

Homework 7: 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 7.1).

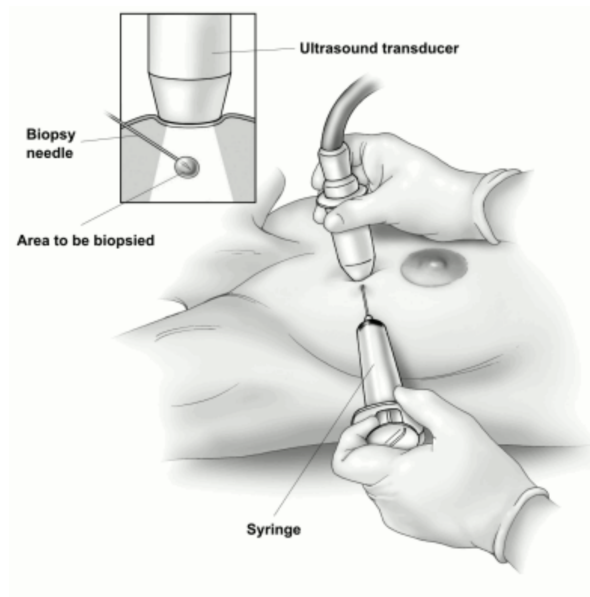


Figure 7.1: Fine Needle Aspiration using ultrasound. © Sam and Amy Collins.

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 7.2. 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.

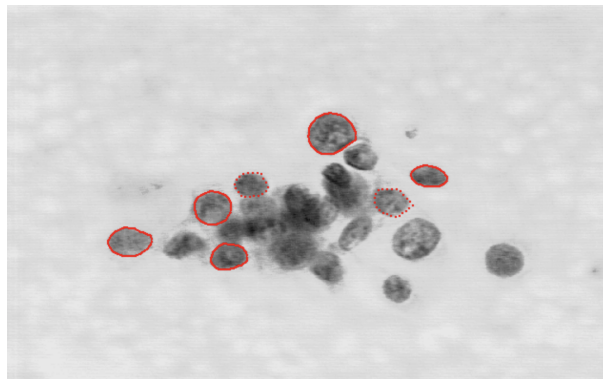


Figure 7.2: Breast sample obtained by FNA.

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

```

1 function C = kmeansclustering(X,K,mu,tol,maxIter)
2 % X = (D x N) data matrix; D = ambient dimension (features)
3 %                               N = number of samples
4 % K = number of clusters
5 % mu = (D x K) matrix containing initial centers
6 % tol = Tolerance parameter for convergence
7 % maxIter = Maximum number of iterations before giving up
8 % C = (1 x N) matrix indicating the clustering.
9
10 C = zeros(1,N);
11 change = tol + 1;
12 it = 0;
13 while change>tol && it<maxIter,
14
15     % ===== Assign points to current centers =====
16
17     % ===== Recalculate centers =====
18
19 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 $(D \times K)$ 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>