Saptarshi Ghose
Machine Learning in Cancer

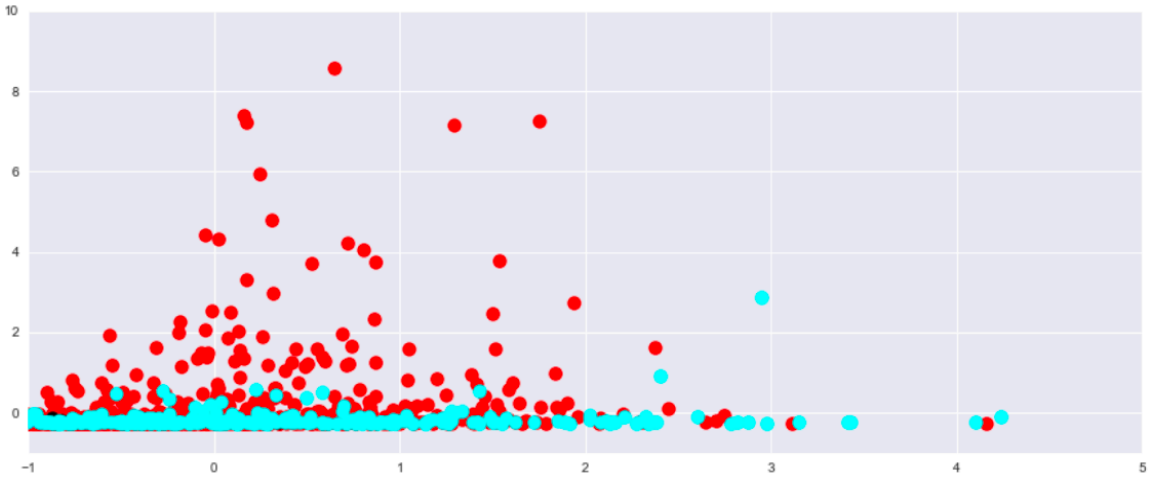
Lab 2 Clustering Write-up

In order to prepare the data for clustering, I first retrieved the relevant data by connecting to the Genomic Data Commons API through the TCGAbiolinks package in R (R script attached). This involved separately querying RNA-Seq data for breast cancer and ovarian cancer, downloading the gene expression data using the GDCdownload method, summarizing the results, transposing them, and writing them to CSVs. At this point, I read in the data as two separate Pandas dataframes in Python (Jupyter notebook attached), normalized the data using StandardScalar preprocessing methods in sklearn, and ran multiple clustering algorithms on the feature sets.

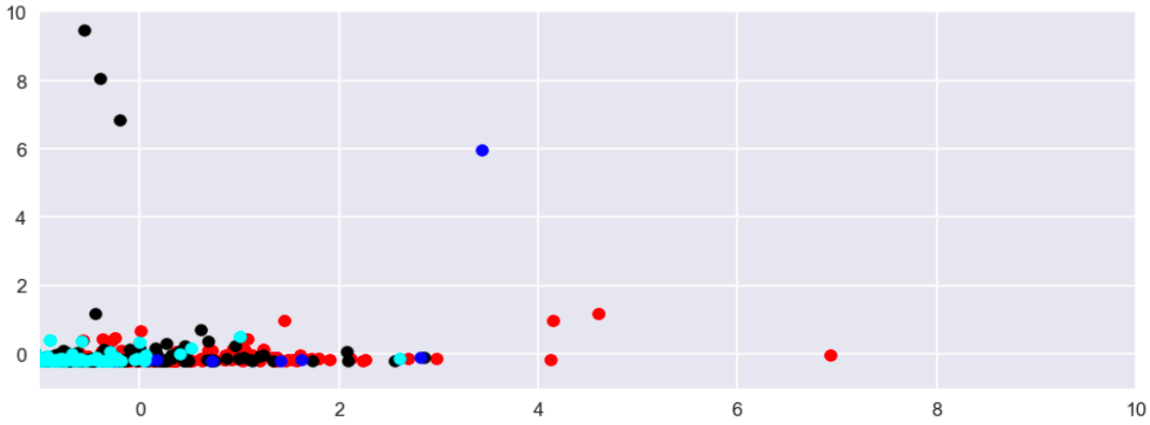
The clustering algorithms I used were K-Means, Agglomerative, AffinityPropogation, and MeanShift. I chose these methods because of their relatively low computation costs – and because they seemed appropriate based on their relevant documentation (https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68). For K-means, which required manual imputation for values for K, I used the so-called "Elbow Method" for finding the optimal K b based on finding the elbow inflection point for sum of squared distances (https://en.wikipedia.org/wiki/Elbow_method_(clustering)). For the breast cancer data, the elbow method results showed that the optimal K = 5. For the ovarian cancer data, the elbow method results showed that the optimal K = 6. Further, I visualized the results of each of these algorithms using scatterplots colored by cluster – with

the exception of Agglomerative Clustering (for which I visualized the results using a dendrogram because it is a hierarchical clustering method).

From here, I proceeded to analyze the clustered results and compared the outputs between the breast cancer data and ovarian cancer data. I noticed that the clusters, across algorithms, seemed to be much more visually defined and separable for the breast cancer data vs. the ovarian cancer data. Illustratively, the Agglomerative Clustering method applied to the breast cancer data showed good separation:



On the other hand, the same algorithm applied to the ovarian cancer data did not produce results that were nearly as promising visually:



The same pattern held true for KMeans clustering methods in the breast cancer data:

vs the ovarian cancer data:

To improve upon this analysis in the future, I would seek to understand more, a priori, of what successful clusters of the relevant gene expression data here might look like (vs. evaluating groupings after the algorithms were run). I would also seek to run additional clustering evaluation metrics (beyond the ones I ran in the notebook) – such as the Silhouette Coefficient (https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-

evaluation) to gain a better quantitative understanding of the performance of the clustering methods. Visual analysis was helpful in evaluating the clusters, but more descriptive evaluation metrics would add another important dimension to further analysis (similar to precision, recall, and F1 metrics for labeled machine learning methods). I would also perform Principle Components Analysis to aide in dimensionality reduction of the unsupervised methods applied to the datasets (http://airccse.org/journal/ijdms/papers/3111ijdms13.pdf). Beyond that, to improve the clustering in the future, I would seek to run additional clustering algorithms that are more attuned to dealing with noisy data, like Hierarchical Density-Based Spatial Clustering of Applications with Noise (which does not require imputation of number of cluster a priori), to see if clustering performance improves.
