Math 345 - Notes
k-means clustering
November 14, 2018

k-means

In this section we discuss a clustering algorithm. The input data are d-dimensional vectors, with numerical entries, and the algorithm clusters them into k similar clusters. The clusters are then identified by their centroids, or their means, hence the name k-means.

k-means Algorithm:

Let $\mathcal{D} = \{\mathbf{x}_1 \ldots, \mathbf{x}_N\}$ be the data set, each a point in d-dimensional space. Fix k.

1. Start with a set of k-means in d-dimensional space (see below how to initialize).

$$\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_k.$$

2. Assign each point to the mean to which it is closest, by using Euclidean distance: for $1 \leq j \leq k$

$$S_j = \{\mathbf{x} \in \mathcal{D} : ||\mathbf{x} - \mathbf{m}_j|| \leq ||\mathbf{x} - \mathbf{m}_i|| \text{ for all } 1 \leq i \leq k\}.$$

3. If there are changes in clustering assignments, recompute the means : for $1 \leq j \leq k$

$$\mathbf{m}_j = \frac{1}{|S_j|} \sum_{\mathbf{x} \in S_j} \mathbf{x},$$

then go to step 2.

4. Stop if there are NO changes in clustering assignments.
5. Output the k-means $\mathbf{m}_j$ and the corresponding clusters S_j.

How to initialize:

- pick k values at random from $\mathcal{D}$ for the initial $\mathbf{m}_j$ $(1 \leq j \leq k)$
- cluster $\mathcal{D}$ into k sets, and compute the means as the initial $\mathbf{m}_j$ $(1 \leq j \leq k)$
- choose the first centroid at random. For $j > 1$, pick the j-th centroid by choosing the point from the data set for which the minimum distance to previously picked centroids is largest. This way, the centroids are far from each other.

Choosing k:

- Most of the time, k is forced from the context.

- If one can choose the best k, one can consider the function

$$f(k) = \sum_{i=1}^{N} [\mathbf{x}_i - m(\mathbf{x}_i)]^2,$$

  where $m(\mathbf{x}_i)$ stands for the centroid of the data point $\mathbf{x}_i$. It is the sum of squared distances from data points to their cluster's centroid. Think of it as a measure of "error" in clustering.

  We choose the k where $f(k)$ "bends", that is, the k that makes the largest impact: it is large enough to capture difference in the clusters, yet it is not too large to overfit.

Remarks:

1. There is no training phase for this algorithm, so k-means is an unsupervised learning algorithm.

2. This algorithm may not lead to an *optimal* clustering.

3. Starting with different initial means may lead to different clusters.

4. Note that $f(k) = 0$ for $k = N$, so that would lead to no error in clustering, but it overfits and it basically does not clustering at all.

5. To visualize the clustering algorithm, try the following websites:

```
http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html
https://www.naftaliharris.com/blog/visualizing-k-means-clustering
```

Example: Let us consider the data set $\mathcal{D} = \{[-1,1], [-1,2], [0,1], [1,1], [2,2], [2,4]\}$. We will look for 2 clusters, by running through the 2-means algorithm.

- Let $\mathbf{m}_1 = [-1, 1]$, $\mathbf{m}_2 = [1, 1]$

- Compute the distance to the means and assign to the two clusters:

data point	dist^2 to $\mathbf{m}_1$	dist^2 to $\mathbf{m}_2$	cluster
[-1,1]	0	4	S_1
[-1,2]	1	5	S_1
[0,1]	1	1	S_2
[1,1]	4	0	S_2
[2,2]	10	2	S_2
[2,4]	18	10	S_2

  Cluster assignment changed, so we recompute means.

- Recompute the means:

$$\mathbf{m}_1 = \frac{[-1,1] + [-1,2]}{2} = \left[-1, \frac{3}{2}\right]$$

$$\mathbf{m}_2 = \frac{[0,1] + [1,1] + [2,2] + [2,4]}{4} = \left[\frac{5}{4}, 2\right]$$

- Compute the distance to the means and assign to the two clusters:

data point	dist^2 to $\mathbf{m}_1$	dist^2 to $\mathbf{m}_2$	cluster
[-1,1]	$1/4$	$81/16+1$	S_1
[-1,2]	$1/4$	$81/16$	S_1
[0,1]	$1+1/4$	$25/16+1$	S_1
[1,1]	$4+1/4$	$1/16+1$	S_2
[2,2]	$9+1/4$	$9/16$	S_2
[2,4]	$9+25/4$	$9/16+4$	S_2

Cluster assignment changed, so we recompute means.

- Recompute the means:

$$\mathbf{m}_1 = \frac{[-1,1]+[-1,2]+[0,1]}{3} = \left[-\frac{2}{3}, \frac{4}{3}\right]$$

$$\mathbf{m}_2 = \frac{[1,1]+[1,2]+[2,4]}{3} = \left[\frac{4}{3}, \frac{7}{3}\right]$$

- Compute the distance to the means and assign to the two clusters:

data point	dist^2 to $\mathbf{m}_1$	dist^2 to $\mathbf{m}_2$	cluster
[-1,1]	$1/9+1/9$	$49/9+16/9$	S_1
[-1,2]	$1/9+4/9$	$49/9+1/9$	S_1
[0,1]	$4/9+1/9$	$16/9+16/9$	S_1
[1,1]	$25/9+1/9$	$1/9+16/9$	S_2
[2,2]	$64/9+4/9$	$4/9+1/9$	S_2
[2,4]	$64/9+64/9$	$4/9+25/9$	S_2

Cluster assignment is NOT changed, so we STOP.

- We output the means:

$$\mathbf{m}_1 = \left[-\frac{2}{3}, \frac{4}{3}\right], \quad \mathbf{m}_2 = \left[\frac{4}{3}, \frac{7}{3}\right]$$

and the clusters:

$$S_1 = \{[-1,1], [-1,2], [0,1]\} \quad S_2 = \{[1,1], [2,2], [2,4]\}.$$