机器学习—聚类系列-层次聚类（Hierarchical Clustering）

blog.csdn.net/xiyujianxia/article/details/80369407

引入Hierarchical Clustering背景：k-means算法却是一种方便好用的聚类算法，但是始终有K值选择和初始聚类中心点选择的问题，而这些问题也会影响聚类的效果。为了避免这些问题，可以选择另外一种比较实用的聚类算法-层次聚类算法。

1、Hierarchical Clustering的作用及类别

1.1Hierarchical Clustering：一如其字面意思，是层次化的聚类，得出来的是树形结构（计算机科学的树是一棵根在最上的树），HC不需要指定具体类别数目的，其得到的是一颗树，聚类完成之后，可在任意层次横切一刀，得到指定数目的 cluster，具体有两种方式一种是在下而上凝聚方法（agglomerative：先将所有样本的每个点都看成一个簇，然后找出距离最小的两个簇进行合并，不断重复到预期簇或者其他终止条件），另一种自上而下分裂方法（divisive：先将所有样本当作一整个簇，然后找出簇中距离最远的两个簇进行分裂，不断重复到预期簇或者其他终止条件）。

2、Agglomerative HC运行的4个步骤详解

2.1、把每个样本归为一类，计算每两个类之间的距离，也就是样本与样本之间的相似度；

2.2、寻找各个类之间最近的两个类，把他们归为一类（这样类的总数就少了一个）.

2.3、重新计算新生成的这个类与各个旧类之间的相似度.

