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?

Answers

1. Would perform poorly on normalized data
2. Use elbow method, that says choose the value of k which cause the average distance between a points in the cluster to the centroid to decrease the most.
3. Yes, we don't often know which feature is more important, if we don't normalize, some feature would have effect on centroid more than the other.

4. Yes, say you use manhattan distance, now this clustering would be called k-mediod clustering which is also a well known clustering method. The original k-means algorithm is only for euclidian distance.
5. the distance between a point in different cluster could be and often is smaller from centroid of another cluster,