Machine Learning worksheet 2 Solution

1. a) 2 Only
2. d) 1, 2 and 4
3. a) True
4. a) 1 only
5. b) 1
6. b) No
7. a) Yes
8. d) All of the above
9. a) K-means clustering algorithm
10. d) All of the above
11. d) All of the above
12. Yes, The k-means algorithm is sensitive to the outliers.

The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster centre closer to the outlier.

An example, is the average of the salaries of the following people:

50k, 20k, 35k, 65k and 1 Million

The average ends up being (50k + 20k + 35k + 65k + 1MM) / 5 = 1170k / 5 = 234k.

If we did not have the 1MM outlier, the average would have been (50k + 20k + 35k + 65k) / 4 = 170k / 4 = 42.5k.

Note that the two average results are wildly different from one another.

Given that k-means clustering is an unsupervised algorithm, it is up to the interpreter to determine whether this makes sense or not for a given data set. There are other clustering algorithms out there that are less sensitive to outliers. Depending on your application it may be worth using a different approach than the k-means algorithm.

13. K means better than other clustering Algorithms because have multiple advantages and is less expansive in comparison to other clustering algorithms. Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. K means have the following advantages

    ->Relatively simple to implement.
    ->Scales to large data sets.
    ->Guarantees convergence.
    ->Can warm-start the positions of centroids.
    ->Easily adapts to new examples.
    ->Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Machine Learning worksheet 2 Solution

14. No, K-Means is a non-deterministic algorithm

    The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.