2.4、重复2和3直到所有样本点都归为一类，结束。

3、Dendrogram工作机制解密：

3.1 计算两个点之间的距离 ，欧式距离（平面几何）。

3.2 计算两个簇之间的距离（Distance between two Cluster），有下面四种方式：

1.Closest Points ：两个簇最近点的距离

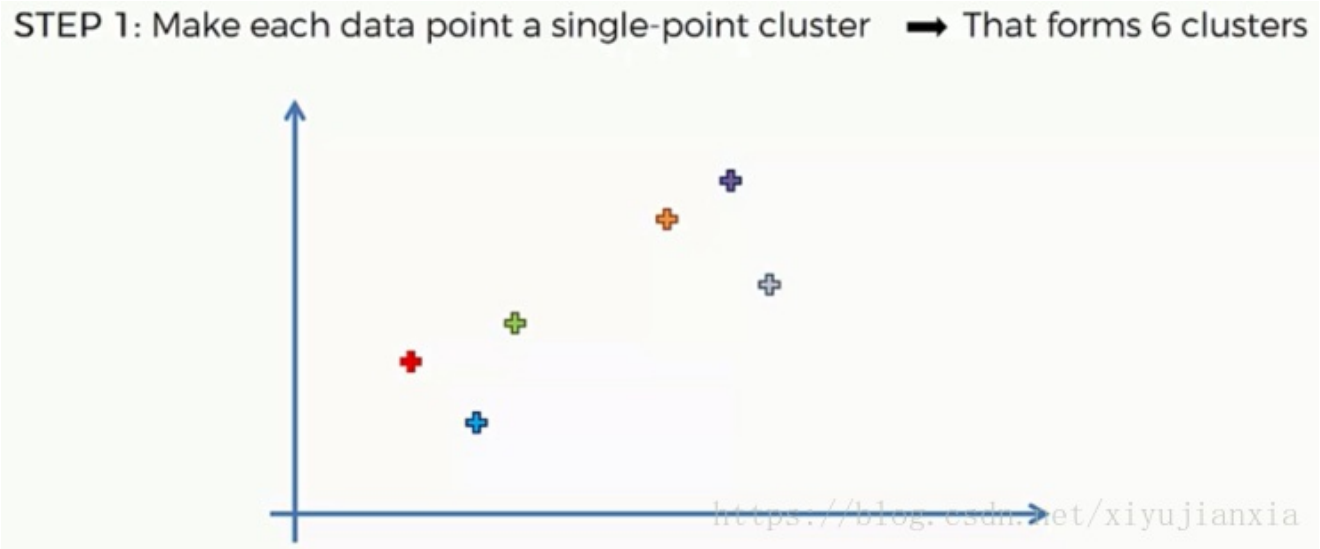
2.Furthest Points ：两个簇最远的点的距离

3.Average Points ：两个簇所有点两两距离平均值。

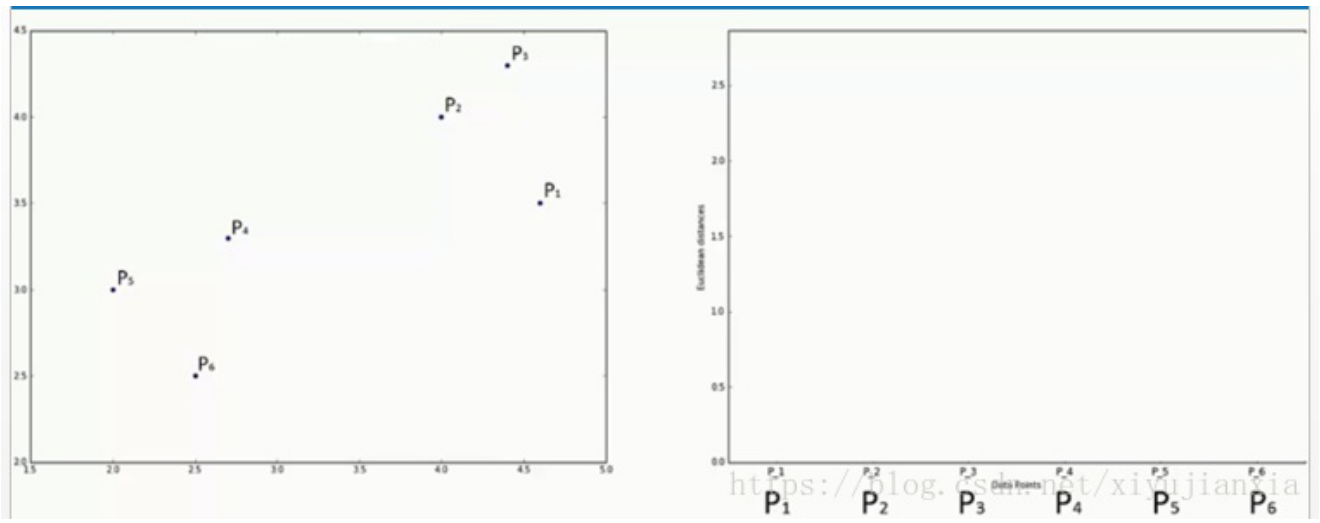
3.Distance Between Centroids 两个簇 簇中心点之间的距离

