QBIO490 Machine Learning Written Responses

1. Looking over your figures, does anything surprise you? Why or why not?

The biggest surprise in the generated figures stems from Figure 3 (PCA plot). This is because I expected there to be a more clear separation between groups/stages of cancer, but it seemed that the groups overlapped significantly and didn’t show any real distinctions (as seen by the colors that overlap with each other in the plot). I also found it interesting that the elbow plots that were generated seemed to have a reasonable “dip”, making the optimal number of clusters easier to determine (Figure 1).

1. Now that you have clusters, what information would you like to know about each cluster? How would you get this information?

Given these clusters, I would like to know about how each of the other clinical variables plays into these groups. For example, I would like to explore how an age cluster might differ from a stage of cancer cluster (which is what my plot covered), and whether there would be similarities between these groups. To do this, I could plot another k-means graph using age\_category instead of stage of cancer (this data would come from CPTAC) and overlay my two plots to observe any potential differences.

1. Brainstorm two ways you could combine RNA and protein information into one figure. Provide two sketches of these figures.

One way of combining these two datasets is by relating the RNA counts for a single patient and the protein up/down-regulation for that patient. You would connect the two using a single barcode, and the expected figure would be a scatterplot that has RNA counts on the x-axis and protein regulation on the y-axis (not sure how to sketch the figure because I’m not 100% what to expect if I were to graph this – but it would be in scatterplot form). This way, you might be able to see whether RNA of a certain gene correlates to protein production of that gene – each point represents a gene in this graph. Another way of combining these datasets is by selecting a single gene and comparing the way this gene’s protein and RNA presence differs in all the patients in the dataset. You would need to relate the two datasets with the gene name, and graph a scatterplot with the x-axis being RNA presence and the y-axis being the protein regulation – each point represents a patient in this graph (again, not sure how to sketch but I’ve done my best to describe the plot appropriately, and it would be in the form of a scatterplot).