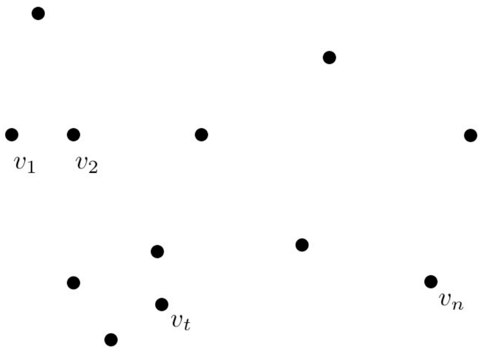
Online K-Means

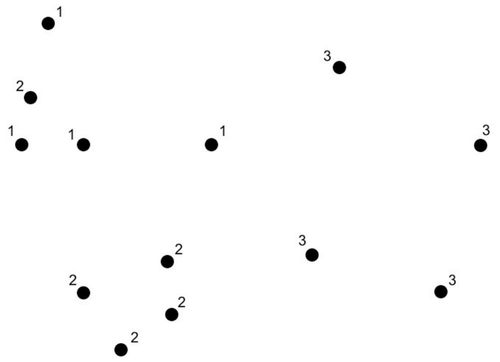
Edo Liberty, Maxim Sviridenko, Ram Sriharsha

K-Means definition



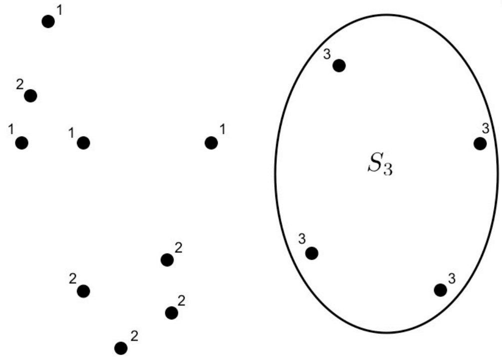
We are given a set of points $v_1, \ldots, v_t, \ldots, v_n$ in Euclidean space.

K-Means definition



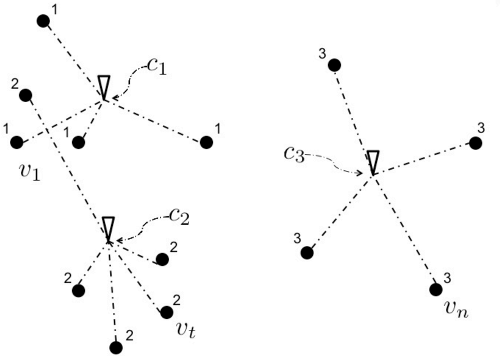
For each point we assign a cluster identifier from the set $\{1, \ldots, k\}$.

K-Means definition



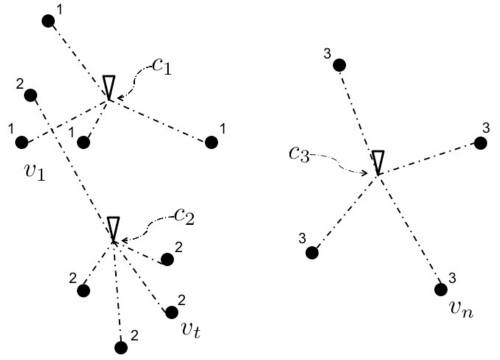
All input points who share the same identifier are called a cluster.

K-Means definition



The assignment cost is the minimal sum of squared distances to cluster centers.

K-Means definition



More accurately $c_i = \frac{1}{|S_i|} \sum_{v \in S_i} v$ and $W = \sum_{i=1}^{k} \sum_{v \in S_i} ||v - c_i||_2^2$.

