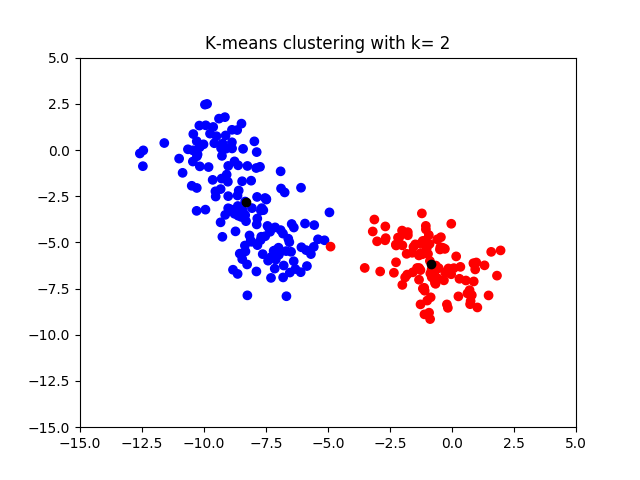
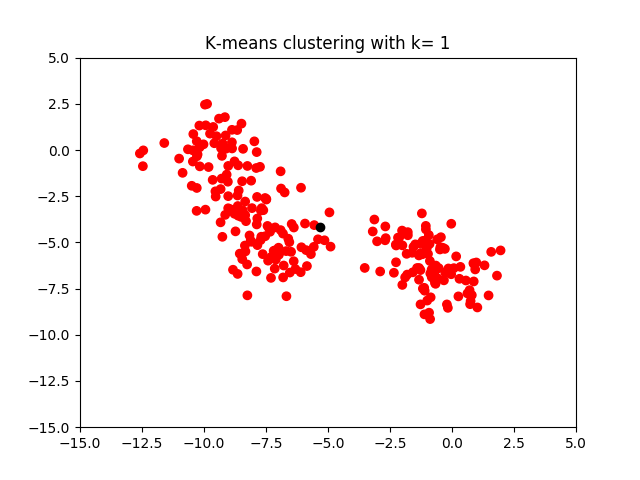
Milo Knowles

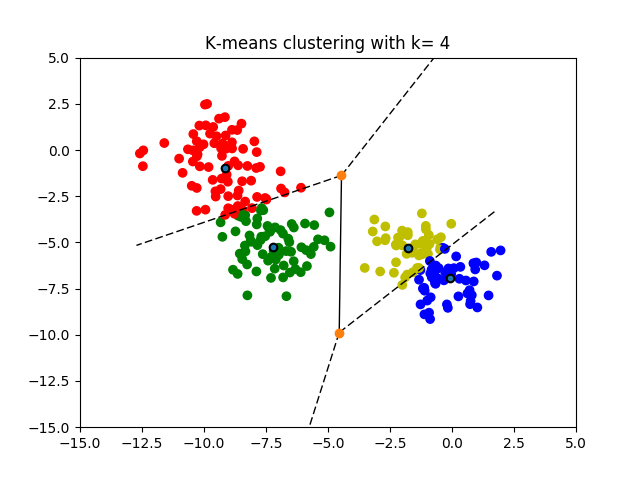
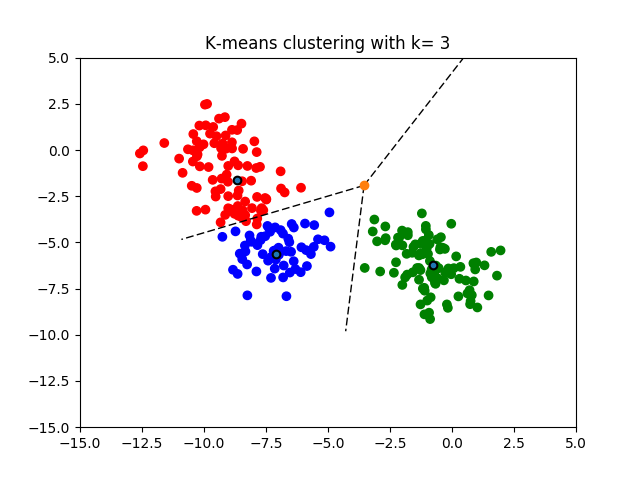
6.036 Project 3

