**10.4 For the *k*-means algorithm, it is interesting to note that by choosing the initial cluster**

**centers carefully, we may be able to not only speed up the algorithm’s convergence, but**

**also guarantee the quality of the final clustering. The *k*-meansCC algorithm is a variant**

**of *k*-means, which chooses the initial centers as follows. First, it selects one center**

**uniformly at random from the objects in the data set. Iteratively, for each object *p* other**

**than the chosen center, it chooses an object as the new center. This object is chosen at**

**random with probability proportional to *dist*.*p*/2, where *dist*.*p*/ is the distance from *p***

**to the closest center that has already been chosen. The iteration continues until *k* centers**

**are selected.**

**Explain why this method will not only speed up the convergence of the *k*-means**

**algorithm, but also guarantee the quality of the final clustering results.**

**10.12 Present conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Give application examples to support your argument.**

Density-based clustering can identify outliers and abnormal (not spherical) cluster shapes.

Density-based clustering also identifies mutually exclusive clusters, as opposed to hierarchical clustering which identifies nested clusters. If your data necessitates outlier identification, abnormal cluster identification, and exclusive clusters, density-based clustering should be used over hierarchical or partition methods. If you are also unsure of the number of clusters in your data, density-based clustering may be a good choice. Unlike partition and hierarchical methods, density-based methods do not require the number of clusters as a parameter. Finally, density-based methods identify clusters based on density rather than similarity/distance. Density-based methods should be used when your goal is to identify where many points are clustered (high density regions) not just identify groupings of points that are similar to each other.

Density-based clustering methods are better suited than partitioning or hierarchical clustering methods when working with spatial data, data that describes the location of objects in physical space. Considering the current pandemic, density-based clustering may be useful for identifying the highly infected areas. There’s no reason to believe the virus would spread in a spherical shape, so we want to be able to identify abnormal shapes. Also, identifying outliers would be useful for knowing where potential new dense regions of infection may start. Clustering may also be useful for identifying areas of a city struggling economically, by identifying areas with a high density of foreclosures or empty commercial buildings, for example. This could help direct city resources to economically stressed areas.

**10.16 Describe each of the following clustering algorithms in terms of the following criteria:**

**(1) shapes of clusters that can be determined; (2) input parameters that must be**

**specified; and (3) limitations.**

**(a) *k*-means**

**(b) *k*-medoids**

**(c) CLARA**

**(d) BIRCH**

**(e) CHAMELEON**

**(f) DBSCAN**