

Unit 4 Unsupervised Learning (2

Course > weeks)

> Lecture 14. Clustering 2 >

4. Computational Complexity of

K-Means and K-Medoids

Audit Access Expires May 11, 2020

You lose all access to this course, including your progress, on May 11, 2020.

4. Computational Complexity of K-Means and K-Medoids Computation Complexity of K-Means and K-Medoids





Video

Download video file

Transcripts

<u>Download SubRip (.srt) file</u> <u>Download Text (.txt) file</u>

Computational Complexity of K-Means

1/1 point (graded)

Remember that the K-Means algorithm is given by

- 1. Randomly select z_1, \ldots, z_K
- 2. Iterate
 - 1. Given $z_1, \ldots z_K$, assign each $x^{(i)}$ to the closest z_j , so that

$$\operatorname{Cost}\left(z_{1}, \ldots z_{K}
ight) = \sum_{i=1}^{n} \min_{j=1,...,k} \left\|x^{(i)} - z_{j}
ight\|^{2}$$

2. Given C_1,\ldots,C_K find the best representatives z_1,\ldots,z_K , i.e. find z_1,\ldots,z_K such that

$$z_j = rac{\sum_{i \in C_j} x^{(i)}}{|C_j|}$$

Assuming that there are n data points $\{x_1,\ldots,x_n\}$, K clusters and representatives,and each $x_i\in\{x_1,\ldots,x_n\}$ is a vector of dimension d, what is the computational complexity for one complete iteration of the k-means algorithm? That is, find the time (or the number of steps) it takes to complete steps 2.1 and 2.2.

Note on Big-O notation

We often describe computational complexity using the "Big-O" notation. For example, if the number of steps involved is $5n^2+n+1$, then we say it is "of

order n^2 " and denote this by $\mathcal{O}\left(n^2\right)$. When n is large, the highest order term $5n^2$ dominates and we drop the scaling constant 5.

More formally, a function $f\left(n\right)$ is of order $g\left(n\right)$, and we write $f\left(n\right)\sim\mathcal{O}\left(g\left(n\right)\right)$, if there exists a constant C such that

$$f(n) < Cg(n)$$
 as n grows large.

In other words, the function f does not grow faster than the function g as n grows large.

The big-O notation can be used also when there are more input variables. For example, in this problem, the number of steps necessary to complete one iteration depends on the number of data points n, the number of clusters K, the dimension d of each vector x_i . Hence, the number of steps required are of $\mathcal{O}\left(g\left(n,K,d\right)\right)$ for some function $g\left(n,K,d\right)$.

<u>Hide</u>

- $\bigcirc \mathcal{O}(n)$
- $\bigcirc \mathcal{O}\left(nK
 ight)$
- $\bigcirc \mathcal{O}\left(nK^2
 ight)$
- $left{lack} \mathcal{O}\left(ndK
 ight)$



Solution:

In line 2.1, we go through each of the n x_i , and iterate through each of the k z_j 's for each x_i (to find the closest z_j). This iteration is $\mathcal{O}(nK)$. And because each x_i has length d, the total iteration is $\mathcal{O}(ndK)$.

Line 2.2 is similar.

Note that because 2.1 and 2.2 both take $\mathcal{O}\left(ndK\right)$, one complete iteration takes $\mathcal{O}\left(ndK\right)$.

Submit

You have used 1 of 2 attempts

1 Answers are displayed within the problem

Computational Complexity of K-Medoids

2/2 points (graded)

Remember that the K-Medoids algorithm is given by

- 1. Randomly select z_1,\ldots,z_K
- 2. Iterate
 - 1. Given $z_1, \ldots z_K$, assign each $x^{(i)}$ to the closest z_i , so that

$$\operatorname{Cost}\left(z_{1}, \ldots z_{K}
ight) = \sum_{i=1}^{n} \min_{j=1,...,k} \operatorname{dist}\left(x^{(i)}, z_{j}
ight)$$

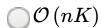
2. Given $C_j \in \left\{C_1, \dots, C_K
ight\}$ find the best representative $z_j \in \left\{x_1, \dots, x_n
ight\}$ such that

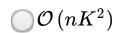
$$\sum_{x^{(i)} \in C_i} \mathrm{dist}\,(x^{(i)},z_j)$$

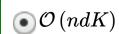
is minimal.

What is the complexity of step 2.1?



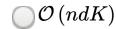








Now what is the complexity of step 2.2?



$$\bigcirc \mathcal{O}\left(nK^2
ight)$$

$$\bigcirc \mathcal{O}\left(nk^2d
ight)$$

$$igodesign{ igodesign{ igonesign{ igodesign{ igonesign{ igodesign{ igonesign{ igodesign{ igonesign{ igo} igodesign{ igo} igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ igo} igoesign{ igo} igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igoesign{ igo} igoesign{ i$$



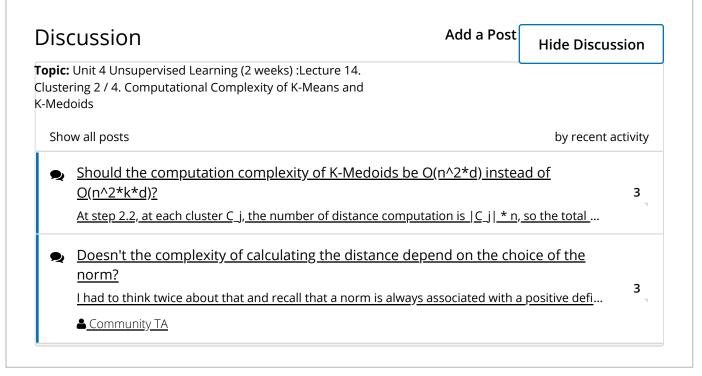
Solution:

Note that step 2.1 of the K-Medoids is the same as that of K-Means, so the time complexity is $\mathcal{O}\left(ndK\right)$. Note that step 2.2 of K-Medoids has an additional loop of iterating through the n points $z_j \in \left\{x_1,\ldots,x_n\right\}$ which takes $\mathcal{O}\left(n\right)$. Thus step 2.2 takes $\mathcal{O}\left(n^2dK\right)$.

Submit

You have used 2 of 3 attempts

1 Answers are displayed within the problem



© All Rights Reserved