Harnoor Singh

Machine Learning

HW2

1. The best approach according to me would be converting the DNA string into a 2D array where the letters ACTG are assigned to some arbitrary 2D value (1, 0), (0,1) etc. Calculating the Hamming distance using the same values will be more useful and lead to better results for the next problem.
2. Result seems to be equal with minor differences. The exact positions might be different but the clustering is pretty closely accurate or similar. The reason is

If we will use Eucledian distances in MDS then graph would be identical to PCA graph. Minimizing the linear distances (MDS) is same as maximizing the linear correlation (in PCA).

1. Best value of k is 3 as seen from PCA and MDS graphs. Both are pretty similar and have good results.

There might be some cases the result might be different for MDS and PCA

When we have data in the form of genes or dna I think MDS might have better results because we can have all type of distances such as log distances and so on. (Our case: hamming distance). Leading to more real life oriented results as compared to Covariance difference in PCA. Making MDS more flexible.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated