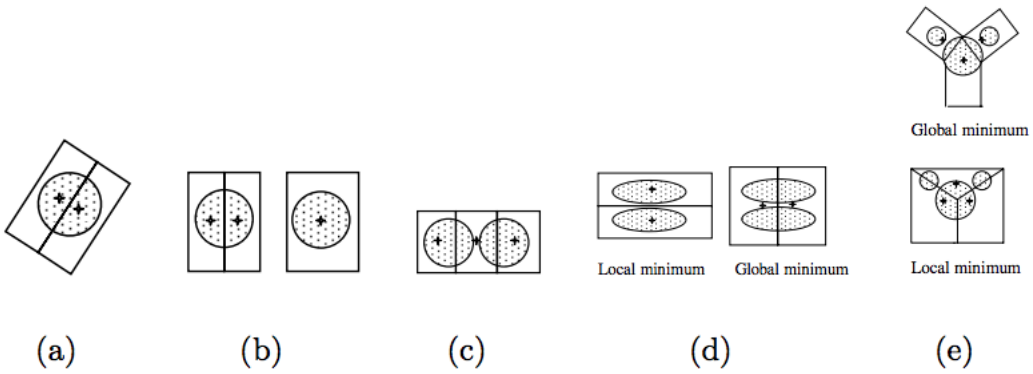
HW3 Solution DM 2018

1.



2.

- One advantage of DBSCAN is that it's relatively resistant to noise and can handle clusters of arbitrary shapes and sizes, since it uses a density-based de_nition of a cluster, while K-means is sensitive to noise and always generate globular shape cluster.

- One disadvantage of DBSCAN is that it has more parameters than K-means, it's always hard to decide the special value for each parameter for a best output.

3. fuzzy c-means has much the same weakness as K-means, it still has the limitations on sensitivity to outliers, difficulty in handling clusters of different sizes and densities and with non-globular shapes.

4.

a. There are ABCDE, 5 items total, the number of association rules can be calculated as below: 3^5-2^(5+1)-1=180
b. There are totally 2^5-1 non-empty subsets of {A,B,C,D,E}, but set {ABCD},{ABCE},{ABCDE} are not frequent itemset (assuming minsup > 0). So the maximum size of frequent itemsets is 4
c. All the size-3 subsets of {A,B,C,D,E} can be derived, so the number of size-3 itemsets that can be derived from this data set is $\binom{5}{3}=10$

d. Set {B,D} or {D,E} have largest support 6.

e. Item A, C

5.

a. Dataset 2
b. 5 closed frequent itemsets
c. Dataset 2
d. Dataset 2
e. Dataset 2