一、 mini batch kmeans 算法简介

在当前大数据的背景下，工程师们往往为了追求更短的计算时间，不得不在一定程度上减少算法本身的计算精度，我说的是在一定程度上，所以肯定不能只追求速度而不顾其它。在 KMeans 聚类中，为了降低计算时间，KMeans 算法的变种 Mini Batch KMeans 算法应运而生。

Mini Batch KMeans 算法是一种能尽量保持聚类准确性下但能大幅度降低计算时间的聚类模型，采用小批量的数据子集减少计算时间，同时仍试图优化目标函数，这里所谓的 Mini Batch 是指每次训练算法时随机抽取的数据子集，采用这些随机选取的数据进行训练，大大的减少了计算的时间，减少的 KMeans 算法的收敛时间，但要比标准算法略差一点，建议当样本量大于一万做聚类时，就需要考虑选用 Mini Batch KMeans 算法。

这种对时间优化的思路不仅应用在 KMeans 聚类，还广泛应用于梯度下降、深度网络等机器学习和深度学习算法。

二、 mini batch kmeans 算法原理

在普通的 K Means 的计算过程中，每次更新各聚类中心点时，需要计算所有点和每个聚类中心点的距离，所以代价特别昂贵。

而在 mini-batch K Means 的计算过程中，每次更新各聚类中心点时，先从所有数据中随机地选取一个小集合（也就是这里的 mini-batch），根据这个

集合中的数据点，来更新各聚类的中心点。下一次更新时，再重新从所有数据点中选取一个随机的小集合，如此重复，直到达到收敛条件。

mini-batch 的思想就是用部分数据，而不是全部数据，来更新模型的参数。那么这 batch size 个样本怎么来的？一般是通过无放回的随机采样得到的。为了增加算法的准确性，我们一般会多跑几次 Mini Batch K-Means 算法，用得到不同的随机采样集来得到聚类簇，选择其中最优的聚类簇。

三、 mini batch kmeans 算法流程

1. 首先抽取部分数据集,使用 K- Means 算法构建出 K 个聚簇点的模型。
2. 继续抽取训练数据集中的部分数据集样本数据,并将其添加到模型中,分配给距离最近的聚簇中心点。
3. 更新聚簇的中心点值。
4. 循环迭代第二步和第三步操作,直到中心点稳定或者达到迭代次数,停止计算操作。