BIF-31306: Hierarchical clustering

Maria Suarez Diez

Clustering is often used to analyze high dimensional data and large -omics datasets (transcriptomics, proteomics or metabolomics). For each variable (gene, protein or metabolite) there are measurements in multiple samples and the goal is to identify variables with similar behavior in the different samples.

Different clustering algorithms implement distinct measures of similarity. Moreover, in most of the cases there is a trade-off between the similarity of the elements in a cluster (also called within cluster similarity or *homogeneity*) and the dissimilarity between the elements in different clusters (also called between cluster similarity or heterogeneity or *separation*). There are many clustering algorithms that implement different choices. Arguably the most famous clustering algorithms in bioinformatics are k-means and hierarchical clustering. This assignment is about **hierarchical clustering**.

Hierarchical clustering takes as *input* an $n \times n$ distance matrix d that contains the distance between the points. Multiple definitions can be used for a distance. A distance fulfills (among other) the following:

- $d(I,J) \geq 0$ and $d(I,J) = 0$ if and only if $I = J$. The distance between two elements I and J is always positive and is zero only for the point itself.
- $d(I,J) = d(J,I)$ symmetric.

Examples of distances are Euclidean, Manhattan or Jaccard. A correlation based distance can be defined as $d_{corr}(I,J) = 1 - corr(I,J)$ where $corr$ represents the correlation (here we will consider Pearsons's).

Output: Once a tree or dendrogram is built (like the one shown in Figure 10.4 of the reading material) clusters are made by splitting the tree. We select the desired number of clusters and we cut the tree at the proper height to recover the clusters. The *output* of a clustering algorithm is an assignment of the data points to the clusters. This can be represented in multiple ways, some options are:

- Labels indicating to which cluster each point is assigned. For example, the sequence [0, 0, 0, 0, 0, 1, 1, 1, 1, 1] indicates the first five elements are in one cluster and the last five are in another cluster.
- Description of the elements in each cluster, using the order on which they were given or the labels. For example: cl1: [0,1,2,3,4] and cl2: [5,6,7,8,9] indicates that elements 0-4 are in one cluster and elements 5-9 are in another. The same can be indicated using the labels: cl1: [‘geneA’,’geneB’, ’geneC’, ’geneD’,’geneE’] and cl2: [’geneF’,’geneG’, ’geneH’, ’geneI’,’geneJ’]
- Groups of elements such as [[0,1,2],[3,4,5],[6,7,8,9]] or [[‘geneA’,’geneB’, ’geneC’],[’geneD’,’geneE’, ’geneF’],[’geneG’, ’geneH’, ’geneI’,’geneJ’]].

Important: Clusters are groups of elements, therefore providing a graphical representation of a dendrogram (tree) is **NOT** the way clustering output should be provided.

Reading Material Read sections 10.1, 10.2 and 10.3 of Jones & Pevzner. Section 10.2 presents hierarchical clustering and section 10.3 presents k-means that will be briefly discussed in the lecture.

Assignment

In this assignment you will implement hierarchical clustering, described in 10.2 of Jones & Pevzner. Three ways are given to compute the distance between clusters, make sure that you use the third one, which is the one the book explicitly indicates should be used.

Use the provided code skeleton to:

- Implement a function to obtain the Euclidean distance between two points $\mathbf{g1} = \{g1_1, g1_2, \cdots, g1_d\}$ and $\mathbf{g2} = \{g2_1, g2_2, \cdots, g2_d\}$ in a d-dimensional space, given by:

$$d_E(\mathbf{g1},\mathbf{g2}) = \sqrt{\Sigma_{i=1}^{d}(g1_i - g2_i)^2} = \sqrt{(g1_1 - g2_1)^2 + (g1_2 - g2_2)^2 + ... + (g1_d - g2_d)^2} \tag{1}$$

- Implement a function to obtain a distance matrix based on d_E for a set of points.
- Implement a function to obtain the correlation based distance between **g1** and **g2** given by:

$$d_{corr}(\mathbf{g1},\mathbf{g2}) = 1 - corr(\mathbf{g1},\mathbf{g2}) = 1 - \frac{d\Sigma_{i=1}^{d}(g1_i \cdot g2_i) - \Sigma_{i=1}^{d} g1_i \cdot \Sigma_{i=1}^{d} g2_i}{\sqrt{d\Sigma_{i=1}^{d} g1_i^2 - (\Sigma_{i=1}^{d} g1_i)^2}\sqrt{d\Sigma_{i=1}^{d} g2_i^2 - (\Sigma_{i=1}^{d} g2_i)^2}} \tag{2}$$

  Note that often the number of dimensions is named n instead of d.