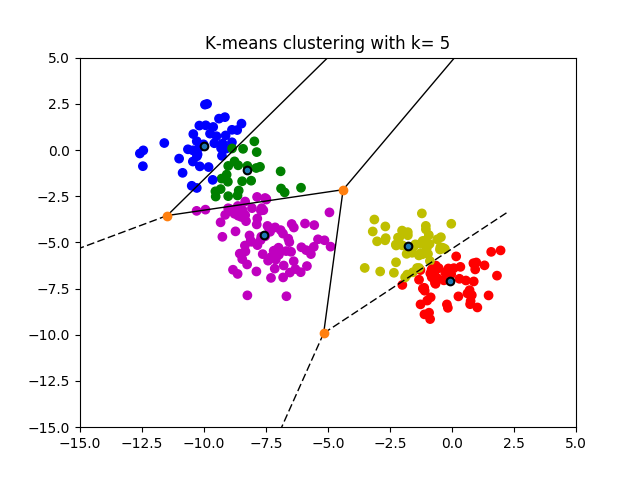
May 5, 2017

**Project 3 Writeup**

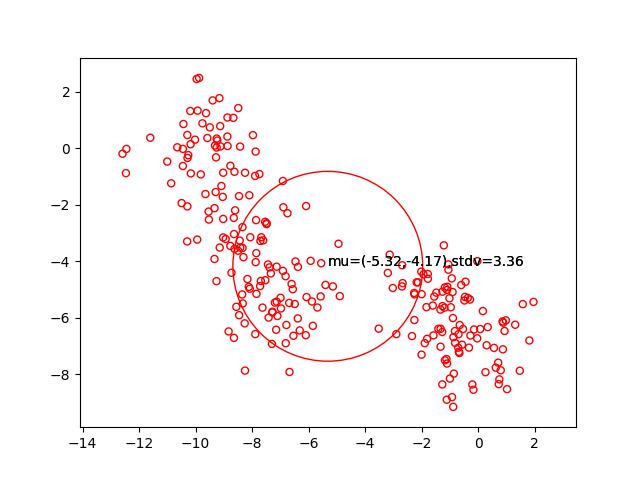
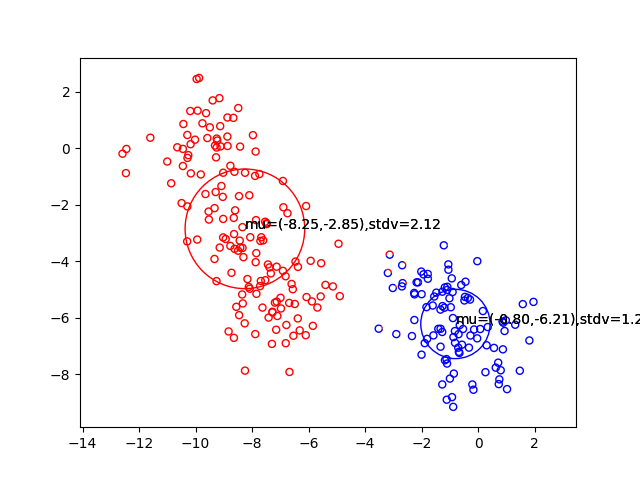
**1.1: K-Means Results (with k=1, 2, 3, 4, 5)**