Prior work (very partial list)

Batch Setting

- Lloyd provides a popular and powerful heuristic [20]
- Ostrovsky, Rabani, Schulman and Swamy prove Lloyds for "well clusterable" inputs [23]
- Arthur and Vassilvitskii, k-means++ provides an expected $O(\log(k))$ approximation [5]
- Kanungo, Mount, Netanyahu, Piatko, Silverman and Wu give a constant approximation ratio with local search [18]
- Bahmani, Moseley, Vattani, Kumar and Vassilvitskii parallelize k-means++ [8]

Streaming setting

- Guha, Meyerson, Mishra, Motwani and O'Callaghan, divide-and-conquer techniques [17].
- Ailon, Jaiswal and Monteleoni [3] build on both [17] and [5].
- Meyerson, Shindler and Wong use techniques similar to facility location [22]

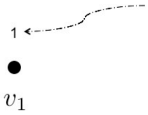
Online

- Charikar, Chekuri, Feder, and Motwani, Online k-centers [9]
- Choromanska and Monteleoni analyze online k-means with experts advise [10]

Online K-Means definition

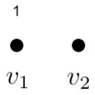
In online k-means we receive one point at a time.

Online K-Means definition



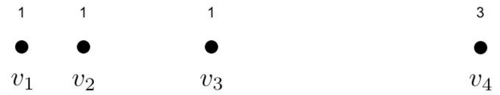
We then immediately assign it a cluster identifier.

Online K-Means definition



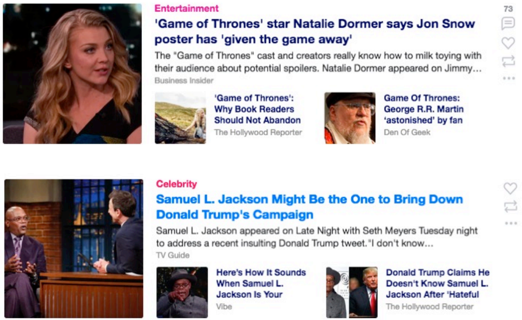
We then receive the next point.

Online K-Means definition



And so on...

Motivation for online k-means

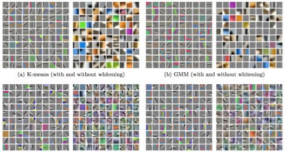


Yahoo show news stories, which are “clusters” of articles. These evolve over time.

Motivation for online k-means

An Analysis of Single-Layer Networks in Unsupervised Feature Learning

Adam Coates   Honglak Lee   Andrew Y. Ng



(a) K-means (with and without whitening)   (b) GMM (with and without whitening)

(c) Sparse Autoencoder (with and without whitening)   (d) Sparse RBM (with and without whitening)

Algorithm	Accuracy (error)
Conv. Neural Network [16]	93.4% (6.6%)
Deep Boltzmann Machine [26]	92.8% (7.2%)
Deep Belief Network [20]	95.0% (5.0%)
(Best result of [11])	94.4% (5.6%)
Deep neural network [27]	**97.13% (2.87%)**
Sparse auto-encoder	96.9% (3.1%)
Sparse RBM	96.2% (3.8%)
K-means (Hard)	96.9% (3.1%)
K-means (Triangle)	97.0% (3.0%)
K-means (Triangle, 4000 features)	**97.21% (2.79%)**

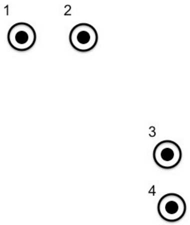
"Surprisingly, we have shown that even the **K-means clustering** algorithm — an extremely simple learning algorithm with no parameters to tune — **is able to achieve state-of-the-art performance** on both CIFAR-10 and NORB datasets when used with the network parameters that we have identified in this work.

Online learning needs online k-means for feature engineering.

Online k-means Algorithm

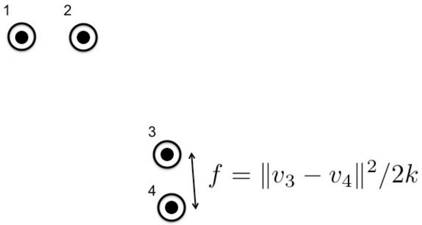
input: V, k
$C \leftarrow$ first $k+1$ distinct vectors in V; and $n = k+1$
(For each of these **yield** itself as its center)
$w^* \leftarrow \min_{v,v' \in C} \|v - v'\|^2/2$
$r \leftarrow 1$; $q_1 \leftarrow 0$; $f_1 = w^*/k$
for $v \in$ the remainder of V **do**
  $n \leftarrow n+1$
  with probability $p = \min(D^2(v, C)/f_r, 1)$
    $C \leftarrow C \cup \{v\}$; $q_r \leftarrow q_r + 1$
  if $q_r \geq 3k(1 + \log(n))$ **then**
    $r \leftarrow r+1$; $q_r \leftarrow 0$; $f_r \leftarrow 2 \cdot f_{r-1}$
  end if
  yield: $c = \arg\min_{c \in C} \|v - c\|^2$
end for

