3.1

a) Suppose M is the minium support threshold and I denote the item set

Since I is not frequent in DB, $Sup(I, DB) = \frac{|\{T \in DB \mid I \subseteq T\}|}{|DB|} < M$

Similarly, Since I is not frequent in ΔDB, $Sup(I, \Delta DB) = \frac{|\{T \in \Delta DB \mid I \subseteq T\}|}{|\Delta DB|} < M$

Therefore, $|\{T \in DB \cup \Delta DB \mid I \subseteq T\}| < M|DB| + M|\Delta DB|$

Since DB and ΔDB are disjoint, $|DB \cup \Delta DB| = |DB| + |\Delta DB|$

$M|DB| + M|\Delta DB| = M(|DB| + \Delta|DB|)$

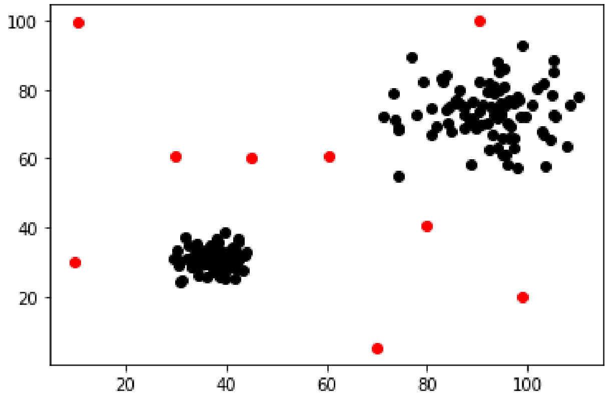
Therefore, $$\frac{|\{T \in DB \cup \Delta DB \mid I \subseteq T\}|}{|DB \cup \Delta DB|} < \frac{M(|DB| + \Delta|DB|)}{|DB \cup \Delta DB|} = M$$

b)

A: All item sets that are frequent in one of DB or ΔDB but not frequent in another should have their support counted. Apriori algorithm for each case only returns the frequent item set in the current data set but not returning the frequent item in DB ∪ ΔDB, so all item sets that are frequent exclusive to one of the data set should be counted.

3.2 A: When determining the value of K, too small would cause the algorithm to be not robust enough. At the same K can't be too large otherwise the result won't be local enough. Therefore, a K value between 10-20 is optimal, and for this example, K is chosen to be 10.

Since LOF determines the ratio of outlier, the threshold for outliers depends on the problem. For this example, I assume 5% of the data are outliers, so the threshold is 2. Below is the screen shot of LOF plot for the data set.


100
80
60
40
20
20
40
60
80
100