K-Means Quiz

1. Facts
    A. Clustering methods show us how many subpopulations there are in the data
    B. Clustering methods do not require labelled data to build a model
2. Clustering algorithms can be characterised by various properties:
    A. Polythetic vs Monothetic: cluster instances share one or a very small number of common attribute values versus cluster instances are 'similar' according to some measure across many or all attribute values
    B. Flat vs Hierarchical: each cluster is a group of instances with no relationship between clusters versus some clusters group other clusters rather than instances directly
    C. Hard vs Soft: an instance can only belong to one cluster versus an instance can belong to more than one cluster
3. The value of K in K-Means refers to the number of clusters the population is divided into.
4. In the K-means algorithm, a cluster is characterised by a 'centroid', which has, for each attribute, the average of that attribute's values over all the instances in the cluster.
5. The K-means algorithm
    A. Start with a centroid for each cluster, e.g. a randomly chosen instance
    B. Iteratively assign each instance to the cluster whose centroid is nearest
    C. Compute the centroid for each cluster from its instances
    D. Terminate when no centroid changes
6. The K-means algorithm minimises the aggregate intra-cluster distance.
7. When the k-means algorithm performs minimisation, it will reach a local minimum or any minimum.
8. When the K-means algorithm terminates, it guarantees that
    A. Every instance is in the cluster with the closest centroid
    B. The aggregate intra-cluster variance is minimised
9. The following are all reasonable ways to select k for k-means, in practice:
    A. Class labels if we have them for the population as a whole but not for each instance
    B. The knee of a scree plot
    C. The maximum of the 2nd derivative of a scree plot
10. Suppose you are allowed to create a clustering algorithm producing K clusters, and suppose you are given R reference classes, and you know which instance is in which class. Suppose also you are allowed to pick K, and you are allowed to pick the method to assign clusters to classes. If you are allowed the type of clustering algorithm and/or assignment of cluster to classes, how would you choose K and how would you choose assignment in order to achieve 100% accuracy, i.e. each instance is in a cluster which is assigned to the class the instance is in?
    A. You can assign multiple clusters to the same class
        a. You choose K to be the number of instances, you cluster each instance to its own cluster, and you assign the cluster to the class the instance is in.
    B. You are allowed to assign a cluster to multiple classes
        a. You choose K to be 1 so all the instances are in one cluster. This cluster is assigned

to all the classes.

C. You are allowed to have clusters that overlap, but you cannot assign multiple clusters to one class or a single cluster to multiple classes
    a. You produce every possible cluster: all the possible subsets of the instances. For each reference class, you find the cluster that contains exactly the instances in that class and assign it to the class. The other clusters are not assigned.

11. Extrinsic evaluation of a clustering solution is measuring the improvement you get on another task if you add the cluster number or label to the input data for that task.