Online k-means algorithm



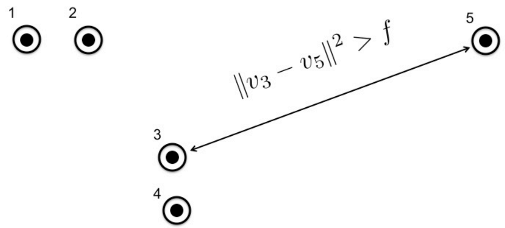
The first $k + 1$ points are assigned to different clusters

Online k-means algorithm



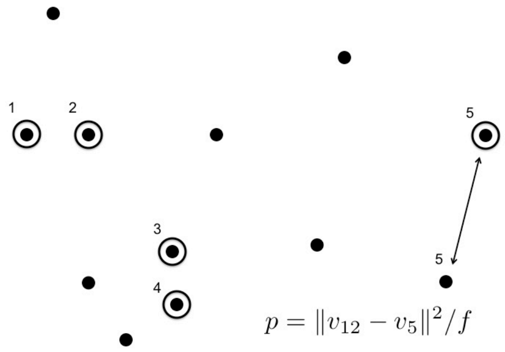
$f \cdot k$ gives a lower bound on the cost of any k-means solution.

Online k-means algorithm



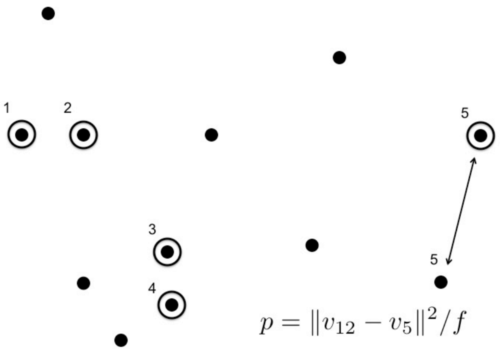
If the cost of assigning a point to an existing cluster is more than f, a new cluster is created

Online k-means algorithm



Otherwise, a new cluster is created with probability p

Online k-means algorithm



Every time $3k(1 + \log(n))$ clusters are added, the value of f is doubled.

Online K-Means immediate observation

We must prove two things about this algorithm

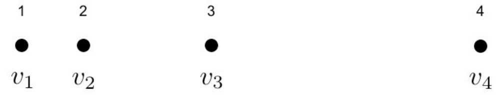
Number of clusters

The algorithm does not create too many clusters.

Cost of clustering

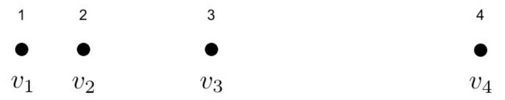
The cost of the clustering is not much worse than optimal

Number of clusters: immediate observation



To be competitive with k-means, online k-means must use **more than k clusters!**

Number of clusters: immediate observation



Let $\gamma = \max_{v,v'} \|v - v'\| / \min_{v,v'} \|v - v'\|$, than $\log(\gamma)$ are needed regardless of k.

Number of clusters

Theorem

Let C be the set of clusters defined by the algorithm. Then

$$\mathbb{E}[|C|] = O\left(k \log n \log \gamma n\right) \ .$$

Where $\gamma = \frac{\max_{v,v'} \|v-v'\|}{\min_{v,v'} \|v-v'\|}$ *is the dataset "aspect ratio".*

Proof idea: there are two phases:

1. **While f is too small:** adding clusters is "too easy". But, f doubles every time $3k(1+\log(n))$ clusters are added.
2. **When f is large enough:** creating new clusters is hard enough such that at most $O\left(k \log n \log \gamma n\right)$ are created in expectation.

