14.1 Clustering motivation

14.2 K-means clustering algorithm

~ K-means always converges, but does not mean it finds the right clusters. You gotta tell the algorithm how many clusters (k) you want to have. It first initialises k centres and then find the closest point.

K-Means clustering algorithm¶

Input: Data points X and the number of clusters K\

Initialization: K initial centers for the clusters

Iterative process:

repeat

Assign each example to the closest center.

Estimate new centers as average of observations in a cluster.

until centers stop changing or maximum iterations have reached.

k-means steps:

1.set random centroids

2.assign samples to closest centroid

3.update centroids

4.repeat 2 points above until convergence

14.3 choosing k (hyper param tuning)

**1. The Elbow method**

**~ Looks at the squared sum of intra-cluster distances (a.k.a inertia)**

~ The inertia decreases as K increases but the problem is that we can't just look for a k that minimizes inertia (If each point is a cluster, the inertia is 0!!)

~ We evaluate the tradeoff "small k" vs "small intra-cluster distances"

Elbow method describes the tradeoff between them and plot a graph

**2. The silhouette method**

**~ Not dependent on the notion of cluster centers**

**~ calculated using the mean intra-cluster distance (a) and mean nearest-cluster distance (b) for each sample**

Chart

Description automatically generated

How to interpret the sihouette graphs!!!!! (Important)

~ Higher values indicate well-separated clusters. Unlike inertia, larger values are better because they indicate that the point is further away from neighbouring clusters.

~ The thickness of each silhouette indicates the cluster size.

~ Unlike inertia, the overall silhouette score gets worse as you add more clusters because you end up being closer to neighbouring clusters.

~ Thus, as with intertia, you will not see a “peak” value of this metric that indicates the best number of clusters.

~ The shape of each silhouette indicates the “goodness” for points in each cluster.

~ The length (or area) of each silhouette indicates the goodness of each cluster.

~ A slower dropoff (more rectangular) indicates more points are “happy” in their cluster.

Is it possible to pick K in a smart way?

Yes! We can use the so-called K-Means++.

Intuitively, it picks the initial centroids which are far away from each other.

In other words, K-Means++ gives more chance to select points that are far away from centroids already picked.

By default sklearn uses this strategy for initialization.

Important things about K-means clustering

~ The centroids live in the same space as of the dataset but they are not actual data points, but instead are average points.

**~ It always converges. Convergence is dependent upon the initial centers and it may converge to a sub-optimal solution.**