HW2: CLUSTERING WITH K-MEANS AND EXPECTATION MAXMIZATION

BY PAUL AN

PART 1 IMPLEMENTATION

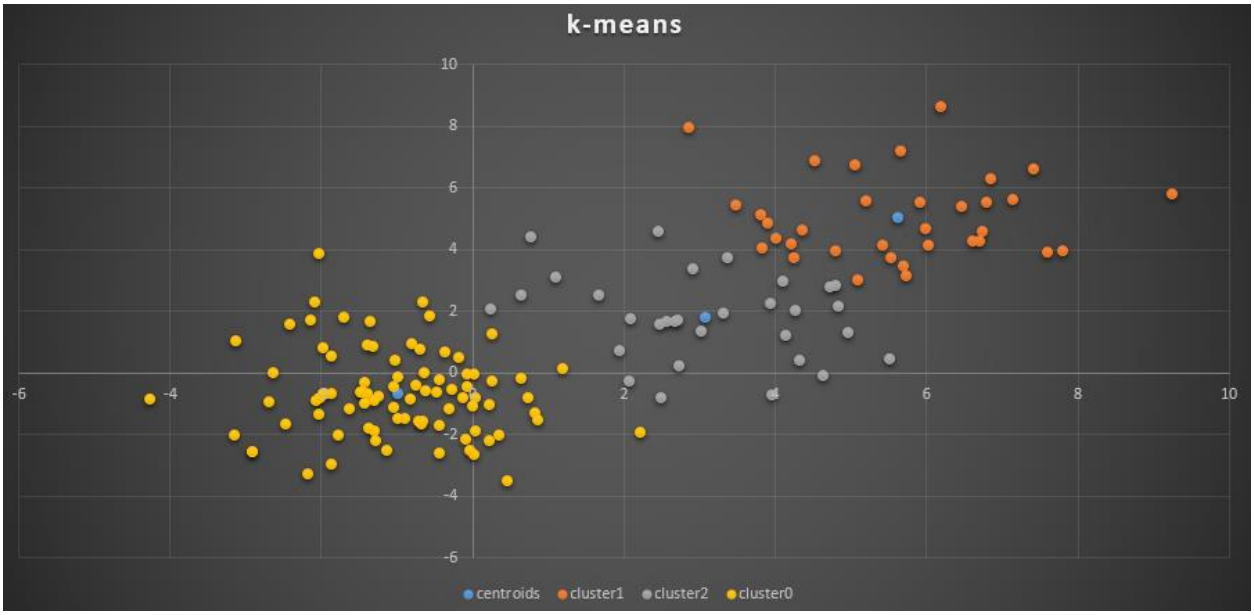
K-MEANS

An algorithm has been developed in python (see the source in kmeans.py) which constructed the centroids as well as the assignments of each point of data. Because the seed of each 3 centroids are randomized at initialization, the location is not always the exact same spot, however they are always close to each other run.

The centroids for one of the runs are:

```
[[-0.97476571808235302, -0.68419304117647095],
[5.6201657349705876, 5.0262263441764716],
[3.083182557032258, 1.7762137380322585]]
```

I have also visualized the assignments (yellow, grey, and orange) as well as the centroid locations (blue).



EXPECTATION MAXIMIZATION

The attempt can be seen in em.py.

After running the algorithm, I obtained the gausses of:

Gauss 1:

$$mean = -0.70189 \quad -1.2379$$
$$intensity = 0.13368$$
$$covar = \begin{bmatrix} 0.48425 & -0.14668 \\ -0.14668 & 0.52999 \end{bmatrix}$$

Gauss 2:

$$mean = 4.41690 \quad 3.3658$$
$$intensity = 0.4313$$
$$covar = \begin{bmatrix} 3.5769 & 2.2749 \\ 2.2749 & 5.1120 \end{bmatrix}$$

Gauss 3:

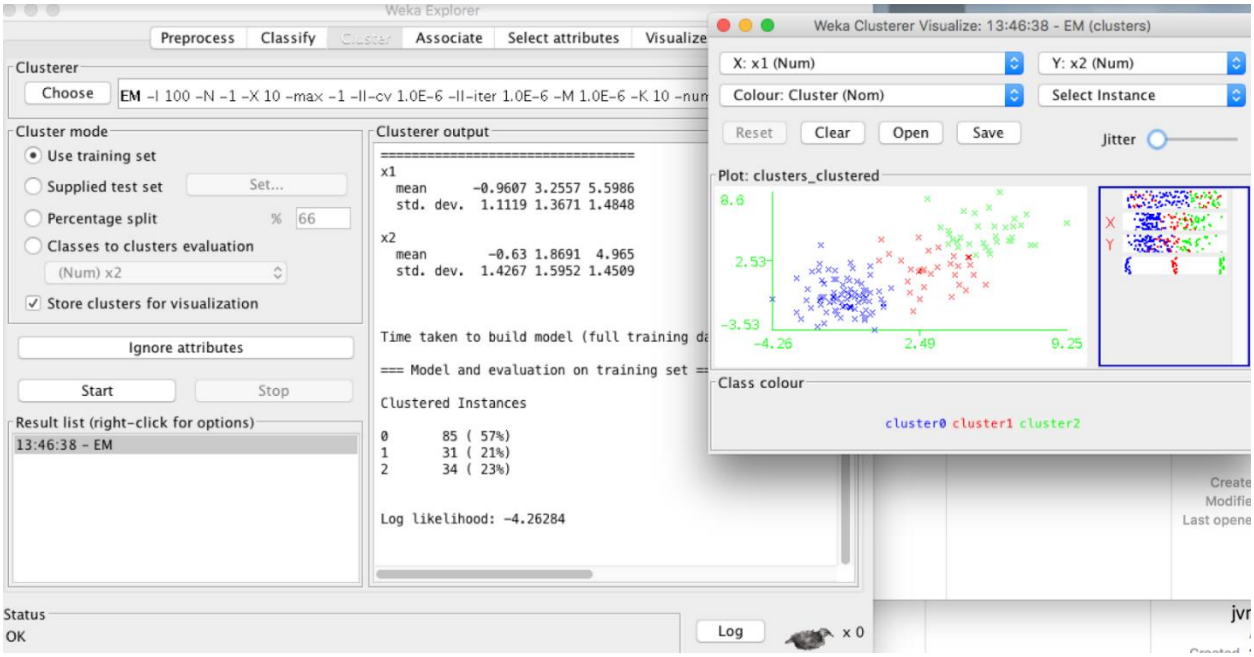
$$mean = -1.0398 \quad -0.3849$$
$$intensity = 0.4350$$
$$covar = \begin{bmatrix} 1.4598 & 0.1252 \\ 0.1252 & 2.4920 \end{bmatrix}$$