4、 Hierarchical Clustering在内部是如何使用Dendrogram进行Clustering的？

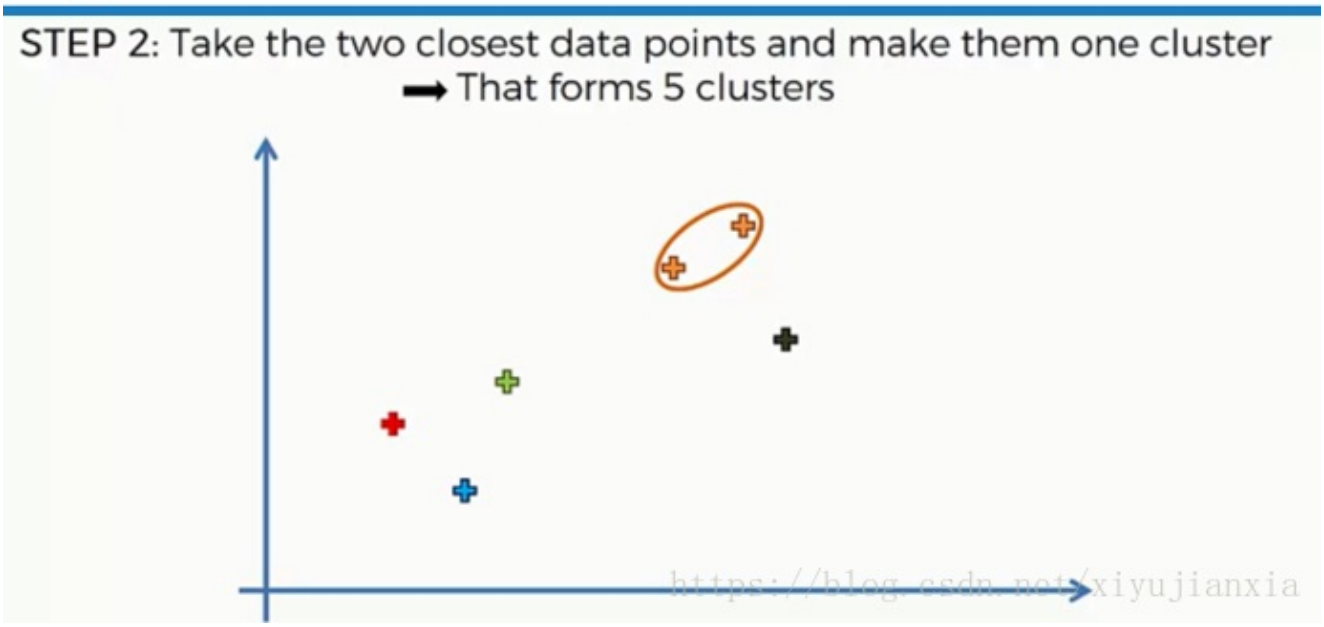
4.1把每一个点当作一个簇（组），例子是6个点



