K-Means++ Initialization

Improved initialization strategy for K-Means

Algorithm Steps

1. **First centroid:** Choose randomly from data points
2. **Subsequent centroids:** Probability proportional to squared distance
3. **Result:** Spreads initial centroids far apart in data space

Selection Probability

$$P(x) \propto D(x)^2$$

D(x) = distance to nearest centroid

⚡ Faster Convergence

Significantly reduces iterations to convergence

✨ Better Quality

Large gains in clustering quality

🎯 Provable Guarantee

O(log k) approximation to optimal solution

🔧 Widely Adopted

Default in scikit-learn and other libraries

⚖️ Trade-off Analysis

Small overhead in initialization → Large gains in quality and speed