Although Excel is unable to show soft clustering allocations, I have displayed the scatterplot of each dataset with its cluster assigned to highest-weight cluster.

PART 2 SOFTWARE FAMILIARIZATION

We can definitely see a cluster at the bottom, but it's difficult to see that there are 3 regions.

Using Weka, I ran the EM (Expected-Maximization) Gaussian Mixture Model, I produced the following plot:



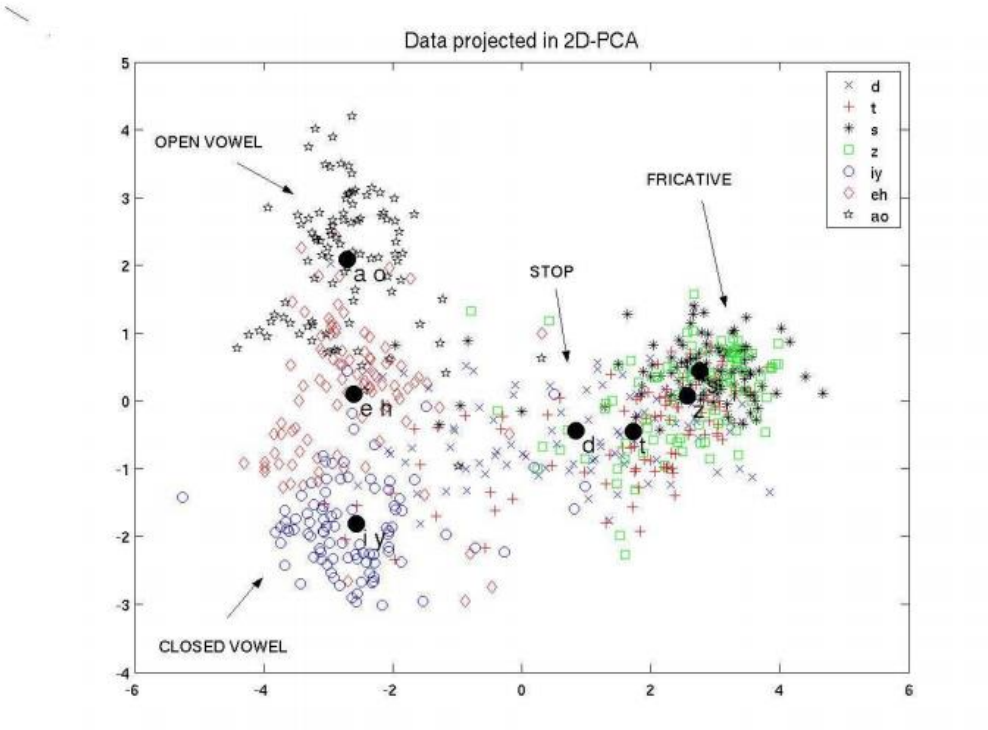
Interestingly, even without supplying the number of clusters, Weka was able to find that there were 3 clusters. Although humans are quite good at finding clusters visually if they are far apart, this would not have been a feat that would have been easy for a human. We also see that although I think iteration of my own EM did an acceptable job in identifying clusters, one of the cluster was rather small and bunched together. We did not see this in Weka's implementation. Perhaps it has to do with the seeds used for that round. As I have used starting parameters of randomized location, intensity of 1, and covariance of identity matrix.

PART 3 APPLICATIONS

Cluster analysis is useful in applications of Voice Recognition, as well as Machine vision. For example, each inflection of speech has a particular signature in the soundwaves, but when we speak, they are all combined together to form words. Using cluster analysis, we are able to identify which signature is most resonating, and string together the pronunciations to recognize what words are being spoken. See table and figure shown below.

PHONETIC GROUPS	SYMBOL	EXAMPLE WORD	POSSIBLE PHONETIC TRANSCRIPTION
STOPS	d	day	dcl d ey
	t	tea	tcl t iy
FRICATIVES	s	see	s iy
	z	zone	z ow n
	v	van	v ae n
	f	fin	f ih n
VOWELS	iy	day	dcl d ey
	eh	tea	tcl t iy
	ao	bought	bcl b ao tcl t

Table 4: {d,t,s,z,iy,eh,ao}: TIMIT lexicon of phonemic and phonetic symbols



This figure has been obtained research done by Julien Neel, ENST Paris, and can be found at http://www.speech.kth.se/prod/publications/files/1687.pdf