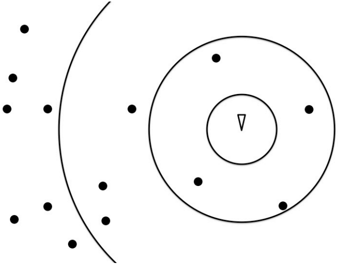
Cost of clustering

Theorem

Let W be the cost of the online assignments of the algorithm and W^* the optimal k-means clustering cost. Then

$$\mathbb{E}[W] = O(W^* \log n) \; .$$

Proof idea: sum expected cost on rings around centers



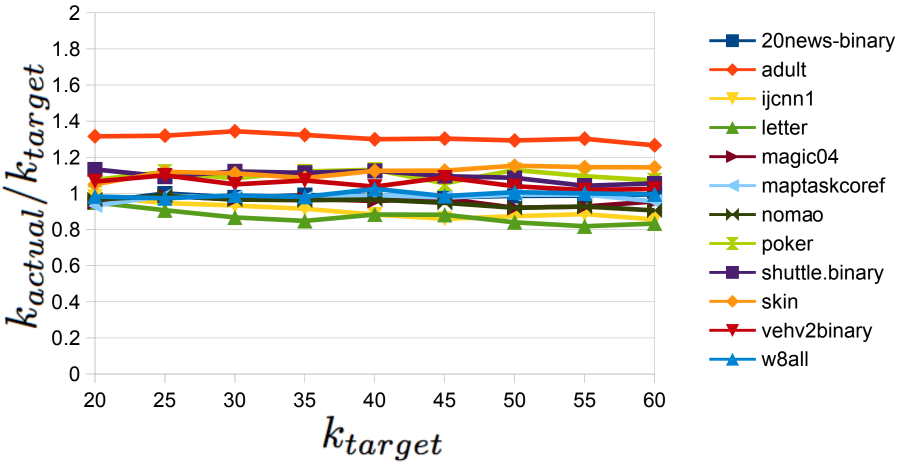
1. After we pick a center from the ring, the cost is at most 16 times optimal.
2. The expected cost before that (or if no center is chosen) is not expected to be high.

Experimental results

Dataset	nnz	n	d	Classification accuracy with raw features	Classification accuracy with k-means features
20news-binary	2.44E+6	1.88E+4	6.12E+4	0.9532	0.9510
adult	5.86E+5	4.88E+4	1.04E+2	0.8527	0.8721
ijcnn1	3.22E+5	2.50E+4	**2.10E+1**	**0.9167**	**0.9405**
letter	2.94E+5	2.00E+4	1.50E+1	0.7581	0.7485
maptaskcoref	6.41E+6	1.59E+5	5.94E+3	0.8894	0.8955
nomao	2.84E+6	3.45E+4	1.73E+2	0.5846	0.5893
poker	8.52E+6	9.47E+5	**9.00E+0**	**0.5436**	**0.6209**
shuttle	2.90E+5	4.35E+4	**8.00E+0**	**0.9247**	**0.9973**
skin	4.84E+5	2.45E+5	**2.00E+0**	**0.9247**	**0.9988**
vehv2binary	1.45E+7	2.99E+5	1.04E+2	0.9666	0.9645
w8all	7.54E+5	5.92E+4	2.99E+2	0.9638	0.9635

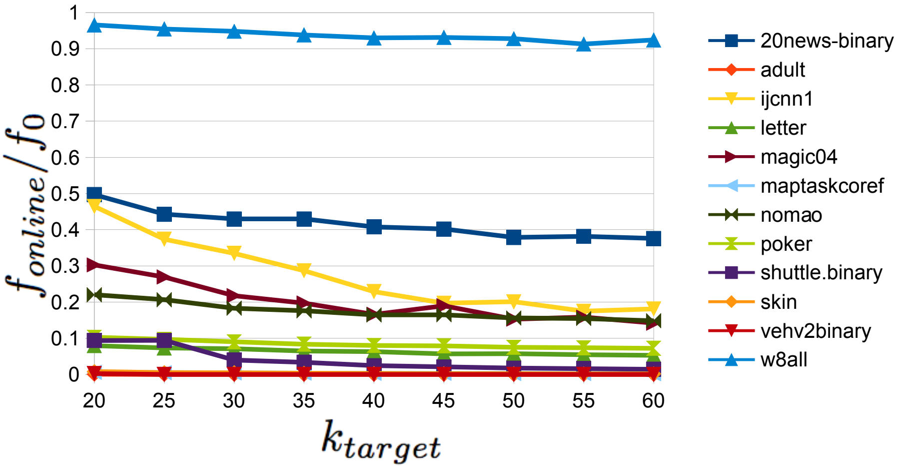
Online k-means gives a boost for online learning, especially in low dimensions.

