Scalable K-means++ Problem Set Answer Key

Zack Nawrocki, Ziping Song, and Jeff Lehmann

December 2020

1 Conceptual Questions

1) In the presentation, we calculated the distances for clustering using the Euclidean distance algorithm. It is also possible to calculate distances, using the Manhattan distance algorithm (you can learn about the formula here https://xlinux.nist.gov/dads/HTML/manhattanDistance.html). When could the Manhattan distance be useful?
(Hint: Page 1 - Abstract: https://bib.dbvis.de/uploadedFiles/155.pdf)

Solution: The Manhattan distance is very useful in higher levels of dimensions. As the number of dimensions increase, data often becomes more sparse, and traditional indexing and algorithmic techniques fail from an efficiency and/or effectiveness perspective.

2) Give an example, different than one discussed in the presentation, of what k-means clustering can be useful for, in real-world situations.

Possible answer: Classifying a newly-discovered virus, that is mutating quickly, into different versions, could take advantage of scalable k-means ++, and similar algorithms. To figure out how dynamic a vaccine would need to be, to be effective for the closely-related versions of the virus, we could use the classification analysis results, as a preprocessor for supervised learning.

3) With k-means algorithms, we utilize averaging. Similar algorithms exist, such as the k-medians algorithm (utilizing medians). Give one situation where this algorithm could be a better approach.

Solution: k-medians algorithms could be a better approach, if a handful of extreme outliers poorly generalize the data, when using k-means algorithms.

4) Why would oversampling, in the Scalable k-means++ algorithm, improve initialization?

Solution: Oversampling allows us to update our clustering distribution less frequently during initialization.

2 Proof Question

1) As discussed in the presentation Theorem 1 is as follows:

If an α approximation algorithm is used in Step 8, then Algorithm k-means// obtains a solution that is an O(α)-approximation to k-means.

To prove this Bahmani proves that the expected cost of adding new centers at each iteration, where ϕX(C) is the cost of the current centers in C, ϕX(C $\cup$ C') is the cost of adding new centers, C', to C, ϕ^* is a constant factor that shows the drop in solution cost after each iteration (Hint: equal to 0 in first iteration) and α is approximately $e^{\frac{-l}{2k}}$, in Theorem 2:

Theorem 2

$$\mathbb{E}[\phi X(C \cup C')] \leq 8\phi^* + \frac{1+\alpha}{2}\phi X(C)$$

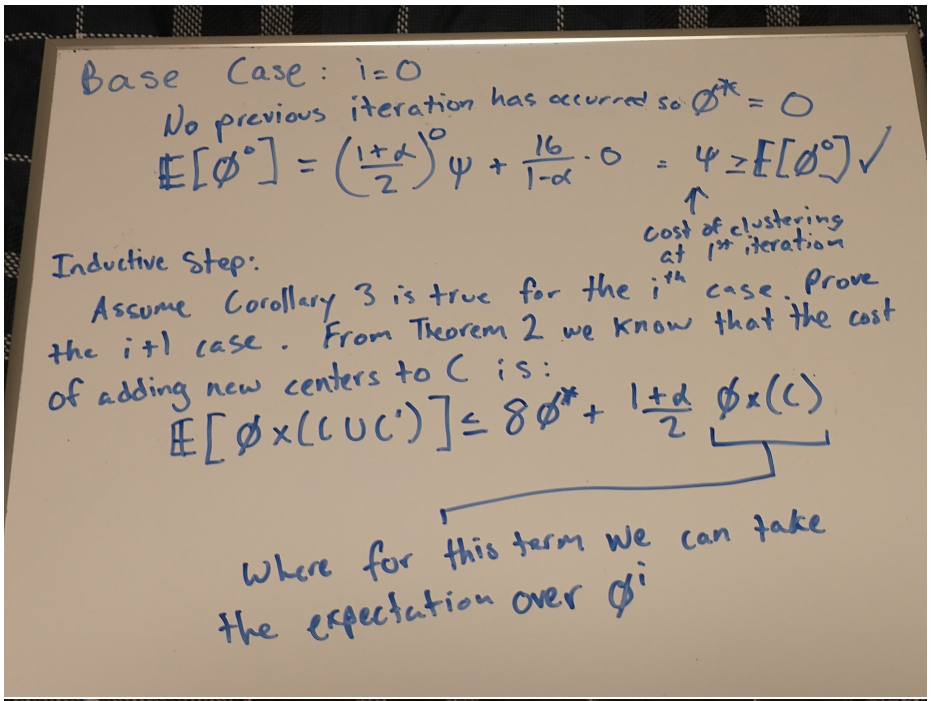
To prove Theorem 1, Bahmani proposes the following Corollary which would prove that there is a constant factor approximation to k-means after O(log ψ) rounds if we prove that the cost of clustering at the ith iteration is:

