spark2.3.1

ml、mllib上都有聚类算法

ML库中包含4种：

1.K-means——[K-means](http://spark.apache.org/docs/latest/ml-clustering.html#k-means)

* 2.LDA——[Latent Dirichlet allocation (LDA)](http://spark.apache.org/docs/latest/ml-clustering.html#latent-dirichlet-allocation-lda)
* 3.二分K-menas—— [Bisecting k-means](http://spark.apache.org/docs/latest/ml-clustering.html#bisecting-k-means)
* 4.GMM（高斯混合模型）——[Gaussian Mixture Model (GMM)](http://spark.apache.org/docs/latest/ml-clustering.html#gaussian-mixture-model-gmm)

MLlib上有6种（基于RDD）：

K-means,二分Kmeans，GMM，PIC，LDA，流式KMeans

* [K-means](http://spark.apache.org/docs/latest/mllib-clustering.html#k-means)
* [Gaussian mixture](http://spark.apache.org/docs/latest/mllib-clustering.html#gaussian-mixture)
* [Power iteration clustering (PIC)](http://spark.apache.org/docs/latest/mllib-clustering.html#power-iteration-clustering-pic)
* [Latent Dirichlet allocation (LDA)](http://spark.apache.org/docs/latest/mllib-clustering.html#latent-dirichlet-allocation-lda)
* [Bisecting k-means](http://spark.apache.org/docs/latest/mllib-clustering.html#bisecting-k-means)
* [Streaming k-means](http://spark.apache.org/docs/latest/mllib-clustering.html#streaming-k-means)

GMM:聚类得到的K类，他们都符合Gaussian分布，但属于参数不同的Gaussian分布。即N（u,Σ）不同：

LDA:文档主题模型，一般用于文档聚类

AdaBoost算法：

实际上，就是根据N个弱的“基分类器”，生成1个强的“最终分类器”

对于N个基分类器:G1,G2,…,Gn

分类器定义为： Gi: Gi(Xi) -> yi ,其中yi是{-1，1} ，即分类值（也可以看做是符号函数sign）

给1个基分类器，定义一个初始权重W={w11,w12,w13 ….w1i …}

让最终生成的“强可学习的分类器”形如：

G(last) —> sign{ Σ（wi\*Gm） }

K次学习，得导K个弱分类器，作“线性组合”：其中基分类器的输出yj = Gm(xj) 即 输入xi， 使用Gm分类器，得到yi， yi就是符号函数

然后,adaBoost的学习过程，就是由初始权重W， 迭代更新，生成“更准的分类器权重放大，不准的分类器权重缩小”的这样一个第k次得到的权重Wk

最终，使误分类=0或者误分类个数<阈值时

**重点： 理解这个权重的迭代过程。**

**注意：刚开始，只有1个基分类器，每1轮由上1个基分类器，学习得到下1个基分类器（并对应权值Wm）,最后将K次更新的基分类器，按权值，线性组合到一起。**