Ramesh

Rough k-means in Q

Version 0.1

January 19, 2024

Contents

Rough k-means in Q 2
1 Introduction 2
1.1 Notation 2
1.2 Invariants 3
1.3 Mathematics to Code 3
2 Computation 3
2.1 The Update Step 3
2.2 The Assignment Step 4

Rough k-means in Q

1 Introduction

We describe the rough k-means algorithm [**?**]. We assume that the reader is familiar with the k-means algorithm [**?**]. We make a minor change to their terminology, refering to the

- lower approximation as the "inner" approximation
- upper approximation as the "outer" approximation

1.1 Notation

1. Let n_J be the number of features. We shall use j as the feature index.
2. Let n_I be the number of objects/instances/observations We shall use i as the instance index.
3. Let n_K be the number of means/centroids. We shall use k as the centroid index.
4. Let X be a set of observations in n_J dimensional space, such that $X_{j,i}$ is the value of the j^{th} feature of the i^{th} instance.

   We store the observations as n_J vectors of length n_I. So, X_j is the vector corresponding to the j^{th} feature and $X_{j,i}$ is the value of the j^{th} feature of the i^{th} observation.
5. Let $\mu_{k,j}$ be the value of the j^{th} feature of the k^{th} centroid
6. Let $I_{k,i}$ be true if instance i is part of the **I**nner approximation of centroid k
7. Let $O_{k,i}$ be true if instance i is part of the **O**uter approximation of centroid k
8. We shall treat the boolean value "true" and the integer 1 interchangeably.
9. We shall treat the boolean value "false" and the integer 0 interchangeably.

10. Let w_I and w_O be the weights assigned to the Inner and Outer approximations
11. Let α be the threshold for determining whether an instance belongs to the outer approximation. This is explained in Section 2.1
12. Identifying i, j, k as feature indexes simplifies the notation. For example, $\sum_k$ is actually $\sum_{k=1}^{k=n_K}$. Similarly, $\forall k$ means for all centroids, numbered $1, \ldots, n_K$
13. We use the following conventions for types

    F4 4-byte floating point

    I4 4-byte signed integer

    B1 1-bit boolean
14. Define $\delta(x, y) = 1$ if $x = y$ and 0 otherwise.

1.2 Invariants

Invariant 1 *An instance can belong to the inner approximation of at most one centroid.*
$\forall i : \sum_k I_{k,i} = 1$

Invariant 2 *If an instance is not part of any inner approximation, it must belong to two or more outer approximations. This implies that an instance cannot belong to only a single boundary region.*
$\sum_k I_{k,i} = 0 \Rightarrow \sum_k O_{k,i} \geq 2$

1.3 Mathematics to Code

The mathematical terms we use are terse, the variable names in the code somewhat more verbose. A mapping is provided in Table 1.

2 Computation

2.1 The Update Step

1. $\forall k : d_k$ is a F4 Vector of length n_I such that $d_{k,i}$ is distance of instance i from centroid k
2. $\bar{d}$ is a F4 Vector of length n_I such that $\bar{d}_i$ is smallest distance of instance i from any centroid i.e., $\bar{d}_i = \min_k d_{k,i}$

Math	Code	Type	Length
d_k	dist[k]	F4 Vector	n_I
$\bar{d}$	best_dist	F4 Vector	n_I
$\bar{k}$	best_clss	I4 Vector	n_I
O_k	is_outer[k]	B1 Vector	n_I
N^O_k	num_in_outer[k]	I4 Vector	n_I
$\hat{I}$	inner	I4 Vector	n_I

Table 1: Math symbols to names in code

3. $\bar{k}$ is an I4 Vector of length n_I such that $\bar{k}_i$ identifies the centroid that is closest to instance i. Note that $\bar{k}_i \in [1, \dots n_K]$
4. $\forall k : O_k$ is a B1 Vector of length n_I such that $d_{k,i} \leq \bar{d}_i \times \alpha \Rightarrow O_{k,i} = \mathrm{true}$
5. N^O is an I4 Vector of length n_I, where $N^O_i = \sum_k O_k[i]$
6. Let $\hat{I}$ be a Vector of length n_I such that $N^O_i \geq 2 \Rightarrow \hat{I}_i = 0$; else, $\hat{I}_i = \bar{k}_i$. What we are doing here is stating that if nobody else has a claim on instance i, then it belongs to the inner approximation of $\bar{k}_i$. In other words,
   (a) $\hat{I}_i = 0 \Rightarrow$ instance i not in inner approximation of any centroid
   (b) $\hat{I}_i = k' \Rightarrow$ instance i in inner approximation of centroid k'

2.2 The Assignment Step

1. The contribution of the inner and outer sets are weighted and then combined into the value of the centroid as follows: $\mu_{k,j} = w_I \times \mu^I_{k,j} + w_O \times \mu^O_{k,j}$
2. $\mu^I_{k,j} = \frac{\sum_i \delta(\hat{I}_i, k) X_{k,i}}{D^I_k}$
3. $\mu^O_{k,j} = \frac{\sum_i (O_{k,i} \times X_{k,i})}{D^O_k}$
4. $D^I_k = \sum_i \delta(\hat{I}, k)$
5. $D^O_k = \sum_i O_{k,i}$