CS 4780/6780: Fundamentals of Data Science © Spring 2019

Homework 6: K-Means Clustering

Instructor: Daniel L. Pimentel-Alarcón Due 04/30/2019

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 6.1).

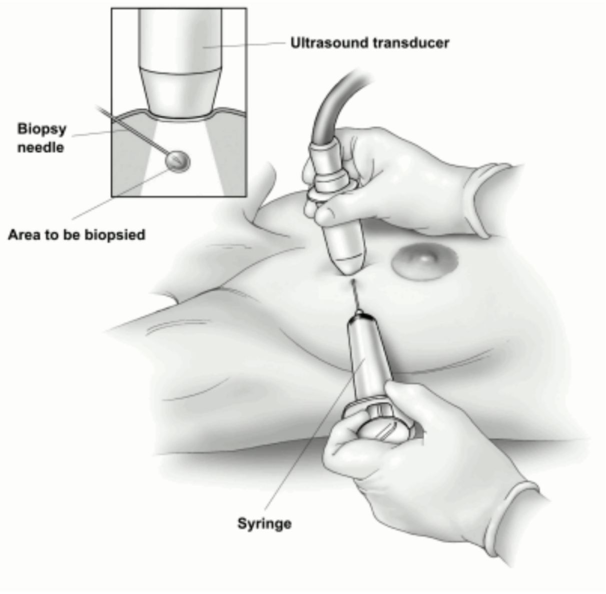

Figure 6.1: Fine Needle Aspiration using ultrasound. © Sam and Amy Collins.

To this end we will use the Wisconsin Diagnostic Breast Cancer dataset, containing information about 569 FNA breast samples [1]. Each FNA produces an image as in Figure 6.2. 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.

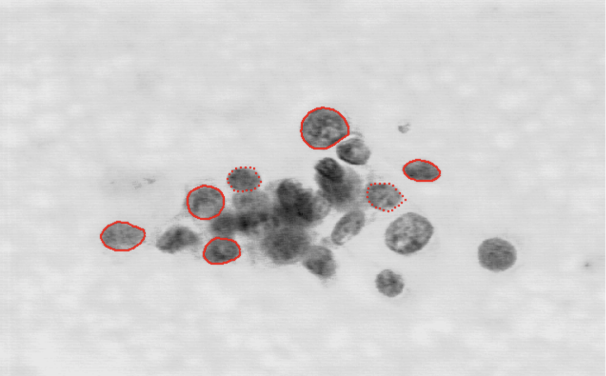

Figure 6.2: Breast sample obtained by FNA.

(a) Implement a function that performs K-means clustering. You can get started with the following code:

```
function C = kmeansclustering(X,K,mu,tol,maxIter)
% X = (D x N) data matrix; D = ambient dimension (features)
%                          N = number of samples
% K = number of clusters
% mu = (D x K) matrix containing initial centers
% tol = Tolerance parameter for convergence
% maxIter = Maximum number of iterations before giving up
% C = (1 x N) matrix indicating the clustering.

C = zeros(1,N);
change = tol + 1;
it = 0;
while change>tol && it<maxIter,

    % ===== Assign points to current centers =====

    % ===== Recalculate centers =====

end
```

(b) 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.

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

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

(e) Run your algorithm, initialized with the centers in the file `mu_init.mat`, containing a (D × K) matrix `mu_init`, where each column represents one of the initial centers. What accuracy do you obtain?

(f) What if you initialize with the *true* centers, obtained using the *true* clustering?

(g) **For extra credit.** Can you could obtain better results using another *unsupervised* learning method? What about a *supervised* one?

References

[1] O. Mangasarian, W. Street and W. Wolberg, *Breast cancer diagnosis and prognosis via linear programming*, Operations Research, 1995. Dataset available at `http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html#diag`