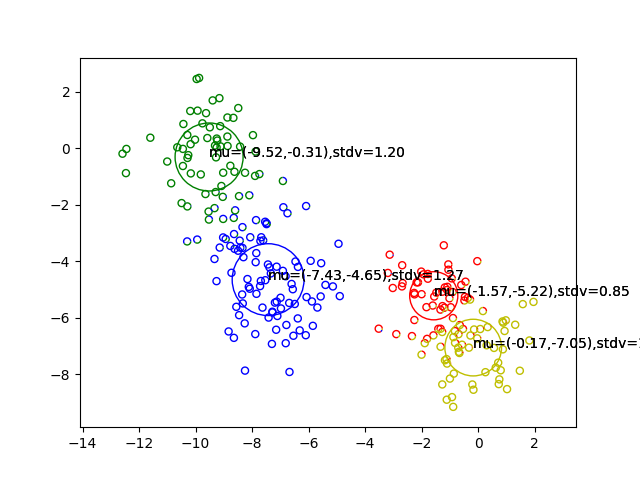
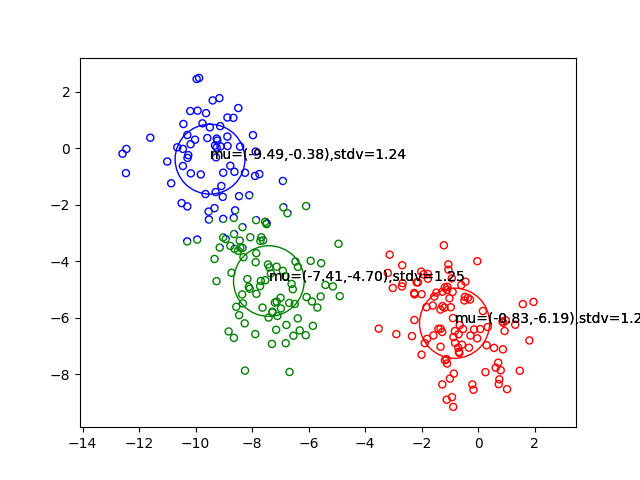


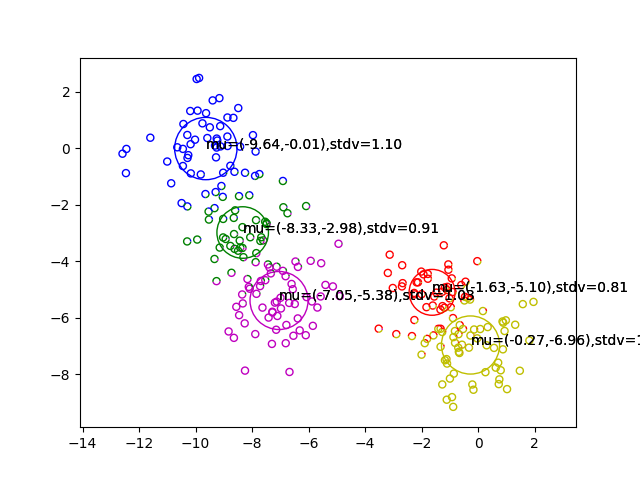




**1.4: EM Results (with k=1, 2, 3, 4, 5)**

****





**Log Likelihoods (for plots above)**

K=1: -1315.32

K=2: -1139.73

K=3: -1072.59

K=4: -1058.96

K=5: -1045.62

**1-5.** For the most part, K-Means and EM produced very similar results for each K=1, 2, 3, 4, 5. Up to an arbitrary coloring, both algorithms seem to cluster points similarly. The difference in coloring is due to the fact that K-Means and EM are non deterministic. Depending on the random initialization of the algorithms, clusters can end up with different labeling, even if they partition points in the same way. In addition, in the EM case, points between clusters have fractional coloring, because this algorithm produces “soft” cluster assignments. With the K-Means algorithm, on the other hand, points are assigned explicitly to a cluster, so colors are complete.

**Part 2: Clustering Census Data**

**2-1.** Our independence assumption will produce overconfident cluster assignments because as more redundant features are added to the data, clusters will appear farther and farther apart. If two features are highly correlated, then points within a cluster are very likely to have similar values of those features. This means that by adding more correlated features, we are simply adding another dimension in which points from the same cluster are close, but points in different clusters are relatively far apart. The Euclidean distance between clusters will be higher in the larger feature space, causing the model to be overconfident in its cluster assignments.