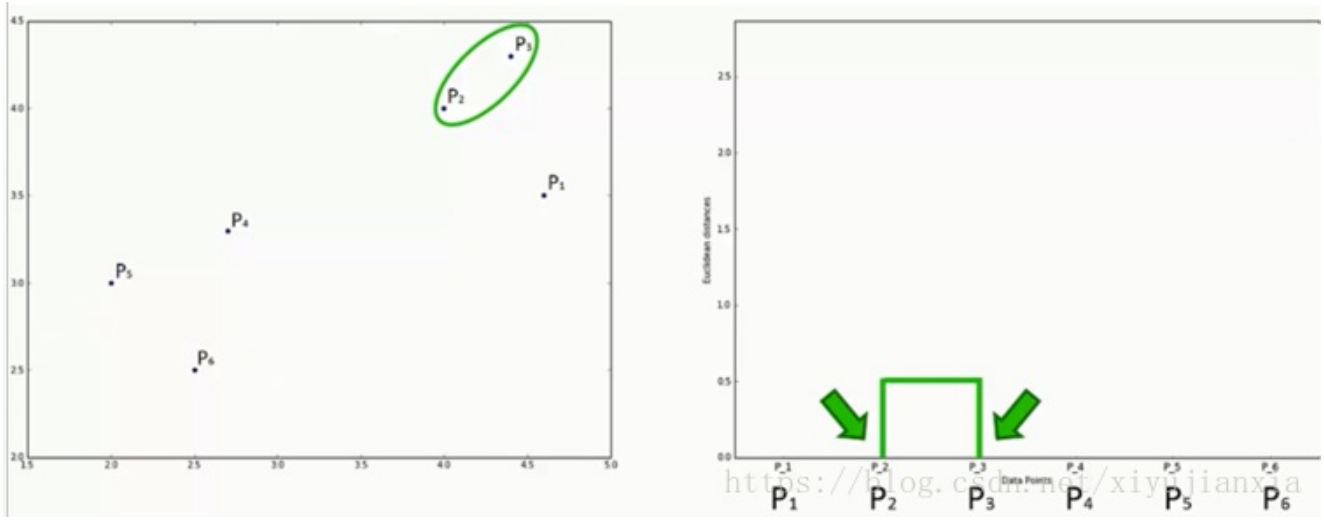
工作过程：开始在P1-P6共6个点



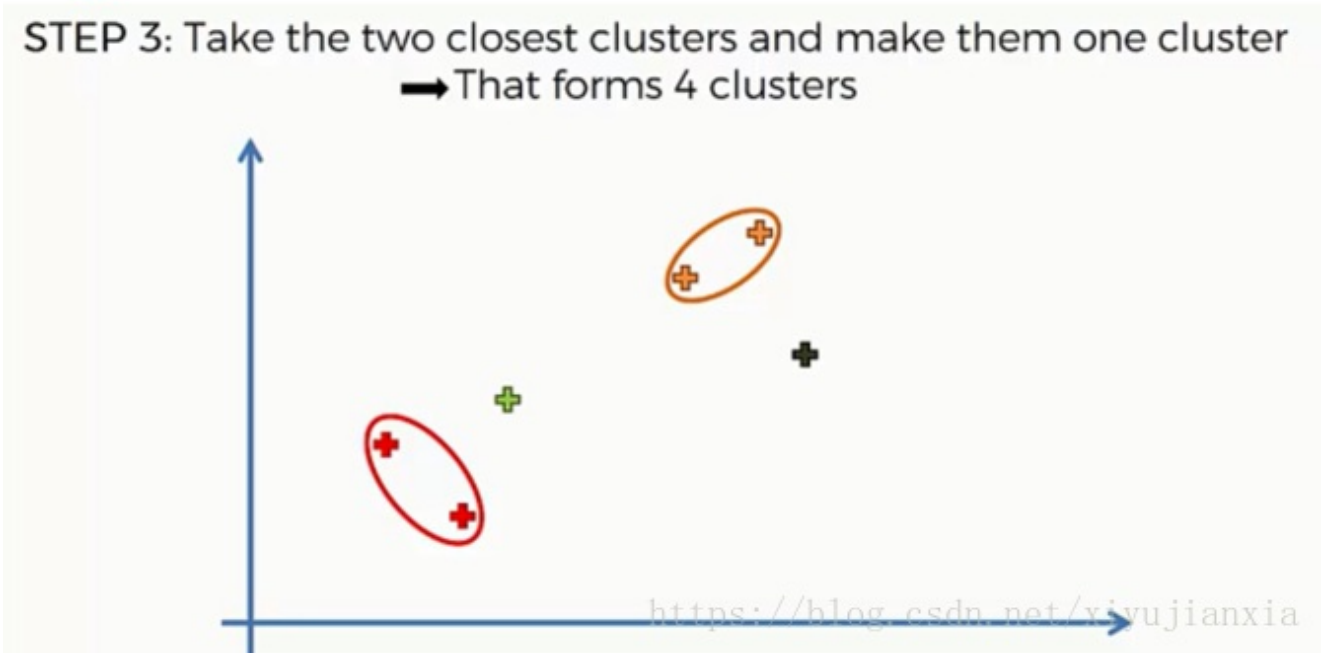
4.2 找到两个最近的点合并为一簇（组），就变成5簇

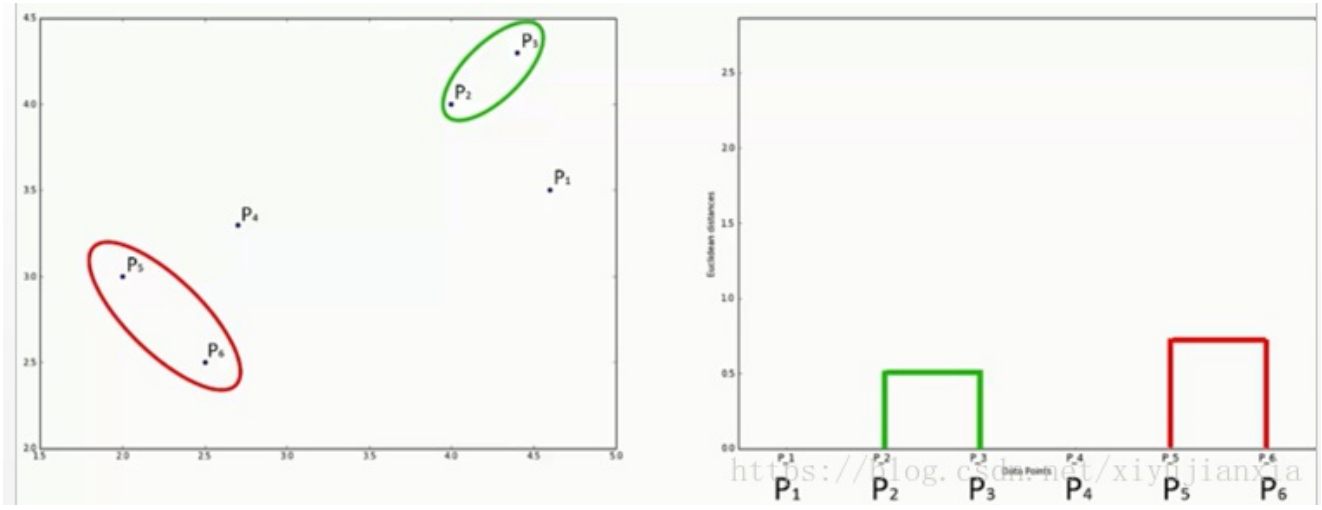