- Implement a function to obtain a distance matrix based on d_{corr} for a set of points.
- Implement a function to perform hierarchical clustering on a distance matrix. Note that three expressions are given in to compute the distance between clusters and in section 10.2 it is stated that the third one is the one that should be used.

Hint In the skeleton a **Clusterset** class has been defined. It contains, for each cluster an identifier or ID, the clusters that fall to the left and to the right in the dendrogram and the distance between these nodes. The ID of the cluster is either -1 if it represents a branching point or contains the IDs of the points (for leave nodes). If you use this class, you will be able to use the functions **get_ordered_elements** (the elements as ordered in the dendrogram) and the **cut_tree** function to cut the dendrogram at a chosen height. Finally, the function **print_tree** produces a very simple representation of the dendrogram. You don't have to use it.

Questions

1. The file *jp_fig10_1a.csv* contains the data in Figure 10.1a of Jones & Pevzner. Use these data to:
   (a) Compute (and print) the Euclidean distance matrix. This should give the matrix in Figure 10.1b BUT the book contains an error. Which gene is not correctly represented in this matrix? (1pt)
   (b) Perform hierarchical clustering in this dataset (based on the Euclidean distance) and generate 3 clusters. (1pt)
   (c) Does your result match what you would obtain from cutting the tree in Figure 10.4 from Jones & Pevzner? (0.5pt)

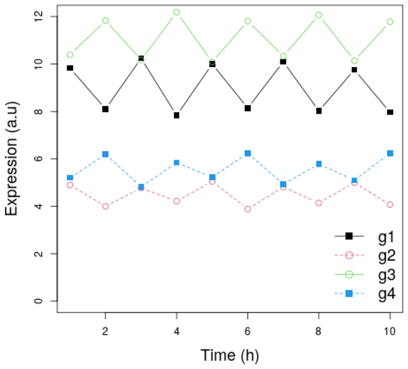

Figure 1: Graphical representation of the data in the file *example_gene_expression_four_genes.csv* with expression values (in arbitrary units) of 4 genes in a time course experiment.

2. The file *example_gene_expression_four_genes.csv* contains expression values (in arbitrary units) of 4 genes in a time course experiment. Data are shown in Fig 1.
   (a) Compute (and print) the Euclidean distance matrix for this dataset. (0.5pt)
   (b) Compute (and print) the correlation based distance matrix for this dataset.(0.5pt)
   (c) Perform hierarchical clustering in this dataset (using the Euclidean distance) and generate a set of two clusters. (0.5pt)

(d) Perform hierarchical clustering in this dataset (using the correlation based distance) and generate a set of two clusters. (0.5pt)

(e) Compare the results in 2c and 2d and explain the differences. (0.5pt)

3. Is the hierarchical clustering algorithm deterministic? (1pt)

4. What is the time complexity of the calculation of the distance matrix when using

   (a) Euclidean distance? (0.5p)

   (b) correlation based distance? (0.5p)

5. What is the time complexity of the hierarchical clustering algorithm you have implemented? (1p)

6. Sometimes datasets contain outliers, that is points with values (in at least some dimension) very different from the values of the other points. How would outliers appear when using hierarchical clustering? (1pt)

7. When analysing gene expression data, hierarchical clustering is often used to find groups of genes that cluster together. However it is also often used to compare the samples and verify that biological replicates behave as expected. The data in *proteomics_data.csv* contains proteomics data from an experiment where three conditions were tested in triplicate[1]. Analyse the data using hierarchical clustering and identify the three groups of replicates. Briefly explain what you did, which distance you used and why. (1pt)

8. **Optional** In chapter 10.2 of Jones & Pevzner three expressions are given for the distance between clusters. Modify your implementation to use d_{avg}. Note that $|C|$ and $|C*|$ indicate the number of elements in C and $C*$ respectively. (1 Bonus point)

[1] https://doi.org/10.1111/1462-2920.15311