K-Means Clustering

Derick Smith

MCSC 6020 G, Fall 2019

What Is A Cluster?

- Collection of points with common similarity?
- Defined with mathematical rigor?
- Philosophical consensus?

What Is A Cluster?

- Collection of points with common similarity?
- Defined with mathematical rigor?
- Philosophical consensus?

What Is A Cluster?

- Collection of points with common similarity?
- Defined with mathematical rigor?
- Philosophical consensus?

Lloyd Algorithm: Minization of Squared Distance.

- Basic algorithm chooses k initial centroids at random.
- Each data point assigned to closest centroid.
- For each the average position per centroid group becomes the new centroid.
- Repeat assignments to closest centroid and update, otherwise, complete.
- $f(n) \in O(kndt)$, k := number of clusters, n:= samples points, d:=dimensions, t:=iterations.
- Exceptionally simple algorithm, easy to implement.

Lloyd Algorithm: Minization of Squared Distance.

- Basic algorithm chooses k initial centroids at random.
- Each data point assigned to closest centroid.
- For each the average position per centroid group becomes the new centroid.
- Repeat assignments to closest centroid and update, otherwise, complete.
- $f(n) \in O(kndt)$, k := number of clusters, n:= samples points, d:=dimensions, t:=iterations.
- Exceptionally simple algorithm, easy to implement.

Lloyd Algorithm: Minization of Squared Distance.

- Basic algorithm chooses k initial centroids at random.
- Each data point assigned to closest centroid.
- For each the average position per centroid group becomes the new centroid.
- Repeat assignments to closest centroid and update, otherwise, complete.
- $f(n) \in O(kndt)$, k := number of clusters, n:= samples points, d:=dimensions, t:=iterations.
- Exceptionally simple algorithm, easy to implement.

Lloyd Algorithm: Minization of Squared Distance.

- Basic algorithm chooses k initial centroids at random.
- Each data point assigned to closest centroid.
- For each the average position per centroid group becomes the new centroid.
- Repeat assignments to closest centroid and update, otherwise, complete.
- $f(n) \in O(kndt)$, k := number of clusters, n:= samples points, d:=dimensions, t:=iterations.
- Exceptionally simple algorithm, easy to implement.

Lloyd Algorithm: Minization of Squared Distance.

- Basic algorithm chooses k initial centroids at random.
- Each data point assigned to closest centroid.
- For each the average position per centroid group becomes the new centroid.
- Repeat assignments to closest centroid and update, otherwise, complete.
- $f(n) \in O(kndt)$, k := number of clusters, n:= samples points, d:=dimensions, t:=iterations.
- Exceptionally simple algorithm, easy to implement.

Helper Function: k-means++.

- First centroid selected at random.
- Farthest centroid from first becomes second.
- $(p+1)^{th}$ centroid is farthest neighbor from i^{th} centroid, $i \in \{1, p\}$.
- Reduction in t, iterations.

Helper Function: k-means++.

- First centroid selected at random.
- Farthest centroid from first becomes second.
- $(p+1)^{th}$ centroid is farthest neighbor from i^{th} centroid, $i \in \{1, p\}$.
- Reduction in t, iterations.

Helper Function: k-means++.

- First centroid selected at random.
- Farthest centroid from first becomes second.
- $(p+1)^{th}$ centroid is farthest neighbor from i^{th} centroid, $i \in \{1, p\}$.
- Reduction in t, iterations.

Helper Function: k-means++.

- First centroid selected at random.
- Farthest centroid from first becomes second.
- $(p+1)^{th}$ centroid is farthest neighbor from i^{th} centroid, $i \in \{1, p\}$.
- Reduction in t, iterations.

Variation: mini-batch.

- The same as Lloyd except performed on random subset of data.
- Converges faster with marginal increase in error.
- Dataset must be "relatively" large.
- Error from true global minimum can be evaluated "quickly" using mini-batch centroids.
- Reduction in t, iterations.

Variation: mini-batch.

- The same as Lloyd except performed on random subset of data.
- Converges faster with marginal increase in error.
- Dataset must be "relatively" large.
- Error from true global minimum can be evaluated "quickly" using mini-batch centroids.
- Reduction in t, iterations.

Variation: mini-batch.

- The same as Lloyd except performed on random subset of data.
- Converges faster with marginal increase in error.
- Dataset must be "relatively" large.
- Error from true global minimum can be evaluated "quickly" using mini-batch centroids.
- Reduction in t, iterations.

Variation: mini-batch.

- The same as Lloyd except performed on random subset of data.
- Converges faster with marginal increase in error.
- Dataset must be “relatively” large.
- Error from true global minimum can be evaluated “quickly” using mini-batch centroids.
- Reduction in t, iterations.

Variation: mini-batch.

- The same as Lloyd except performed on random subset of data.
- Converges faster with marginal increase in error.
- Dataset must be “relatively” large.
- Error from true global minimum can be evaluated “quickly” using mini-batch centroids.
- Reduction in t, iterations.

Alternative: DBScan Algorithm.

- A predefined ε is chosen for maximum radius of membership.
- For some $x_p \in c_k$, if $||x_p - x_q|| < \varepsilon$ then $x_q \in c_k$.
- Capable of grouping non-globular clusters.
- Worst-case time complexity, $O(n^2)$.

Alternative: DBScan Algorithm.

- A predefined ε is chosen for maximum radius of membership.
- For some $x_p \in c_k$, if $||x_p - x_q|| < \varepsilon$ then $x_q \in c_k$.
- Capable of grouping non-globular clusters.
- Worst-case time complexity, $O(n^2)$.

Alternative: DBScan Algorithm.

- A predefined ε is chosen for maximum radius of membership.
- For some $x_p \in c_k$, if $||x_p - x_q|| < \varepsilon$ then $x_q \in c_k$.
- Capable of grouping non-globular clusters.
- Worst-case time complexity, $O(n^2)$.

Alternative: DBScan Algorithm.

- A predefined ε is chosen for maximum radius of membership.
- For some $x_p \in c_k$, if $||x_p - x_q|| < \varepsilon$ then $x_q \in c_k$.
- Capable of grouping non-globular clusters.
- Worst-case time complexity, $O(n^2)$.

Examples.

Example

Varying k-means Clustering Parameters.

Example

Image Compression: 3D RGB Plot.

Examples.

Example

Varying k-means Clustering Parameters.

Example

Image Compression: 3D RGB Plot.

Summary

- Clustering is a broad concept.
- Lloyd just one way of defining a cluster.
- Exceptionally popular in unsupervised analysis applications.