Week 9 Tutorial on K-means++

Sarat Moka

K-means++ is a simple yet powerful randomized seeding scheme for Lloyd's K-means approach. It picks the first center uniformly at random and then iteratively samples each new center with probability proportional to the squared distance to its nearest existing center. In expectation this initialization yields an $O(\log K)$-approximation to the optimal clustering, at only an $O(nK)$ overall cost—almost the same as plain random init but with dramatically improved solution quality.

References:

- The standard K-means algorithm was proposed by Stuart Lloyd (Bell Labs) in 1957 and was published as a journal article in 1982: S. Lloyd (1982), "Least squares quantization in PCM," in IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137
- Arthur, D. and Vassilvitskii, S. (2007). "K-means++: The advantages of careful seeding." In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1027–1035.

1 Problem setup

Let $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a set of n data points, and let $K \in \mathbb{N}$ be the desired number of clusters. In classical K-means we wish to find a set of centers

$$\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\} \subset \mathbb{R}^d$$

that approximately minimize the within-cluster sum of squared distances (the *inertia*):

$$J(\mathcal{C}) \;=\; \sum_{i=1}^{n} \min_{1 \le j \le K} \|\mathbf{x}_i - \mathbf{c}_j\|^2.$$

The standard algorithm alternates between assigning each x_i to its closest center and recomputing each μ_j as the mean of its assigned points. Its final solution, however, can be very sensitive to the choice of the initial centers.

2 K-means++ initialization

K-means++ is a simple and effective randomized seeding procedure that boosts the chance of finding a good local optimum. We summarize it in Algorithm 1.

Algorithm 1 K-means++ Initialization

Require: Data set $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, number of clusters K.
Ensure: Initial centers $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$.

$\mathcal{C} \leftarrow \emptyset$
Choose $\mathbf{c}_1$ uniformly at random from D
$\mathcal{C} \leftarrow \{\mathbf{c}_1\}$
for $j = 2$ to K **do**
  for all $\mathbf{x} \in D$ **do**
    $d_{\mathbf{x}} \leftarrow \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{x} - \mathbf{c}\|^2$
  end for
  Draw $\mathbf{c}_j$ from D with probability

$$\mathbb{P}(\mathbf{x} \text{ is chosen}) \;=\; \frac{d_{\mathbf{x}}}{\sum_{\mathbf{x}' \in D} d_{\mathbf{x}'}}\,.$$

  $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{c}_j\}$
end for
return $\mathcal{C}$

After obtaining $\mathcal{C}$, one simply runs the standard K-means (Lloyd's) iterations until convergence.