Vincent La
903178639-vla6-hw3

K Means Clustering

Table 1: Clustering with 3 Centers using All Features

Percentage Cluster	Case	Control	Unknown
Cluster 1	26.7%	25.7%	26.1%
Cluster 2	25.3%	29.3%	24.1%
Cluster 3	47.9%	45.0%	49.7%
	100%	100%	100%

Table 2: Clustering with 3 Centers using Filtered Features

Percentage Cluster	Case	Control	Unknown
Cluster 1	34.9%	32.4%	41.4%
Cluster 2	32.3%	33.3%	25.7%
Cluster 3	32.8%	34.3%	32.9%
	100%	100%	100%

GMM Clustering

Table 3: Clustering with 3 Centers using All Features

Percentage Cluster	Case	Control	Unknown
Cluster 1	27.2%	25.4%	26.2%
Cluster 2	23.9%	27.7%	26.5%
Cluster 3	48.9%	46.9%	47.2%
	100%	100%	100%

Table 4: Clustering with 3 Centers using Filtered Features

Percentage Cluster	Case	Control	Unknown
Cluster 1	33.5%	31.7%	27.8%
Cluster 2	33.2%	30.2%	28.5%
Cluster 3	33.3%	38.0%	33.7%
	100%	100%	100%

Discussion on K-means and GMM

In both K-means and GMM I noticed that when using all features, the percentage makeup of Case, Control, and Unknown is relatively imbalanced. However, when using filtered features, notice that for Case and Control, the Clustering Algorithm derives clusters of almost equal size.

Table 5: Purity Value for Different Number of Clusters

K	K-Means All Features	K-Means Filtered Features	GMM All Features	GMM Filtered Features
2	0.78552	0.56606	0.71773	0.87202
5	0.48454	0.40945	0.54962	0.55226
10	0.37473	0.38565	0.23048	0.50431
15	0.20174	0.32459	0.29284	0.36081

The first pattern that I see is somewhat obvious what as you increase K purity decreases. This makes sense as it is almost definitional that when you decrease K you will increase purity, and thus similarly when you increase K purity should be expected to decrease.

Some patterns other that I see from this is that in general GMM has higher purity than K-Means, and Filtered Features generally has higher purity than using all features.