Homework #2: Clustering

Download a data set GDS4168 from http://file.biolab.si/lectures/bcm-dm/GDS4168.tab. The documentation on this data set is available on GEO repository at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4168.

1. Cluster the data using hierarchical clustering. Use either Euclidean distance or cosine similarity to assess the distance between human subjects. Use Ward linkage to determine the distances between clusters. How well do resulting clusters correspond to patients' phenotypes? Which of the two distance matrices should we use? Evaluate the results qualitatively by displaying the dendrogram. Provide also any quantitative estimate of the correspondence between the clusters and phenotypes. You may use the Box Plot widget to select cluster for subgroup and class as the variable for the latter.
2. Use k-means clustering on this data. What is the proposed number of clusters according to the silhouette score? Report on the correspondence of clusters to the patient phenotypes. Again, you can use the Box Plot to quantify the correspondence.
3. Comment on the three approaches (hierarchical clustering with two different distance metrics and k-means clustering): if there are any differences, what are they, and what do you think are the reasons for them?

You may want to additionally explore the differences between these methods using the t-SNE widget. We have not described its functionality yet. For now, it should suffice to know that the widget can map the multi-dimensional data in two dimensions. The data items that are close to each other in the original space will, in t-SNE visualization, be close to each other in the resulting two-dimensional map.

Submit the homework as a short report in PDF where you answer to above questions. The report should include the title of the homework, your name, and your email. It should be one page long (this limit is strict!), use 11 pt Arial or Calibre or similar, with line spacing of 1.2. It should start with a short paragraph describing the data set. It should also include the screenshot of the workflow with the content of any widgets that support the answers to the above questions. Submit your homework as a PDF document to bzupan@gmail.com with the subject "DM-HW2". Name this document as lastname-firstname-2.pdf (like smith-mary-2.pdf; notice there are no spaces in the name) where lastname is your last name and firstname your first name. The deadline is 11:00 am on Thursday, February 10.