Notes - k-means

Samuele Nicolò Straccialini

January 2025

Some pseudocodes and notes on k-means clustering algorithms.

1 Initializations

Let k the number of clusters, X an $N \times M$ matrix of N datapoints in M dimensions.

1.1 random

Algorithm 1 random initialization

Require: k, X
  Select k points at random

1.2 random-data

Algorithm 2 random-data initialization

Require: k, X
  for datapoint **do** $x \in X$
    assign x to one of the k clusters
  end for

1.3 greedy

Algorithm 3 greedy initialization

Require: k, X
  choose μ_0 randomly from X
  for $i = 1, ..., k-1$ **do**
    for datapoint $x \in X$ **do**
      $D(x) = \min_j \|x - \mu_j\|^2$ ▷ Squared distance to closest centroid
    end for
    $\mu_i = \arg\max D(x)$ ▷ Select point with max distance
  end for

1.4 k-means++

Algorithm 4 k-means++ initialization

Require: k, X
  choose μ_0 randomly from X
  for $i = 1, ..., k-1$ **do**
    for datapoint $x \in X$ **do**
      $D(x) = \min_j \|x - \mu_j\|^2$ ▷ Squared distance to closest centroid
    end for
    $P(x) = \frac{D(x)}{\sum_{x' \in X} D(x)}$
    Select μ_i based on the probability distribution $P(x)$
  end for

2 Clustering

2.1 lloyd

2.2 hartigan

2.3 extended-hartigan

2.4 safe-hartigan

2.5 binary-hartigan