K Means Clustering

K means clustering is an unsupervised clustering method

Algorithm

Repeat this procedure till centroids are no longer moving
1. Initiaze k centroids randomly,
2. For each data points identify which centroid is closest to that point,
3. group the datapoints based on closest centroid,
4. shift the centroid to the mean of those datapoints.

Drawbacks

1. K means clustering is good at finding clusters that are spherical in shape, but they perform really bad on basically almost everything else, take moon shape for an example.
2. It doesn't work a lot of times, sometimes you would just have a centroid sitting somewhere between two or more clusters, some centroid with no cluster to call home eh!. Fortunately, this issue is solved in k++ means.

K++ means

The only difference in this algorithm is how you initialize the centroids.

1. Select a datapoint randomly, call it the 1st centroid
2. for the following centroids, choose a datapoint with probability proportional to the sum distance from each centroid that has been assigned so far

Questions

1. What would happen if you multiply the distances from each centroid instead of adding in case of k++ means
2. How do you select an optimal value for k?
3. Should you normalize the attributes?
4. Can you use other distances instead of euclidian?
   [fillers]
5. why does k means not work for moon shape clusters?