Homework #1 K-means

In this homework you will use K-means clustering to try to diagnose breast cancer based solely on a Fine Needle Aspiration (FNA), which as the name suggests, takes a very small tissue sample using a syringe (Figure 1).

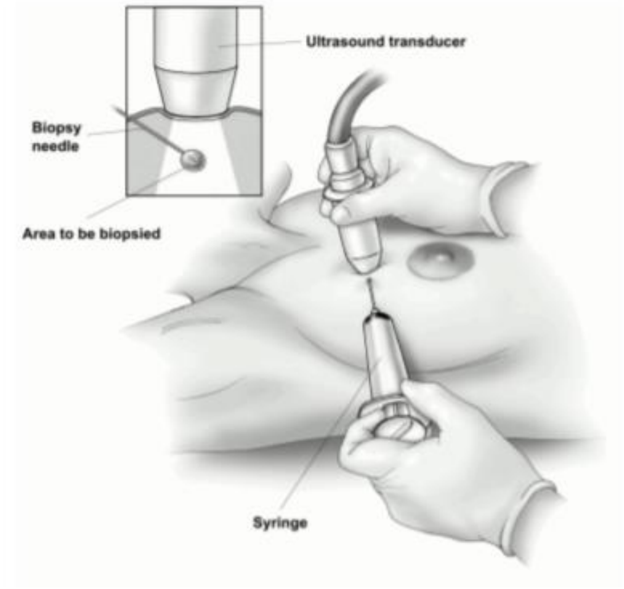

Figure 1: Fine Needle Aspiration using ultrasound.

To this end we will use the Wisconsin Diagnostic Breast Cancer dataset, containing information about 569 FNA breast samples. Each FNA produces an image as in Figure 2. Then a clinician isolates individual cells in each image, to obtain 30 characteristics (features), like size, shape, and texture. You will use these 30 features to cluster benign from malign FNA samples.

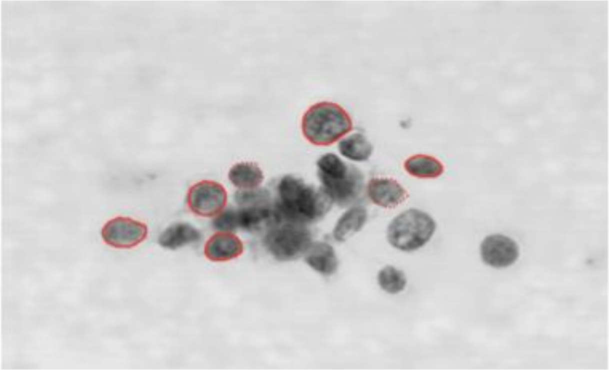

Figure 2: Breast sample obtained by FNA.

a) Load the Wisconsin Diagnostic Breast Cancer dataset (**breast_data.csv**). You should obtain a data matrix with D = 30 features and N = 569 samples. Run K-means clustering on this data.

b) The file **breast_truth.csv** contains a column vector in $\{0, 1\}^{569}$ indicating the true clustering of the dataset (0 = benign, 1 = malign). What is the accuracy of your algorithm?

c) Run your algorithm several times, starting with different centers. Do your results change depending on this? Explain.

NOT: You should upload your source code, the executable file of your source code and the answers of the questions given above as a zip file entitled with your Student ID_Name_Surname.