Unit 4 - lecture 14: Clustering 2

Objectives:

- Understand the limitation of the K- means algorithm
- Understand how K - Medoids algorithm is different from the K - Means algorithm
- Understand the computational complexity of the K-means and the K - Medoids algorithms
- Understand the importance of choosing the right number of clusters
- Understand elements that can be supervised in unsupervised learning

limitations of the K - Means Algorithm:

The cost we introduced last time is:

$$\text{cost}\left(C_1, \dots C_k, z^{(1)}, \dots z^{(k)}\right) = \underbrace{\sum_{j=1}^{k}}_{\text{go through all the clusters from } 1 \to k} \overbrace{\sum_{i \in C_j}}^{\text{in each cluster}} \underbrace{\| x^{(i)} - z^{(j)} \|^2}_{\text{measure the distance between the representative and each point}}$$

This K-mean Algorithm works to find the best partitioning, which optimizes this cost.

```
1. Randomly initialize $z^{(1)} \dots z^{(k)}$
2. Iterate until no change in cost
    2a. for $i = 1 \dots m$    $C_j = \{ i \mid \text{s.t. } z^{(j)} \text{ is closest to } x^{(i)} \}$
    2b. for $j = 1 \dots k$    $z^{(j)} = \frac{\sum x^{(i)}}{C_j}$
```

The limitations of this algorithm:

- the z's are actually not guaranted to be members of the original set of points x.
- In order for us to say that the representative is actually a centroid of the points, we had to utilize the squared Euclidean distance. So if we want to use another methode, this algorithm will not cut.

Introduction to the K - Medoids Algorithm:

```
1. Randomly initialize $\{ z^{(1)} \dots z^{(k)} \} \subseteq \{ x^{(1)}, \dots x^{(m)} \}$
2. Iterate until there is no change in cost
    2.a. for $i = 1 \dots m$    $C_j = \{ i \mid \text{s.t. } z^{(j)} \text{ is closest to } x^{(i)} \}$
    2.b. for $j = 1 \dots k$    $z^{(j)} \in \{ x^{1} \dots x^{(m)} \} \mid \text{s.t. } \sum_{c \in C_j} \text{dist}(x^{(i)}, z^{(j)})$ is minimal.
```

Computation complexity of K-Means and K-Medoids:

Differences between these two algorithms?

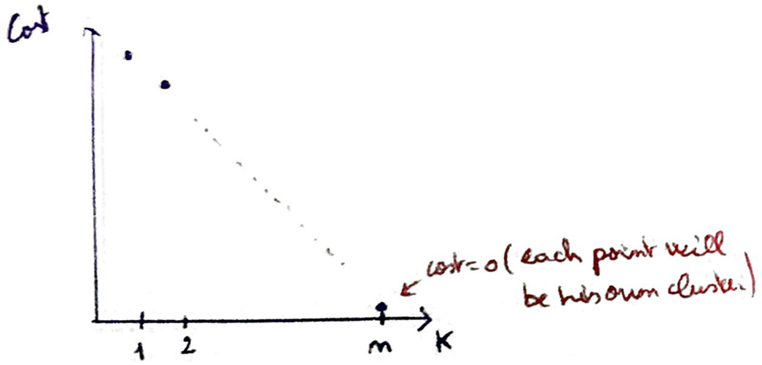
(2.b) in K-Medoids clearly seems to us more escpensive than the computations that we are doing in K-Means. If we are going to use O notation which tell us about the order of growth, let's at these algorithms and compare them in terms of their time complexity. (for one iteration)

K-Mean : $O(\overset{\text{all points}}{m}.\underset{\text{nb of clusters}}{k}.\overset{\text{size of cluster}}{d})$

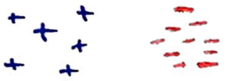
K-Medoids : $O(m^2 kd)$

Whenever you are selecting clustering algorithms which fits for your application, you may want to think about different consideration when you're finding the best clustering algorithm for your needs.

Determining the Number of clusters:



Let's say we have a supervised task



The k will always be related to the performance on the final supervised task.

People always feel that unsupervised means that we don't provide our system with any knowledge However, even though it's unsupervised in the sense that we don't provide any annotated points, we as people who develop those algorithms actually provide quite a bit of indirect supervision.

We decide whiche similarity measure to use, how many clusters to give.