2.3 K-Means Clustering

b.

Percentage Cluster	Case	Control	Unknown
Cluster 1	9.93852%	7.38397%	11.96145%
Cluster 2	90.06148%	92.08861%	84.69388%
Cluster 3	0%	0.52743%	3.34467%
	100%	**100%**	**100%**

Table 1: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	99.28279%	0%	29.64103%
Cluster 2	0.71721%	97.78481%	70.25641%
Cluster 3	0%	2.21519%	0.10256%
	100%	**100%**	**100%**

Table 2: Clustering with 3 centers using filtered features

2.4 GMM Clustering

b.

Percentage Cluster	Case	Control	Unknown
Cluster 1	4.91803%	6.11814%	24.4898%
Cluster 2	0%	0%	2.72109%
Cluster 3	95.08197%	93.88186%	72.78912%
	100%	**100%**	**100%**

Table 3: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	77.35656%	0%	26.97436%
Cluster 2	3.27869%	97.78481%	42.87179%
Cluster 3	19.36475%	2.21519%	30.15385%
	100%	**100%**	**100%**

Table 4: Clustering with 3 centers using filtered features

2.5 Streaming K-Means Clustering

c.

Percentage Cluster	Case	Control	Unknown
Cluster 1	26.84426%	13.81857%	42.97052%
Cluster 2	5.53279%	69.51477%	17.80045%
Cluster 3	67.62295%	16.66667%	39.22902%
	100%	**100%**	**100%**

Table 5: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	76.33197%	0%	51.89744%
Cluster 2	7.17213%	100%	1.12821%
Cluster 3	16.4959%	0%	46.97436%
	100%	**100%**	**100%**

Table 6: Clustering with 3 centers using filtered features

2.6 Discussion on k-means and GMM

a. Based on the result from clustering, filtered features recorded more accurate in case and control than all features. K-Means of filtered features recorded 99% for case and 98% for control when all features recorded 10% for case and 7% for control. However, unknown cluster of filtered features recorded not even close to 1% which is worse than 3% of all features result. Same observation is made on GMM. Filtered features of GMM scored 77% for case and 98% for control when all features scored only 5% for case and 6% for control. Similar to K-Means, unknown cluster of filtered features for GMM didnt do better than all features. It scored 30% when all features scored 73%

b.

k	K-Means All features	K-Means Filtered features	GMM All Features	GMM Filtered features
2	0.47831	0.66126	0.47831	0.43532
5	0.61958	0.40531	0.51139	0.84201
10	0.60982	0.41187	0.63259	0.87892
15	0.70174	0.89169	0.68547	0.88444

Table 7: Purity values for different number of clusters

Observed pattern is that purity values increased when K values went up except K=2 for K-Means filtered features. As seen above, purity value recorded best for both types of K-Means and GMM when K value was 15 which is highest.

3. Advanced phenotyping with NMF

b.

k	NMF All features	NMF Filtered features
2	0.47831	0.66161
3	0.47831	0.60849
4	0.47939	0.72335
5	0.47939	0.68506

Table 8: Purity values for different number of clusters

c.

Percentage Cluster	Case	Control	Unknown
Cluster 1	21.20902%	22.67932%	32.82313%
Cluster 2	78.79098%	77.32068%	67.17687%
Cluster 3	0%	0%	0%
	100%	**100%**	**100%**

Table 9: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	49.07787%	1.05485%	11.69231%
Cluster 2	3.9959%	0%	35.58974%
Cluster 3	46.92623%	98.94515%	52.71795%
	100%	**100%**	**100%**

Table 10: Clustering with 3 centers using filtered features

d.

$$1.\ W_{ij} \leftarrow W_{ij} + \eta_{ij}(XH^T - WHH^T)_{ij}$$

$$2.\ W_{ij} \leftarrow W_{ij}\frac{(XH^T)_{ij}}{(WHH^T)_{ij}}$$

$$3.\ H_{ij} \leftarrow H_{ij} + \mu_{ij}(W^TX - W^TWH)_{ij}$$

$$4.\ H_{ij} \leftarrow H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}}$$

Above equation is from http://www.almoststochastic.com/2013/06/nonnegative-matrix-factorization.html