Online k-means algorithm



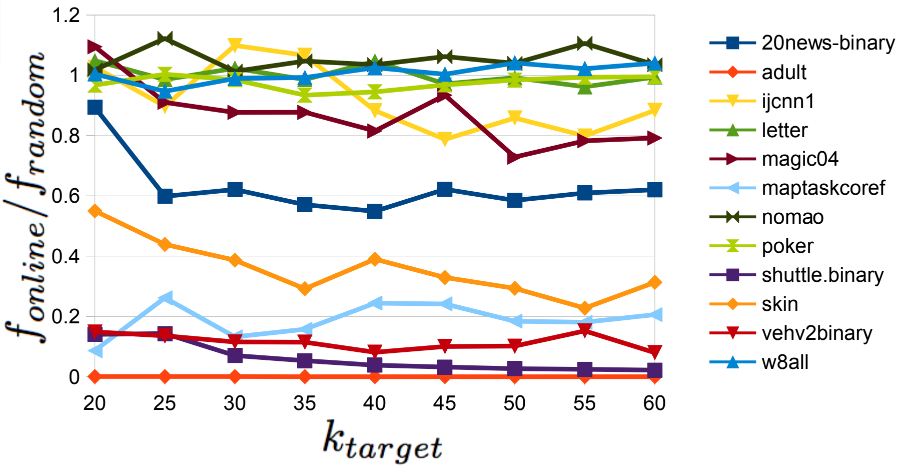
The number of retuned clusters is well concentrated.

Online k-means algorithm



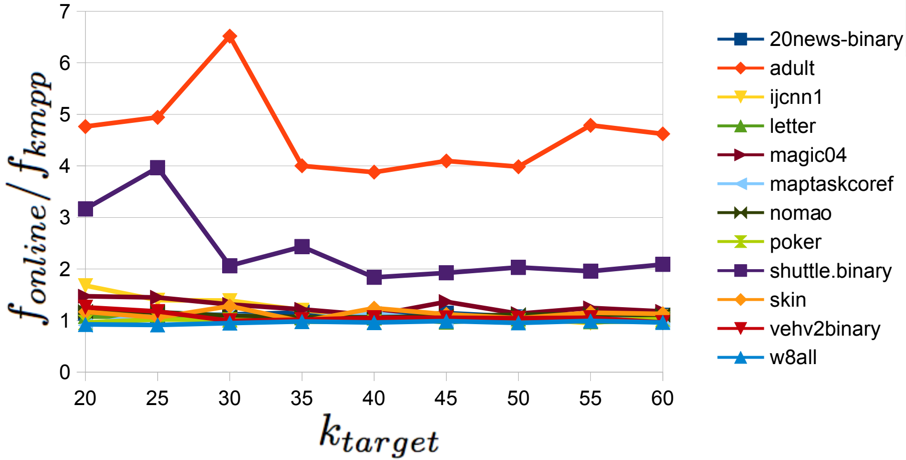
The total error goes reduces with k (as expected)

Online k-means algorithm



Interestingly, uniformly selecting centers improve at the same rate.

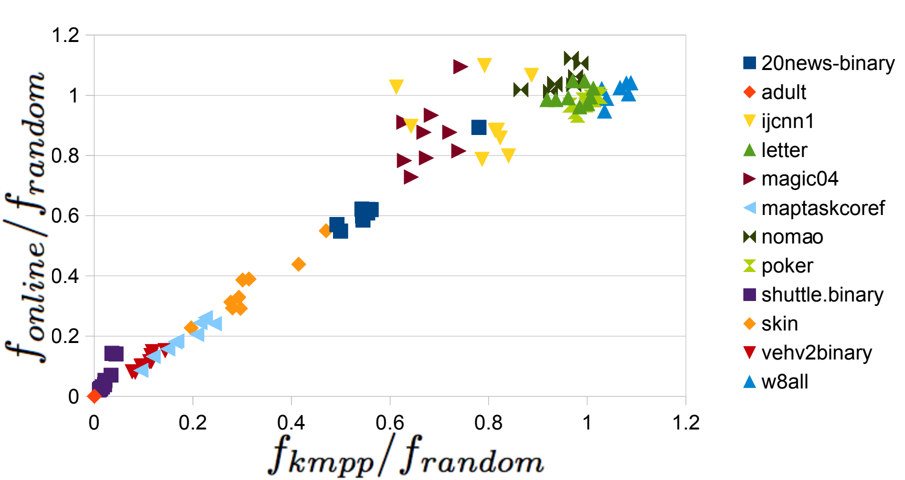
Online k-means algorithm



In comparison to k-means++, this algorithm is consistently worse.

YAHOO!

Online k-means algorithm



Nevertheless, in most scenarios it performs as well (even though it's online!)

Thank you

Marcel R. Ackermann, Marcus Märtens, Christoph Raupach, Kamil Swierkot, Christiane Lammersen, and Christian Sohler.
Streamkm++: A clustering algorithm for data streams.
ACM Journal of Experimental Algorithmics, 17(1), 2012.

Ankit Aggarwal, Amit Deshpande, and Ravi Kannan.
Adaptive sampling for k-means clustering.
In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, pages 15–28, 2009.

Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni.
Streaming k-means approximation.
In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *NIPS*, pages 10–18. Curran Associates, Inc., 2009.

Aris Anagnostopoulos, Russell Bent, Eli Upfal, and Pascal Van Hentenryck.
A simple and deterministic competitive algorithm for online facility location.
Inf. Comput., 194(2):175–202, 2004.

David Arthur and Sergei Vassilvitskii.
k-means++: the advantages of careful seeding.
In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *SODA*, pages 1027–1035. SIAM, 2007.

Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit.
Local search heuristics for k-median and facility location problems.
SIAM J. Comput., 33(3):544–562, 2004.

Kevin Bache and Moshe Lichman.
UCI machine learning repository, 2013.

Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii.
Scalable k-means++.
PVLDB, 5(7):622–633, 2012.

Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani.
Incremental clustering and dynamic information retrieval.
In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 626–635, New York, NY, USA, 1997. ACM.

Anna Choromanska and Claire Monteleoni.
Online clustering with experts.

In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, pages 227–235, 2012.

Adam Coates, Andrew Y. Ng, and Honglak Lee.
An analysis of single-layer networks in unsupervised feature learning.
In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík, editors, *AISTATS*, volume 15 of *JMLR Proceedings*, pages 215–223. JMLR.org, 2011.

Sanjoy Dasgupta.
Topics in unsupervised learning.
Class Notes CSE 291, 2014.

Zvi Drezner and Horst W. Hamacher.
Facility location - applications and theory.
Springer, 2002.

Rong-En Fan.
Libsvm data: Classification, regression, and multi-label., 2014.

Dimitris Fotakis.
On the competitive ratio for online facility location.
Algorithmica, 50(1):1–57, 2008.

Dimitris Fotakis.
Online and incremental algorithms for facility location.
SIGACT News, 42(1):97–131, 2011.

Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan.
Clustering data streams: Theory and practice.
IEEE Trans. Knowl. Data Eng., 15(3):515–528, 2003.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu.
A local search approximation algorithm for k-means clustering.
In *Symposium on Computational Geometry*, pages 10–18, 2002.

Percy Liang and Dan Klein.
Online EM for unsupervised models.
In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pages 611–619, 2009.

Stuart P. Lloyd.
Least squares quantization in pcm.
IEEE Trans. Inf. Theor., 28(2):129–137, September 1982.

Adam Meyerson.
Online facility location.
In *FOCS*, pages 426–431. IEEE Computer Society, 2001.

Adam Meyerson, Michael Shindler, and Alex Wong.
Fast and accurate k-means for large datasets.
NIPS, 2011.

Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy.
The effectiveness of lloyd-type methods for the k-means problem.
J. ACM, 59(6):28, 2012.

Jens Vygen.
Approximation algorithms for facility location problems.
Lecture Notes, Technical Report No. 05950, 2005.