BINF 610 Applied Machine Learning (Spring 2023)
HW3
Due: April 25

PCA + Clustering

In this assignment, you are asked to do clustering of protein cellular localization data, hoping that the amino acid composition features of proteins are somehow related to their cellular localization, and therefore proteins of the same subcellular locus are grouped in to the same cluster.

The data contains protein sequences in FASTA format, with the subcellular locus information in the header line. Your task in this assignment is to use the dipeptide features for clustering. To do so, you need to first convert each protein into a 400-dimension vector, with each component of the vector representing the occurrence frequency of the corresponding dipeptide (amino acid pair) in the protein sequence.

1) Use k-means clustering algorithm on the dataset. Use inertia and Silhouette score to identify an optimal k. Show plot like Figure 9-8 and Figure 9-9.

2) Repeat part 1 on the data with reduced dimension (2-dim) by PCA. In addition, do the following.
- What is the explained variance ratio respectively for the first 2 principal components?
- Draw the plot showing cluster boundaries (Voronoi tessellation) like Figure 9-3 in the textbook.
- Discuss the clustering quality as compared with visual inspection.
- Whether the optimal k and the corresponding inertia and Silhouette scores from part 1 with full dimension and from part 2 with reduced dimension are consistent.

3) Evaluate the clustering performance. For k = 10, evaluate the clustering results as compared with the ground-truth labels. For each cluster, label it with the locus having the maximum number of proteins in the cluster. Then, produce the confusion matrix for the 10 subcellular loci.