k-Medians algorithm

Procheta Sen

UNIVERSITY OF LIVERPOOL

Representative-based algorithms

- Let k be the number of clusters
- Let $D = \{\overline{X}_1, \ldots, \overline{X}_n\}$ be the dataset
- The goal is to determine k representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ that minimise the following objective function

$$\sum_{i=1}^{n} \left[\min_j d(\overline{X}_i, \overline{Y}_j) \right]$$

i.e. the sum of the distances of the objects to their closest representatives needs to be minimised.

Representative-based algorithms

We obtain specific algorithms by specifying

- the way of choosing representatives, and
- the distance function $d(\cdot\,,\cdot)$

In general representatives do not necessarily belong to the dataset.

General k-representatives approach

- **Initialise**: pick initial k representatives
- **Iteratively refine**:
  - (**assign step**) Assign each object to its closest representative using distance function $d(\cdot,\cdot)$. Denote the corresponding clusters $C_1, \ldots, C_k$
  - (**optimise step**) Determine the optimal representative $\overline{Y}_j$ for each cluster C_j that minimises its local objective function $\sum_{\overline{X}_i \in C_j} \left[d\left(\overline{X}_i, \overline{Y}_j\right) \right]$

k-Means algorithm

- Representatives are chosen **not necessarily** from the dataset
- The distance function is **squared Euclidean distance**

k-Means algorithm

The objective function $$\min_{\overline{Y}_1,\ldots,\overline{Y}_k} \sum_{i=1}^{k} \sum_{\overline{X}\in C_i} \|\overline{X} - \overline{Y}_i\|^2$$

where C_i consists of the objects that are closest to $\overline{Y}_i$.

We want to minimise the total distance between data objects and their cluster representatives $\overline{Y}_1, \ldots, \overline{Y}_k$

This objective function is called the *within cluster sum of squares* (WCSS) objective

k-Means algorithm

Assume that the clusters $C_1, C_2, \ldots, C_k$ are fixed.

Find the set $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ of representatives such that

$$f_{C_1,\ldots,C_k}(\overline{Y}_1, \ldots, \overline{Y}_k) = \sum_{i=1}^{k} \sum_{\overline{X} \in C_i} \|\overline{X} - \overline{Y}_i\|^2 \text{ is minimised.}$$

$$\frac{\partial f_{C_1,\ldots,C_k}(\overline{Y}_1, \ldots, \overline{Y}_k)}{\partial \overline{Y}_i} = -\sum_{\overline{X} \in C_i} 2(\overline{X} - \overline{Y}_i) = 0 \qquad \overline{Y}_i = \frac{1}{|C_i|} \sum_{\overline{X} \in C_i} \overline{X}$$

Just compute the centroid (mean) of each cluster and that will give you the new **optimal** cluster representatives

k-Means algorithm

What if **Manhattan distance** is more appropriate for our application than the **Euclidean distance**?

Manhattan distance (or L^1 distance) $$ManDist(\overline{X}, \overline{Y}) = \|\overline{X} - \overline{Y}\|_1 = \sum_{i=1}^{d} |x_i - y_i|$$

k-Medians algorithm

The objective function $$\min_{\overline{Y}_1,\ldots,\overline{Y}_k} \sum_{i=1}^{k} \sum_{\overline{X}\in C_i} \|\overline{X} - \overline{Y}_i\|_1$$

where C_i consists of the objects that are L^1-closest to $\overline{Y}_i$.

We want to minimise the total L^1-distance between data objects and their cluster representatives $\overline{Y}_1, \ldots, \overline{Y}_k$

k-Medians algorithm

Assume that the clusters $C_1, C_2, \ldots, C_k$ are fixed.
Find the set $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ of representatives such that

$$f_{C_1,\ldots,C_k}(\overline{Y}_1, \ldots, \overline{Y}_k) = \sum_{i=1}^{k} \sum_{\overline{X} \in C_i} \|\overline{X} - \overline{Y}_i\|_1 \text{ is minimised.}$$

Let's consider a fixed cluster C. What is the new representative Y that minimises $\sum_{\overline{X} \in C} \|\overline{X} - \overline{Y}\|_1$?

Let $\overline{X} = (\overline{X}^{(1)}, \overline{X}^{(2)}, \ldots, \overline{X}^{(d)})$ and $\overline{Y} = (\overline{Y}^{(1)}, \overline{Y}^{(2)}, \ldots, \overline{Y}^{(d)})$.

k-Medians algorithm

$$\arg\min_{\overline{Y}} \sum_{\overline{X} \in C} \|\overline{X} - \overline{Y}\|_1$$

$$\sum_{\overline{X} \in C} \|\overline{X} - \overline{Y}\|_1 = \sum_{\overline{X} \in C} \sum_{i=1}^{d} \left| \overline{X}^{(i)} - \overline{Y}^{(i)} \right| = \sum_{i=1}^{d} \sum_{\overline{X} \in C} \left| \overline{X}^{(i)} - \overline{Y}^{(i)} \right|$$

Since the coordinates are independent, to minimise the later sum we can minimise each of its terms independently.

More specifically, if for $i = 1, \ldots, d$, the number $\overline{Y}'^{(i)}$ minimises $\sum_{\overline{X} \in C} \left| \overline{X}^{(i)} - \overline{Y}^{(i)} \right|$, then

$\overline{Y}' = (\overline{Y}'^{(1)}, \overline{Y}'^{(2)}, \ldots, \overline{Y}'^{(d)})$ minimises $\sum_{\overline{X} \in C} \|\overline{X} - \overline{Y}\|_1$

k-Medians algorithm

Think about them as the i-th coordinates of the objects in cluster C

Given s numbers $\overline{X}_1^{(i)}, \overline{X}_2^{(i)}, \ldots, \overline{X}_s^{(i)}$, the number

$$\overline{Y}'^{(i)} = \text{median}(\overline{X}_1^{(i)}, \overline{X}_2^{(i)}, \ldots, \overline{X}_s^{(i)})$$

minimises $\displaystyle\sum_{t=1}^{s} \left| \overline{X}_t^{(i)} - \overline{Y}^{(i)} \right|$.

Def. A median of a sequence of numbers is any value such that at most half of the values is less than the proposed median and at most half is greater than the proposed median.

k-Medians algorithm

Hence an object $\overline{Y}' = (\overline{Y}'^{(1)}, \overline{Y}'^{(2)}, \ldots, \overline{Y}'^{(d)})$ that minimises

$$\arg\min_{\overline{Y}} \sum_{\overline{X} \in C} \|\overline{X} - \overline{Y}\|_1$$

is a **median object** of (the objects in) the cluster C,

where $\overline{Y}'^{(i)}$ is the median of the the i-th coordinates of the objects in the cluster C.

k-Medians algorithm

`k-MediansClustering` (Number of clusters: k, Dataset: $\{\overline{X}_1, \ldots, \overline{X}_n\}$)

1. Initialisation phase

Choose k cluster representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ from the dataset randomly

2. Assignment phase

Assign all objects in the dataset to the L^1-closest representative. The resulting clusters: $C_1, \ldots, C_k$

3. Optimisation phase

Compute the new representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ as the **medians** of the current clusters $C_1, \ldots, C_k$

Repeat Phases 2 and 3 until convergence. (Convergence is either “no objects have moved among clusters” or “fixed number of iterations specified by user”)