CS 6250 Homework 3

Tian Tan

Feb.24, 2017

2.*FeatureConstruction*

2.3 K-Means Clustering

Percentage Cluster	Case	Control	Unknown
Cluster1	9.1016 %	17.7215%	22.3356 %
Cluster2	90.9836%	81.7511 %	74.3197 %
Cluster3	0.0000 %	0.5274 %	3.3447 %
	100%	100%	100%

Table 1: K-Means Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster1	99.4877 %	100.0000 %	32.4102 %
Cluster2	0.0000 %	0.0000 %	67.4872 %
Cluster3	0.5123 %	0.0000 %	0.1026 %
	100%	100%	100%

Table 2: K-Means Clustering with 3 centers using filtered features

2.4 Clustering with Gaussian Mixture Model

Percentage Cluster	Case	Control	Unknown
Cluster1	6.4549 %	14.6624 %	20.0113 %
Cluster2	0.1025 %	1.4768 %	2.9478 %
Cluster3	93.4426 %	83.8608 %	77.0408 %
	100%	100%	100%

Table 3: GMM Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster1	18.9549 %	0.0000 %	13.3333 %
Cluster2	0.0000 %	100.0000 %	0.0000 %
Cluster3	81.0451 %	0.0000 %	86.6667%
	100%	100%	100%

Table 4: GMM Clustering with 3 centers using filtered features

2.5 Discussion on k-means and GMM

a. Briefly discussion:

From the results of 2.3b and 2.4b, we can find that for Kmeans and GMM methods, filtering features lead to a higher purity. When looking carefully at the clustering results, I find that kmeans and GMM methods does show more accuracy using filtered features compared to all features. For instance, kmeans put case, control and unknown patients into cluster 2 when using all features, but put case and control into cluster 1 and unknown patients into cluster 2 when using filtered features. GMM behaves similarly, where it predicts case,control and unknown patients into cluster 3 with all features, and predict case and unknown into cluster 3 and control into cluster 2 with filtered features. Since filtered features model has higher purity than all features model, we can tell that model does show more accuracy in classification but results in a different classification way.

b.

k	K-Means All features	K-Means Filtered features	GMM All Features	GMM Filtered feat
2	0.47831	0.56559	0.479831	0.65213
5	0.54745	0.87020	0.53118	0.85185
10	0.62310	0.88716	0.61768	0.84666
15	0.67408	0.88577	0.67110	0.87816

Table 5: Purity values for different number of clusters

Discussion: increasing k values can increase purity for both kmeans and GMM methods for all features and filtered features; when k is small(2 and 5), GMM has higher purity than kmeans; when k becomes larger(10 and 15), K means performs better than GMM with respect to purity.

3.*AdvancedphenotypingwithNMF*

b.

c.

k	NMF All features	NMF Filtered features
2	0.47831	0.57113
3	0.47831	0.54448
4	0.47858	0.5529
5	0.49159	0.56836

Table 6: NMF Purity values for different number of clusters

Percentage Cluster	Case	Control	Unknown
Cluster1	4.8156 %	9.3882 %	6.9728 %
Cluster2	89.8565 %	85.1266 %	88.5487 %
Cluster3	15.3279%	5.4852 %	4.4785 %
	100%	100%	100%

Table 7: NMF Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster1	8.1967 %	58.6354 %	20.4102 %
Cluster2	76.8443 %	12.2601 %	52.5128 %
Cluster3	14.9590 %	29.1045 %	27.0769 %
	100%	100%	100%

Table 8: NMF Clustering with 3 centers using filtered features

d. MU rule uses gradient descent algorithm

$$
\begin{aligned}
f(W^t, H^t) &= \frac{1}{2}||\boldsymbol{V} - \boldsymbol{WH}||_2^2 && (1)\\
\nabla_W f(W, H) &= (WH - V)H^T && (2)\\
\nabla_H f(W, H) &= W^T(WH - V) && (3)\\
H_{ij}^{t+1} &= H_{ij}^t - \frac{H_{ij}^t}{W^TWH}\nabla_H f(W, H) && (4)\\
&= \frac{\left(H^tW^TWH^t - H^t(W^TWH^t - W^TV)\right)_{ij}}{(W^TWH^t)_{ij}} && (5)\\
&= H_{ij}^t \frac{(W^\top V)_{ij}}{(W^\top WH^t)_{ij}} && (6)\\
W_{ij}^{t+1} &= W_{ij}^t - \frac{W_{ij}^t}{W^TWH}\nabla_W f(W, H) && (7)\\
&= \frac{\left(W^t(W^tHH^\top) - W^t(W^tH - V)H^T\right)_{ij}}{4(W^tHH^\top)_{ij}} && (8)\\
&= W_{ij}^t \frac{(VH^\top)_{ij}}{(W^tHH^\top)_{ij}} && (9)\\
& && (10)
\end{aligned}
$$