) vs those without the cancer (ADH). (Reference: Pubmed ID 15864312).

We have downloaded this gene set for you and the "Entrez gene IDs" of the gene set are in the following file:

http://veda.cs.uiuc.edu/courses/fa16/cs466/assign/POOLA_INVASIVE_BREAST_CANCER_UP.txt

(This has been tested on the newest version of DAVID, if you get an error message when uploading the gene list, confirm that you selected the correct identifier type.)

(Hint: There are 128 genes in the "overlap set" of the best association for this dataset.)

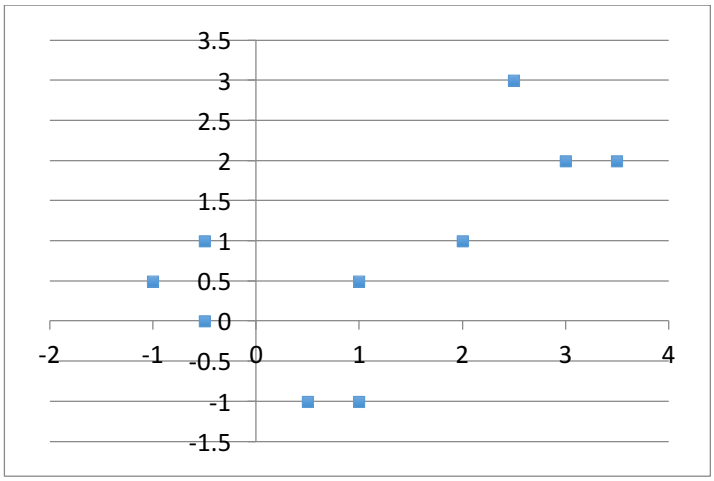
(b) Repeat the above exercise for the gene set "ACEVEDO_LIVER_CANCER_UP", available from http://veda.cs.uiuc.edu/courses/fa16/cs466/assign/ACEVEDO_LIVER_CANCER_UP.txt

This is a set of 973 genes that were shown to be up-regulated in hepatocellular carcinoma (HCC) compared to normal human liver samples (Reference: PMID 18413731).

Problem 3 [25 points]:

Consider the ten data points (in 2-D) listed below. A plot of the ten points is also shown below, for your convenience.

x	y
1	0.5
2.5	3
2	1
3	2
3.5	2
-0.5	0
-0.5	1
-1	0.5
1	-1
0.5	-1



Show the steps (and final result) of the Lloyd algorithm for K-means clustering for this data set, with K=3 and with initial cluster centers set to (0,0), (2,3) and (1.5,-1).

Problem 4 [25 points]:

Consider the five points listed in the table below. A plot of these five points is shown below, for your convenience. Show the steps (intermediate results in tree/forest form) and final result of the Hierarchical Clustering algorithm, as discussed in class, applied to this data set. Define the distance between two clusters as the minimum distance between a pair of points, one in each cluster.

Table:

	X	Y
Point 1	1	2.5
Point 2	1	2
Point 3	3	2
Point 4	3	1
Point 5	3.5	2.5

Plot of points in table