Week 13

August 8, 2021

Unsupervised learning

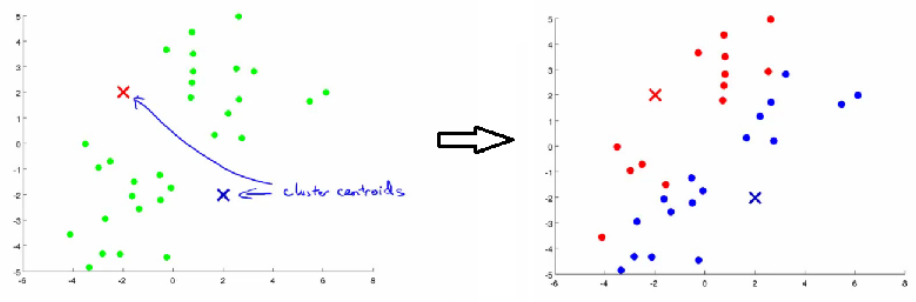
- Try to figure out the structure of the data.
- The clustering algorithm organizes data based on data characteristics.
- Market segmentation is categorizing clients into different market categories.
- Social network analysis.
- Computer clusters and data centers are organized for network structure and location.
- Understanding galaxy creation through astronomical data analysis.

K-means algorithm

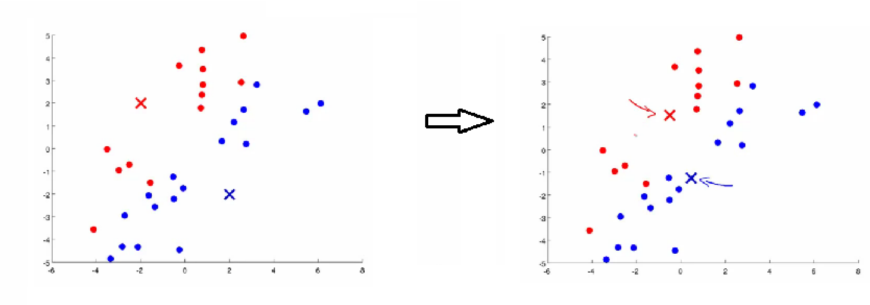
- Would you like an algorithm to automatically arrange data into coherent clusters?
- By far the most used clustering algorithm is K-means.

Algorithm overview:

1. Assign k locations at random as cluster centroids.
2. Go through each example and assign each point to one of the k clusters based on which center it is closest to.



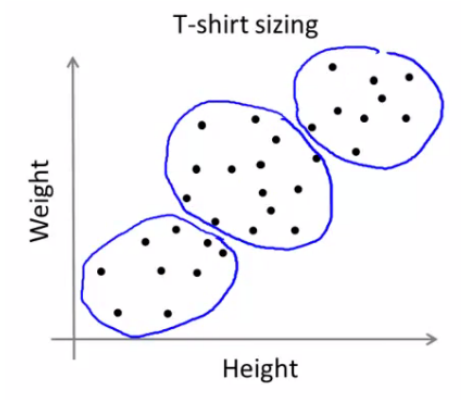
3. Move to the average of the similarly allocated data-points for each centroid.



4. Repeat 2) and 3) until convergence.

K-means for non-separated clusters

- So far, we've looked at K-means, which has well-defined clusters.
- However, K-means is frequently used on datasets with poorly defined clusters.
- As an example, consider t-shirt sizes. How large do you make them if you want three sizes (S,M,L)?
- As a result, three clusters are formed, even if they are not actually there.
- This is an example of market segmentation; create items that are tailored to the demands of your subpopulations.



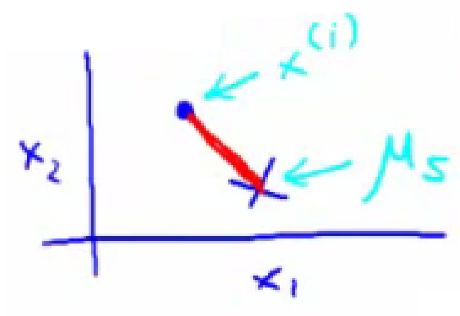
K means optimization objective

- K-means, like the supervised learning functions we've examined, has an optimization goal.
- While K-means is running, we keep track of two sets of variables.

