Dimensionality reduction and clustering tutorial

In this tutorial, we will use PCA and a couple of clustering techniques (hierarchical and k-means).

1) First, as a warm-up, we will look at a dummy genetic data set, for which I simulated 4000 loci of DNA that could differ among individuals, listing the sequences for a couple, then producing 3 generations of offspring.

   On Latte download the dummy genetic data.
   The file g4.mat contains the sequences of the final generation.
   The file g.mat contains the sequences of all individuals from all 4 generations.
   Each column is an individual, each row corresponds to one base.

   The goal is to build a family tree, aligning each individual in g4 or g with family units.

   a) The MATLAB command "linkage" and "dendrogram" will be needed to produce hierarchical clusters of the 4^{th} generation individuals.

      You should also carry out PCA on the 4^{th} generation individuals and see (by using plot3 the first 3 principle components) how the clustering appears.

      Finally, you can use kmeans (try different numbers of clusters) to see how well 4^{th} generation individuals are assigned by family.

   b) Repeat a) but by using the entire set of individuals across generations (g.mat) to assign ancestors to each family. (For ease, the increasing generations from the 1^{st} increase monotonically with row number in the matrix).

2) The folder "Taste Data" contains an array with neural spike times from multiple trials of different taste stimuli. If you run the codes "open_taste_data.m" then "reformat_taste_data.m" you will produce an array (analysis_data) with each row representing the number of spikes by a given neuron in a given time bin in a particular trial. Each row is a separate trial to be categorized.

   An array of taste ids (id_data) is also produced.

   Look at the codes to see how this is achieved.

   The goal of this assignment is to figure out a method that best groups sets of trials according to which taste stimulus was used. Of course, the grouping must be carried out using only the array "analysis_data" and then you can test whether trials with the same taste stimulus are grouped together by comparing with the array "id_data".

You will use kmeans to cluster data, and/or the hierarchical clustering (for example the latter may work well if palatable tastes cluster together and unpalatable tastes cluster together).

Use PCA to plot in 3D the first 3 principle components of the data, color coded according to the clusters you have sorted the data into.

The code “reformat_taste_data.m” can be modified so that the spike train from a single neuron can be counted in separate bins (with each bin then treated like a separate neuron’s spike count on that trial). Try running the code with different bin widths and bin separations until you get an optimal clustering of the taste stimuli, using either the kmeans or the hierarchical clustering and devising a metric that indicates how well the data cluster correctly.