CS313 Exercise 6 Cover Page

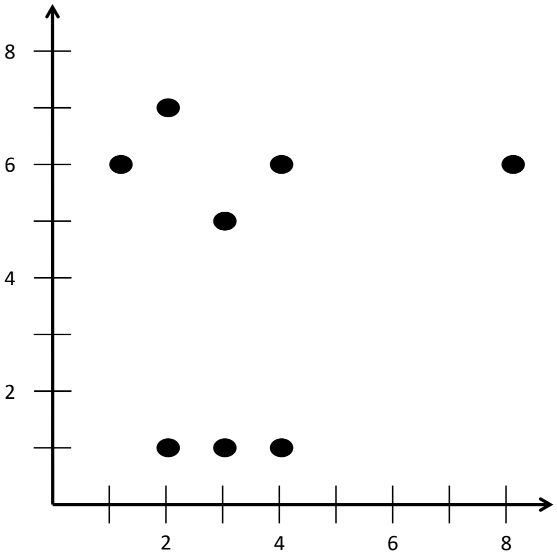
Name(s): ________________________________

In the *TIME* column, please estimate the time you spent on the parts of this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

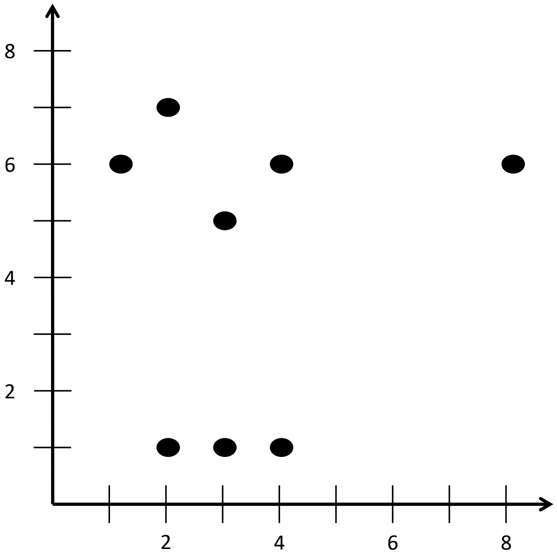
PART	TIME	SCORE
Exercise		

Task 1: Clustering with CAST

Consider the following eight points shown below with coordinates (2,1), (3,1), (4,1), (1,6), (3,5), (2,7), (4,6), and (8,6). Cluster these points with the CAST algorithm using an affinity threshold of 4.1 and starting with the point (8,6). How many clusters result and which points cluster together?



Now cluster these points with the CAST algorithm using an affinity threshold of 4.1 and starting with the point (4,1). How many clusters result and which points cluster together?



Task 2: Hierarchical and *K*-Means Clustering

If the hierarchical clustering algorithm is re-run, will it necessarily yield the same clustering result?

If the *k*-means clustering algorithm is re-run, will it necessarily yield the same clustering result?

In terms of *n*, the number of points/genes, and *m*, the number of dimensions/experiments, what is the runtime of the centroid linkage hierarchical clustering algorithm?

Why is *k*-means often re-run for multiple iterations?

In the remainder of this exercise, you will be looking at gene expression data. In particular, you will be searching for genes that appear to be similarly expressed (transcriptionally) as evinced by the expression assays. We will be using clustering algorithms, such as hierarchical clustering and *k*-means clustering, to find such similarly expressed groups of genes. For this assignment, we will be using the Morpheus site:

https://software.broadinstitute.org/morpheus

Task 3: Analysis of Gene Expression Data

To begin, you will need to download a file containing the results of gene expression assays. The `yeast.txt` file in the `Exercise6` directory contains expression data from a series of experiments performed on our beloved yeast organism. Go to the Morpheus site and open this `yeast.txt` file.

As this file is opened, click on the cell with value 0.33 so that only numerical data is contained in the blue cells. The first row (in red) contains the name of each experiment and the first two columns (in green) contain the names of yeast genes. You can confirm that Morpheus correctly opened the yeast file by checking at the top of the web page that it says "2,467 rows by 79 columns”. These numbers indicate that the yeast file contains information on the expression of 2,467 yeast genes as measured in 79 different experiments. A summary of the 79 experiments can be found here:

http://cs.wellesley.edu/~cs313/projects/project5/YeastCols.html

The rightmost column contains only the ORF name for each yeast gene. Since we also want to show the more descriptive annotation for each yeast gene, go to the “View” menu, select “Options”, click on the “Annotations” tab, and in the “Row annotations” drop-down menu, make sure both “id” and “NAME” are selected. Notice how the rightmost column now contains more descriptive information about each yeast gene.

Most of the window should contain a lot of red, white, or blue squares, corresponding to the measured expression level of each gene (row) in each experiment (column). Try moving the mouse over individual colored squares. Do you see the expression value for the gene in the relevant experiment as well as the gene name at the top of the window?

Are you ready to cluster? I can't hear you. ARE YOU READY TO CLUSTER? Ok, then. In the “Tools” menu, select “Hierarchical Clustering”. You can use the default options, such as “Metric” and “Linkage method”, but for the “Cluster” drop-down menu you should select “Rows and Columns” so that we will be finding groups (clusters) of similarly expressed *genes*, as well as finding groups (clusters) of similar *experiments*. Finally, click “OK” to perform the clustering.

When clustering completes, there should be a tree dendrogram at the top of the experiments (columns) and a tree dendrogram on the side of the genes (rows). Try selecting different subtrees (by clicking with the mouse) within these dendrograms to view groups of experiments or groups of genes that clustered together.

