**K-Means**

When you are running a k-means algorithm, there is a centroid that has no point or with zero points assigned to it. In that case, the common thing to do is eliminate that cluster centroid. If you do that you will end up with k-1 clusters instead of k clusters.

Some times you really need K clusters, and at that case you should reinitialize that cluster centroids.

ci: index of cluster (1,2,…,K) to which example xi is currently assigned

uk: cluster centroid k(uk є Rn)

uck: cluster centroid of cluster to which example xi has been assigned.

Optimization objective:

J(c1, c2, …, cm, u1, u2, …, uK) = 1/m \* sum(||xi – uci||2)(i from 1 up to m);

this cost function could help us to make k-means runs well and may help us to avoid the bad local optima.

min J(c1, c2, …, cm, u1, u2, …, uK)

K-means Algorithm

Randomly initialize K cluster centroids u1, u2, u3, …, uK є Rn

Repeat{

for i = 1:m

ci := index of cluster (1,2,…,K) to which example xi is currently assigned

(cluster assignment step, minimize J(…) wrt c1, c2, …, cm, holding u1, u2, … uk, fixed)

for I = 1:K

uk := average (mean) of points assigned to cluster k.

(move centroid, minimize J(…) wrt u1, u2, …, uk)

}

**How to avoid local optima**

Should have K < m. Randomly pick K training examples. Set u1, u2, …, uK equal to these K examples.

**Avoid local optima**

If you want to avoid the local optima, you should initialize K-means lots of times and run K-means lots of times to try to make sure we get as good a solution, as a good local or global optima as possible.

**Randomly initialization**

Concretely:

For i = 1:iterationNumber{

Randomly initialize K-means.

Run K-means. Get c1, c2, …, cm, u1, u2, …, uK.

Compute cost Function J(c1,c2, …, cm, u1, u2, …, uK)

}

Finally, pick clustering that gave lowest cost J(…).

Maybe k also equals to 2-10.

Note: Typically iterationNumber is 50-1000.

Choosing the value of K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

Andrew Ng: The better way to think about how to choose the number of clusters is to ask, for what purpose are you running K-means?