CSE8803: Big Data Analytics in Healthcare
Homework 3

Jeff McGehee

Deadline: Feb 26, 2017, 11:55 PM AoE

1 Programming: Rule based phenotyping

a. See submission
b. See submission

2 Programming: Unsupervised phenotyping via clustering

2.1 Feature Construction

a. See submission

2.2 Evaluation Metric

a. See submission

2.3 K-Means Clustering

a. See submission
b. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2.

2.4 Clustering with Gaussian Mixture Model (GMM) [5 points]

a. See submission
b. Compare clustering for the $k = 3$ case with the ground truth phenotypes that you

Percentage Cluster	Case	Control	Unknown
Cluster 1	38%	3%	39%
Cluster 2	2%	3%	2%
Cluster 3	60%	94%	59%
	100%	**100%**	**100%**

Table 1: KMeans Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	34%	36%	38%
Cluster 2	25%	23%	20%
Cluster 3	41%	41%	42%
	100%	**100%**	**100%**

Table 2: KMeans Clustering with 3 centers using filtered features

computed for the rule-based PheKB algorithms. Specifically, for each of *case*, *control* and *unknown*, report the percentage distribution in the three clusters for the two feature construction strategies. Report the numbers in the format shown in Table 1 and Table 2.

Percentage Cluster	Case	Control	Unknown
Cluster 1	36%	34%	36%
Cluster 2	9%	11%	10%
Cluster 3	55%	55%	54%
	100%	**100%**	**100%**

Table 3: GMM Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	55%	55%	57%
Cluster 2	39%	38%	36%
Cluster 3	6%	7%	7%
	100%	**100%**	**100%**

Table 4: GMM Clustering with 3 centers using filtered features

2.5 Discussion on k-means and GMM [6 points]

a. Briefly discuss what you observe in 2.3b and 2.4b.

In all cases, neither Kmeans nor GMM did a great job of isolating the classes into their own clusters. In the filtered case, KMeans had relatively evenly distributed clusters, while GMM had two large clusters and one small cluster for both the filtered and unfiltered data.

b. Re-run k-means and GMM from the previous two sections for different k (you may run it each time with different k). Report purity for filtered and all features for each k by filling up Table 6. Discuss patterns observed, if any.

Purity is at its best for low K, and oddly is very high for Kmeans All Features. It looks like purity may decreas exponentially in K.

k	K-Means All features	K-Means Filtered features	GMM All Features	GMM Filtered features
2	0.96	0.57	0.63	0.73
5	0.37	0.40	0.44	0.45
10	0.32	0.46	0.32	0.37
15	0.23	0.34	0.21	0.40

Table 5: Purity values for different number of clusters

3 Advanced phenotyping with NMF

a. See submission.
b. Run NMF clustering for $k = 2, 3, 4, 5$ and report the purity for two kinds of feature construction. [5 points]

k	NMF All features	NMF Filtered features
2	0.59	0.74
3	0.57	0.76
4	0.52	0.74
5	0.61	0.76

Table 6: Purity values for different number of clusters

c. Perform the comparison of clustering for the $k = 3$ case with the ground truth phenotypes that you computed for the rule-based PheKB algorithms. Specifically, for each cluster, report the percetage of *case*, *control* and *unknown* in Table 7 and Table 8 for two feature construction strategies. [5 points]

Percentage Cluster	Case	Control	Unknown
Cluster 1	0%	0%	2%
Cluster 2	20%	68%	14%
Cluster 3	80%	32%	84%
	100%	**100%**	**100%**

Table 7: NMF with 3 centers characteristics using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	100%	3%	38%
Cluster 2	0%	0%	10%
Cluster 3	0%	97%	52%
	100%	**100%**	**100%**

Table 8: NMF with 3 centers characteristics using filtered features