Let's search for a particular gene. In the search field at the top of the window, let’s search for the yeast gene *rpn12* that codes for a component of the 26S proteasome lid. When you search for this gene, a small blue bar should appear in the scroll bar for the genes at the far right of your window. Scroll down the rows to this blue bar to find the *rpn12* gene. Try selecting a branch of the gene (row) dendogram so that the *rpn12* gene and about 20 to 30 of its neighbors are selected (i.e., about 30 genes which cluster with the *rpn12* gene and appear to be similarly expressed in the 79 experiments).

Are the functions of these ~30 genes related to the function of the *rpn12* gene? What is the functional role of most of these ~30 genes?

Are there yeast genes which have a function similar to that of *rpn12* but which do not appear to be similarly expressed to *rpn12* in the 79 experiments?

In what type of experiments does *rpn12* (and those genes which are similarly expressed) appear to be more highly expressed (red) than control and less expressed (blue) than control?

Now search for the yeast gene *hxk1* that codes for a hexokinase. Are the functions of the genes that cluster with *hxk1* related to the function of *hxk1*?

Let’s close the tab, i.e., clear the window, and start over from the beginning but this time clustering with “KMeans Clustering” rather than “Hierarchical Clustering”. Don’t forget, in the “Options” to show the full gene name annotations, as before. When you perform clustering with the *k*-means algorithm, you can use the default “Metric” but cluster both the “Rows and Columns” into 20 clusters and let’s use a million maximum iterations rather than the default value.

When clustering completes, near the top of the window highlighted in gray should be a “k_means_20” at the beginning of a row of 79 colored squares. Since we clustered into 20 clusters, each of the 79 squares should be one of 20 different colors (each corresponding to a different cluster), though it’s hard to tell because the 20 different colors are similar shades of each other. If you click on the “k_means_20” gray rectangle, you can sort the experiments by which of the 20 clusters they were grouped into. Similarly, a second “k_means_20” gray rectangle appears on the right side of the window on top of a column of 2,467 squares that are each one of 20 different colors corresponding to the 20 clusters that the genes (rows) are grouped into. Again, clicking on this gray rectangle will result in ordering the rows by their cluster.

Again, find the *rpn12* yeast gene in the data and look at the genes that cluster with this gene. Do the same genes cluster with the *rpn12* gene in this case (using the *k*-means clustering algorithm) as in the case when you clustered the data using a different approach (hierarchical clustering algorithm)?

Try re-clustering the data a few times and experiment with different parameters or methods (e.g., use a different number of clusters in the *k*-means algorithm, or perform hierarchical clustering using "Complete linkage" or "Single linkage" instead of "Average linkage"). Do the genes which cluster with the *rpn12* yeast gene change? How confident are you in the clustered results?

Task 4: Cancer Classification using Gene Expression Data

One of the many challenges in diagnosing and treating cancers is that cancers that appear clinically similar can be genetically heterogeneous. Though a common feature of cancers is the loss of function of multiple tumor suppressor genes, pathologically similar cancers can result from different, independent gene defects. The different gene defects can have different implications for prognosis and treatment of the cancer. For this part of the assignment, we will be dealing with two different forms of acute leukemia, namely acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The two leukemias appear very similar morphologically. However, because the chemotherapy regimens differ for AML and ALL patients, the ability to distinguish between them is critical for successful treatment.

You will be analyzing gene expression data from experiments based on 38 patients with either AML or ALL. The experiments were performed by extracting RNA samples from bone marrow cells of the patients and assaying gene expression levels. The data is available in the `ALL-AML.txt` file.

The data corresponds to the measured gene expression of 654 human genes in 38 experiments (one experiment for each patient). Using Morpheus, you should cluster this data using the *k*-means clustering algorithm. Try clustering the experiments (columns) into 2 clusters using the "Metric" of "One minus kendall's correlation" with a maximum number of iterations at one million.

Have a look at how the clustering algorithm grouped the 38 experiments into 2 clusters. Do the AML patients predominantly cluster together in one of the groups and the ALL patients predominantly cluster together in the second group? Re-run the *k*-means clustering algorithm a couple more times and see if the results change (i.e., do the same patients cluster together in the 2 groups). Do your results indicate that these experiments can be used to distinguish between different forms of acute leukemia? If a new patient were diagnosed with acute leukemia, and if an experiment were performed on that patient's bone marrow RNA, how might the results of the new experiment be used to help guide the patient's diagnosis?

Finally, re-cluster your data using the *k*-means clustering algorithm again using the “One minus kendall’s correlation” metric and one million maximum iterations, but cluster the experiments (columns) into 3 groups rather than 2. Do one or two of your clusters correspond predominantly to AML? Do one or two of your clusters correspond predominantly to ALL?

The researchers who first performed these experiments (they clustered the data just as we have been doing in this exercise) found that ALL experiments tended to cluster into 2 of the 3 different groups. After examining the groups more closely based on immunophenotype data, they found that the 2 ALL clusters corresponded to patients with "T-lineage" ALL and patients with "B-lineage" ALL (cells which express different levels of particular antigens).

Task 5: Differentially Expressed Genes

As with the yeast data, hopefully you are displaying the full gene name annotations via "Options".

Are there particular genes which you see which seem highly expressed in AML patients and less expressed in ALL patients or vice versa? Try searching for the gene "adipsin" and for the gene "TCL1". Is the expression of these genes informative in determining different classes of cancer?

The researchers who originally performed this study had the computer choose only 50 genes that appeared informative, and they then clustered the expression data for these 50 genes. With these 50 genes (less than one tenth of the number we used in Task 4), they made no mis-classifications. Of the 50 genes chosen by the computer, most turned out to be closely related to the particular type of leukemia. For example, some of the genes are known oncogenes (c-MYB, E2A, HOXA9). Also, some genes (CD11c, Cd33, and MB-1) encode cell-surface proteins for which antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells.