Convergence of online k-means

Sanjoy Dasgupta Gaurav Mahajan Geelon So

Dept. of Computer Science and Engineering, UC San Diego

Background

Lloyd's method (Lloyd, 1982) is an algorithm to k-means cluster a dataset. At each step, the algorithm proposes k centers, say $W_1, \ldots, W_k \in \mathbb{R}^d$. Each data point is mapped to its closest center, partitioning the dataset into k clusters. The update sets each center to the mean of its cluster data.

Bottou and Bengio (1995) define the **online Lloyd's** method where the data come in a stream. At each step, the center W_i is set to the mean of all previous data points that had mapped to the ith center: it maintains a counter N_i and if W_i is the center closest to the next data point X, the update is:

$$N_i \leftarrow N_i + 1 \quad \text{and} \quad W_i \leftarrow W_i - \frac{1}{N_i} \cdot \Big(W_i - X\Big).$$

Motivating question: the convergence properties of algorithms such as online Lloyd's is unknown. What is the behavior of *online k-means algorithms* given a never-ending stream of data from an underlying data distribution?

The online k-means algorithm

Algorithm *online k-means*
Initialize: k arbitrary distinct centers $W = (W_1, \ldots, W_k) \in \mathbb{R}^{k \times d}$

1. **for** iteration $n = 0, 1, 2, \ldots$
2. **do** sample data point $X \sim p$
3. identify closest center $i \leftarrow \arg\min_{j \in [k]} \|W_j - X\|$
4. update closest center $W_i \leftarrow W_i - H_i \cdot \Big(W_i - X\Big)$

Theorem

Let p have bounded support on $\mathbb{R}^d$ and f its k-means cost function. Assume that the set of stationary points $\{\nabla f = 0\}$ is topologically nice. If the sequence of (stochastic) learning rates $(H_i^{(n)})_{n=0}^{\infty}$ satisfies some mild and reasonable conditions, then the iterates $W^{(n)}$ asymptotically converge:

$$\limsup_{n \to \infty} \inf_{\nabla f(w) = 0} \|W^{(n)} - w\| = 0 \quad \text{a.s.}$$

In fact, a well-conditioned variant of online Lloyd's asymptotically converges.

A step of online k-means

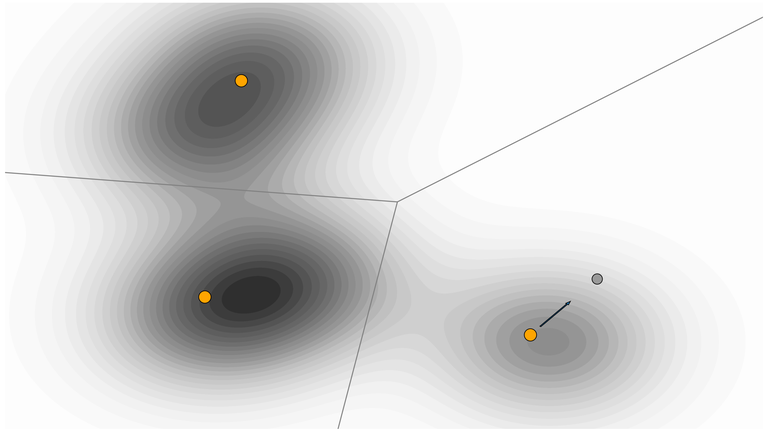

Figure 1. Three centers, the orange dots, partition $\mathbb{R}^2$. A random data point X, the gray dot, lands in the lower right Voronoi cell. An online k-means update will nudge that center W_i toward X. In expectation, the center W_i will move toward the mean of its Voronoi cell.

The k-means cost function

Let $w = (w_1, \ldots, w_k) \in \mathbb{R}^{k \times d}$ be a k-tuple of centers. Define $V_i(w)$ be the Voronoi cell corresponding to points with w_i as the closest center.

$$f(w) = \frac{1}{2} \sum_{i \in [k]} \int_{V_i(w)} \|w_i - x\|^2 p(x)\, dx.$$

Connection to gradient descent

The gradient of f with respect to the ith center turns out to be:

$$\begin{aligned} \nabla_{w_i} f(w) &= \int_{V_i(w)} (w_i - x)\, p(x)\, dx \\ &= P_i(w) \cdot \Big(w_i - M_i(w)\Big), \end{aligned}$$

where $P_i(w)$ and $M_i(w)$ are the mass and the mean of the ith Voronoi cell. Notice that a gradient descent step pushes w_i toward $M_i(w)$.

Our contributions

Computation of the gradient

Computing the derivative of the k-means cost function with respect to w is not straightforward because the both the domain of integration and the integrand depend on w. Surprisingly, the shifting boundaries of the Voronoi cells do not contribute to the gradient: points that jump between two cells under small perturbations must be fairly equidistant to either centers. They contribute nothing to the first-order change in the k-means cost, which is what the gradient computes.

Convergence with non-uniform and stochastic learning rates

Standard techniques from optimization can analyze SGD with uniform learning rates, but are insufficient for the variant performed by online k-means, which has center-specific learning rates. We extend the techniques from Bertsekas and Tsitsiklis (2000) to cover non-uniform learning rates.

The key property that we shall require for convergence is that if a center W_i is far from the mean of its Voronoi cell, then with constant probability, it is updated at a rate not too much slower than the rest of the centers.

Online Lloyd's algorithm

To design an online version of Lloyd's algorithm with asymptotic guarantees, we start from the interpretation of Lloyd's algorithm as preconditioned gradient descent. Then, we define an online Lloyd's algorithm as its stochastic analog, which concurrently keeps an estimate of the preconditioner. We prove the consistency of our estimator to the Lloyd preconditioner, lending our algorithm the interpretation of a natural extension of Lloyd's.

References

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 2000.

Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems*, 1995.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982.

The 25th International Conference on Artificial Intelligence and Statistics