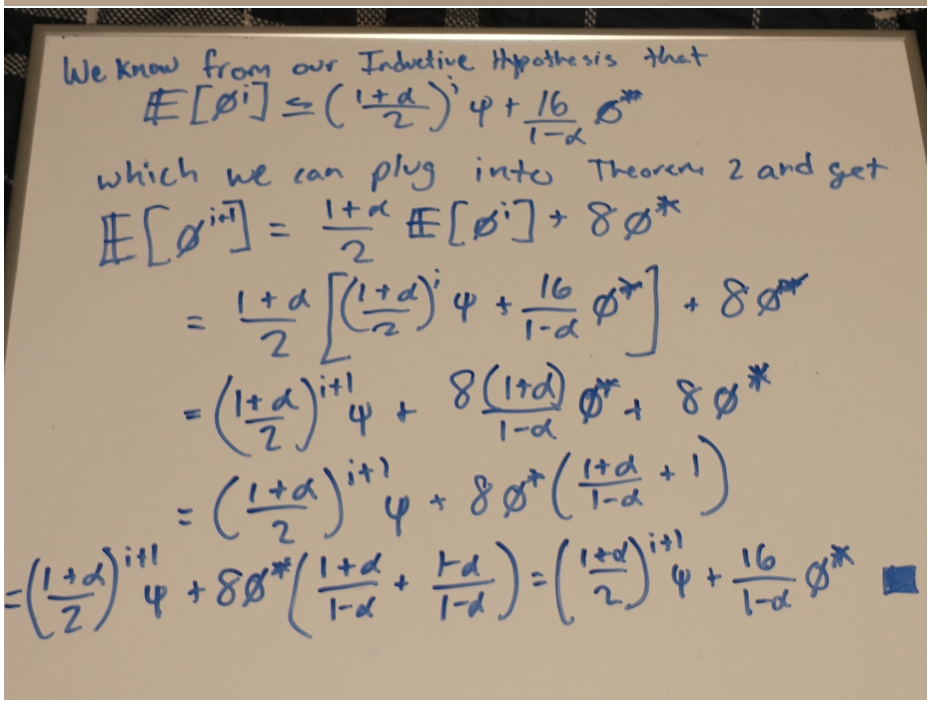
Corollary 3

$$\mathbb{E}[\phi^{(i)}] \leq (\frac{1+\alpha}{2})^{(i)}\psi + \frac{16}{1-\alpha}\phi^*$$

Prove Corollary 3 using induction which therefore proves the approximation guarantee of Theorem 1. (Hint use Theorem 2 in your inductive step)

Solution:


Base Case: i=0
No previous iteration has occurred so $\phi^* = 0$
$\mathbb{E}[\phi^0] = \left(\frac{1+\alpha}{2}\right)^0 \psi + \frac{16}{1-\alpha} \cdot 0 = \psi \geq \mathbb{E}[\phi^0]$ ✓
cost of clustering at 1st iteration
Inductive Step:
Assume Corollary 3 is true for the i^{th} case. Prove the i+1 case. From Theorem 2 we know that the cost of adding new centers to C is:
$\mathbb{E}[\phi_X(C \cup C')] \leq 8\phi^* + \frac{1+\alpha}{2}\phi_X(C)$
Where for this term we can take the expectation over ϕ^i


We know from our Inductive Hypothesis that
$\mathbb{E}[\phi^i] \leq \left(\frac{1+\alpha}{2}\right)^i \psi + \frac{16}{1-\alpha}\phi^*$
which we can plug into Theorem 2 and get
$\mathbb{E}[\phi^{i+1}] = \frac{1+\alpha}{2}\mathbb{E}[\phi^i] + 8\phi^*$
$= \frac{1+\alpha}{2}\left[\left(\frac{1+\alpha}{2}\right)^i \psi + \frac{16}{1-\alpha}\phi^*\right] + 8\phi^*$
$= \left(\frac{1+\alpha}{2}\right)^{i+1}\psi + \frac{8(1+\alpha)}{1-\alpha}\phi^* + 8\phi^*$
$= \left(\frac{1+\alpha}{2}\right)^{i+1}\psi + 8\phi^*\left(\frac{1+\alpha}{1-\alpha} + 1\right)$
$= \left(\frac{1+\alpha}{2}\right)^{i+1}\psi + 8\phi^*\left(\frac{1+\alpha}{1-\alpha} + \frac{1-\alpha}{1-\alpha}\right) = \left(\frac{1+\alpha}{2}\right)^{i+1}\psi + \frac{16}{1-\alpha}\phi^*$ ■