合并最近的两个点，绘制一个矩形，高度代表两个簇之间的距离

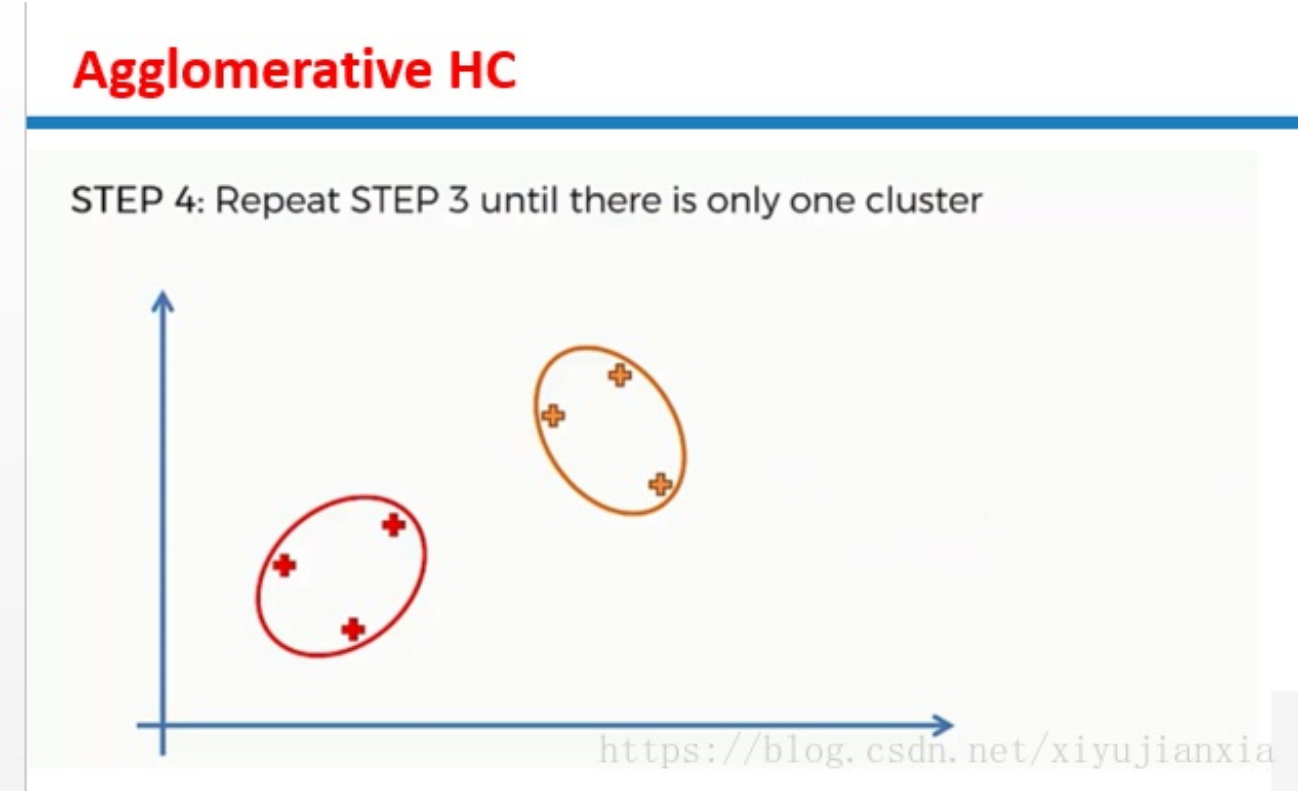


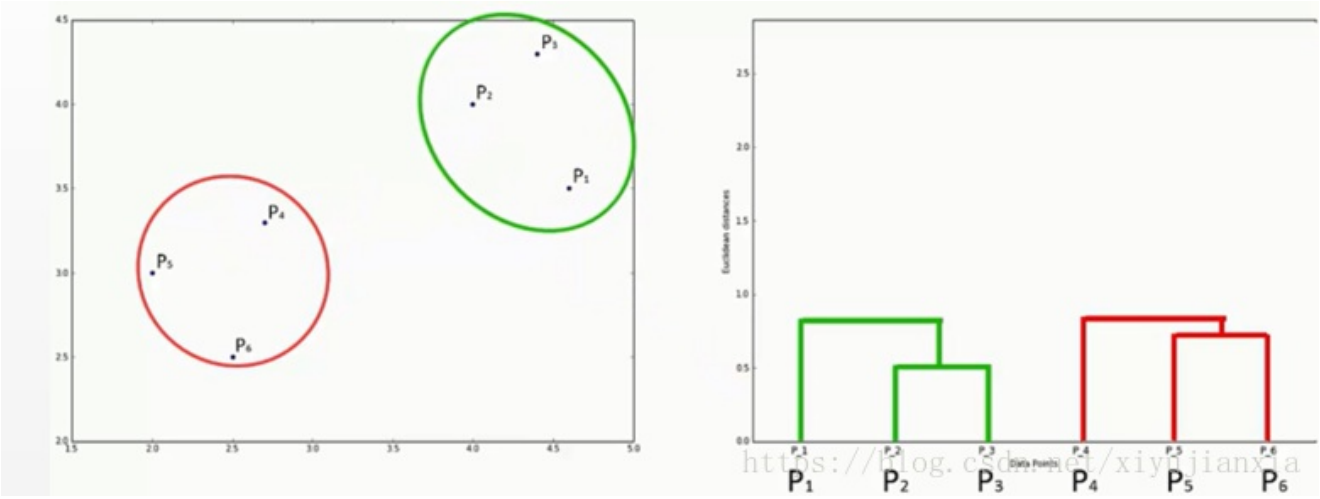