**2-2.** **CMM: Posterior Distribution**

**2-3a.** **CMM:** **Maximum Likelihood Estimate for .**

**2-3b.** **Maximum Likelihood Estimate for .**

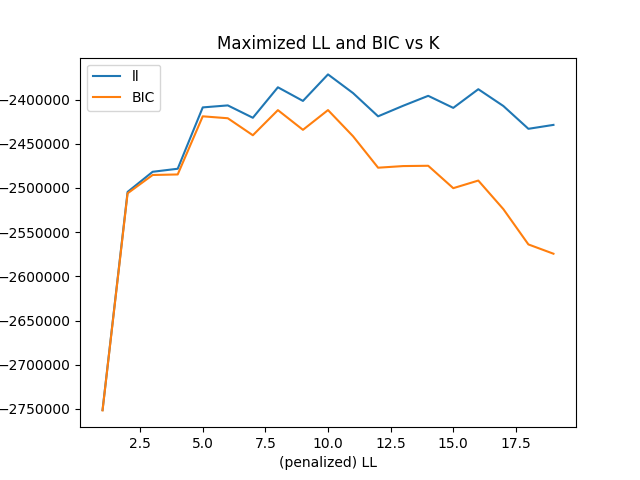
**2-5.** Over the iterations, log likelihood increases. This makes sense, because our model is finding the maximum likelihood estimates of the parameters. Because log is a monotonic function, as we increase the likelihood of our model, the log likelihood also increases. For example, with *k=2*, the model begins with a log-likelihood score of -4.656e6 but ends with a log-likelihood score of *-2.736e6* after *7* iterations. In addition, as the *k* value increases, and more clusters are used, the log likelihood improves more, since the model is better able to fit to the data.

**2-6b.**

**Best K Value (LL): 10**

**Best K Value (BIC): 8**

Based on the plot below, it appears that the best *K* values when considering log-likelihood and BIC are different. It makes sense that these *K* values do not agree, and that BIC favors a smaller *K*, because it penalizes model complexity. As *K* increases, the number of parameters required to fit for the model also increases.



**7a.** Using K=8, clusters seem to primarily be divided by age: there is a cluster representing each of the 0-12, 13-19, 20-29, 30-39, 40-49, and 65+ age groups. For the 30-39 age group, there are two clusters, one male with Western European origin and U.S citizenship and one female with Asian origin and no citizenship. For the 20-29 demographic, however, there are two female clusters: one for women of Hispanic origin who are not U.S citizens, and one for women of Western European origin who are born in the U.S.

With K=5 clusters, there were some key similarities and differences between the clusters from the K=8 case. One cluster that remained for K=5 was the age 20-29 females of Hispanic origin. All of the age groups except for the 13-19 group were represented (likely because the 13-19 group is the smallest age range). This seems to indicate that with even fewer clusters, clustering by age still occurs.

A key difference between the K=8 and K=5 case was that with 5 clusters, clusters seemed to “lose” ancestry diversity – all but one demographic were from Western Europe. With fewer clusters, there is less differentiation by ancestry. In addition, all 5 clusters (except for one with age 0-12) had high school education level. In the previous case when we had 8 clusters, 4 of the possible education levels were represented.

**7b.** After running the algorithm three times with K=8, it was clear that the model is very sensitive to initialization, and therefore unstable. For example, even though most of the same age groups are represented as clusters, genders changed arbitrarily between clusters. In all three runs of the algorithm, there was always a 65+ age cluster, a 20-29 age Hispanic origin cluster, a 13-19 female Western European origin cluster, and a 0-12 male cluster. In two of the 3 runs, a 30-39 female Asian origin cluster appeared. However, the 40-49 age group was not present in every run of the algorithm.

**7c.** If the model was trained on census data for each state, we would end up with some *K* clusters, each represented a demographic. Each state would likely have some number of it’s citizens in each of the clusters (or perhaps would not be represented in some of the clusters). By examining the number of people from each state that are represented in each of the clusters, we could put states into a *K* dimensional “cluster space”, where each dimension of the space represents a cluster. States with a high representation in some cluster would have a large value in the dimension corresponding to that cluster. Once we represent each state as a vector in the *K* dimensional cluster space, we can compare the “closeness” of two states based on their cosine similarity. This is analogous to the way that we compared documents using cosine similarity. If two states are close in the cluster space, this means that they are represented similarly among demographics; this means that the states have similar populations.