- c^i is the index of clusters $1, 2, ..., K$ to which x^i is currently assigned.
- μ_k, is the cluster associated with centroid k.
- μ_c^i, is the cluster centroid of the cluster to which example x^i has been assigned to.
- We may write the optimization objective using this notation:

$$J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K) = \frac{1}{m}\sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

i.e. squared distances between training example x^i and the cluster centroid to which x^i has been assigned to.

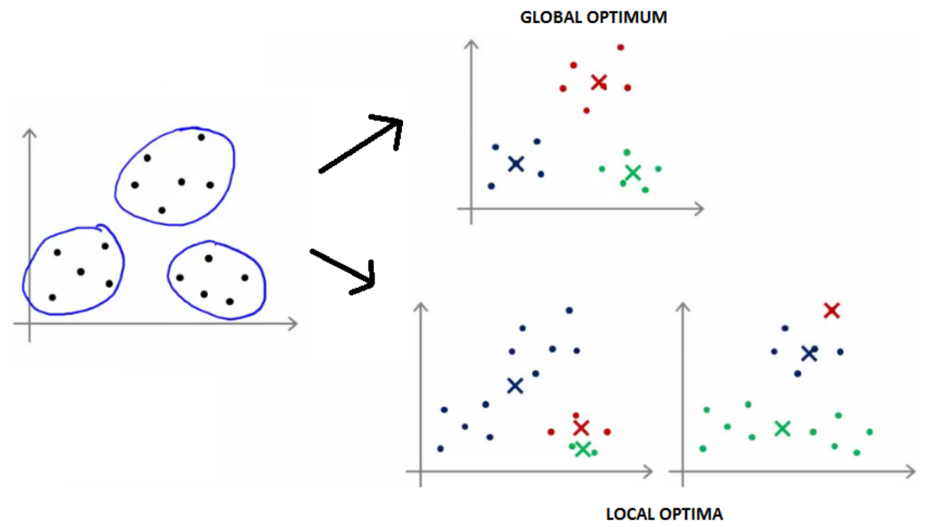


When we look at the k-means method:

- The cluster assigned step is minimizing $J(...)$ with respect to $c_1, c_2...c_i$ i.e. find the centroid closest to each example. Doesn't change the centroids themselves.
- The move centroid step. We can show this step is choosing the values of μ which minimizes $J(...)$ with respect to μ.
- So, we're partitioning the algorithm into two parts: First part minimizes the c variables. Second part minimizes the J variables.

Random initialization

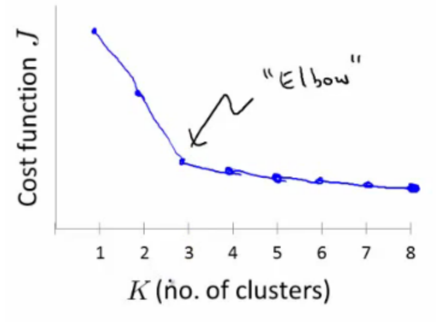
Depending on the starting setting, K means might converge to different solutions.



- Randomly initialize K-means.
- For each n (e.g. 100) random initialization run K-means.
- Then compute the distortion on the set of cluster assignments and centroids at convergent.
- End with n ways of cluster the data.
- Pick the clustering which gave the lowest distortion.

Elbow method

- How do we choose the number of clusters K?
- Vary K and compute cost function at a range of K values.
- $J(...)$'s minimum value should decrease as K rises (i.e. you decrease the granularity so centroids can better optimize).
- Look for the "elbow" on the graph (K vs $J()$).


Cost function J
"Elbow"
K (no. of clusters)