4.3 继续上面的步骤，找到两个最近的点合并为一簇（组），就变成4簇：



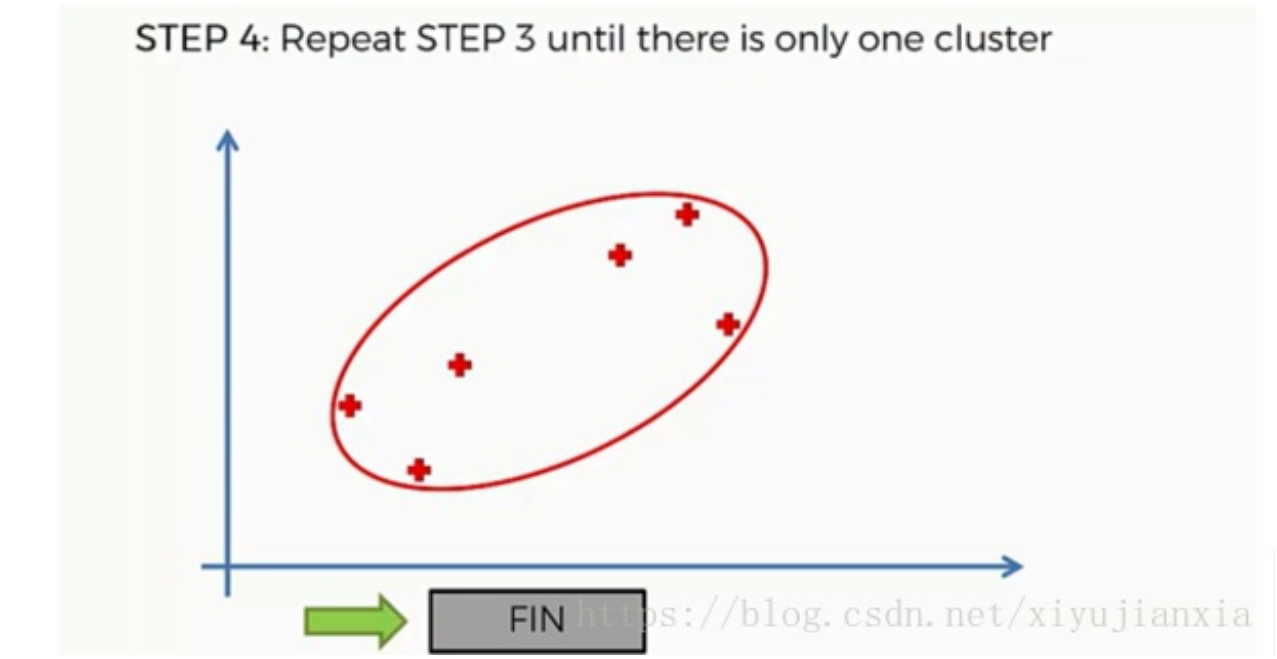


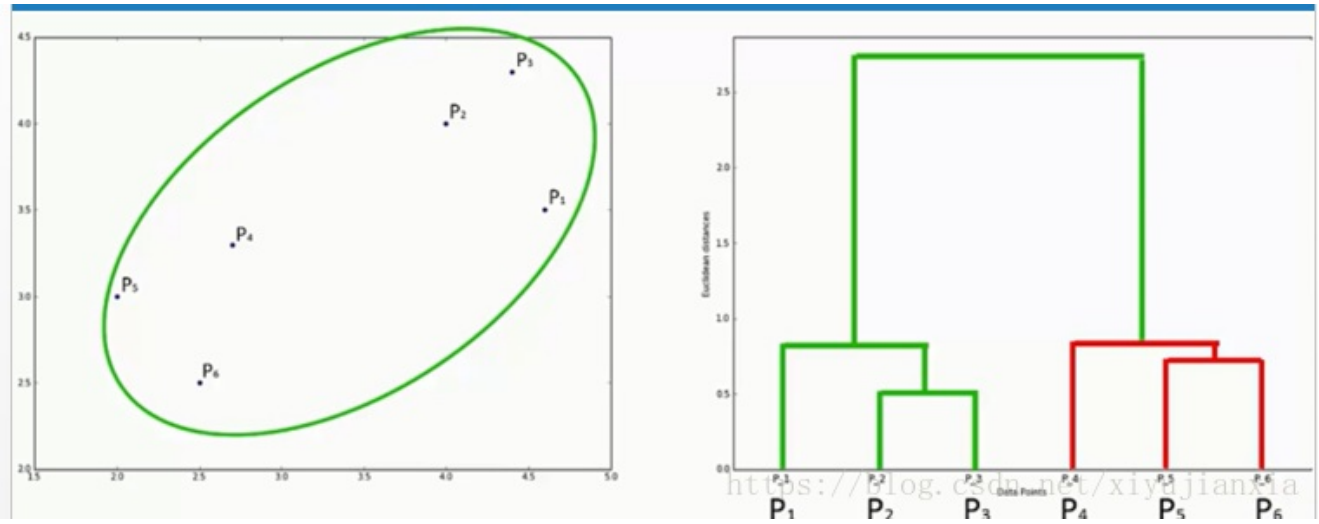
4.4 继续上面的步骤，找到两个最近的点合并为一簇（组），就变成2簇：





4.5 继续上面的步骤，直到变成一个簇（组），结束：





5、寻求最优的Clusters的个数

采用什么标准来决定Clusters的个数？ 最大距离，横向切过一条线，所有关联的的簇（组）。

Optimal numbers of clusters