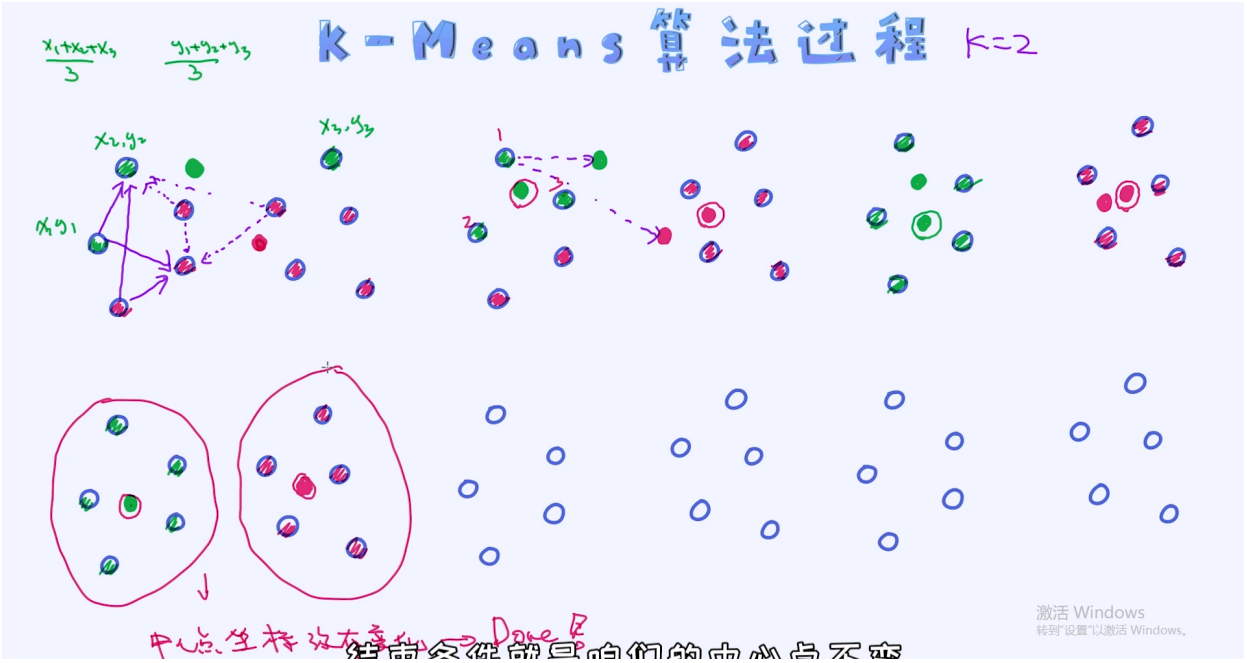
无监督学习算法输入的样本是不带标签的

聚类算法

k means算法是一种聚类算法也是常用的无监督学习算法



k-Means 算法

循环迭代式的算法

初始化：

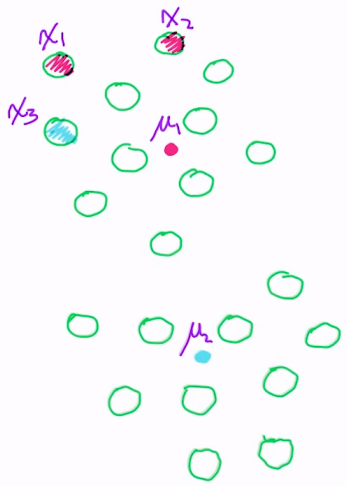
随机选择k个点，作为初始中心点，每个点代表一个group

交替更新：

计算每个点到所有中心点的距离，把最近的距离记录下来并把对应的group赋给当前的点

针对于每一个group里的点，计算其平均并作为这个group的新的中心点。

k-Means的目标函数



$$D = \{x_1, x_2 \cdots x_N\}$$

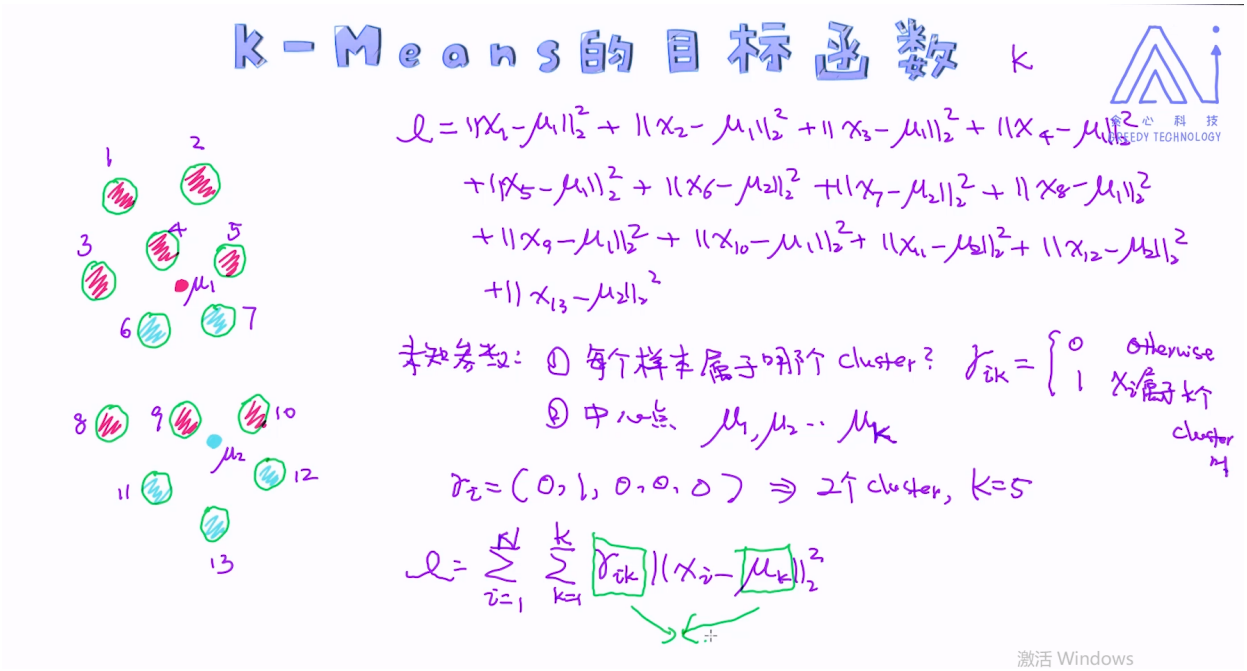
$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{2} \|x_i - \mu_j\|_2^2 \leftarrow \text{Objective Function.}$$

$$\downarrow$$

$$\|x_1 - \mu_1\|_2^2 + \|x_1 - \mu_2\|_2^2 + \|x_2 - \mu_1\|_2^2 + \|x_2 - \mu_2\|_2^2$$

$$+ \cdots$$

$$\|x_1 - \mu_1\|_2^2 + \|x_2 - \mu_1\|_2^2 + \|x_3 - \mu_2\|_2^2 + \cdots$$



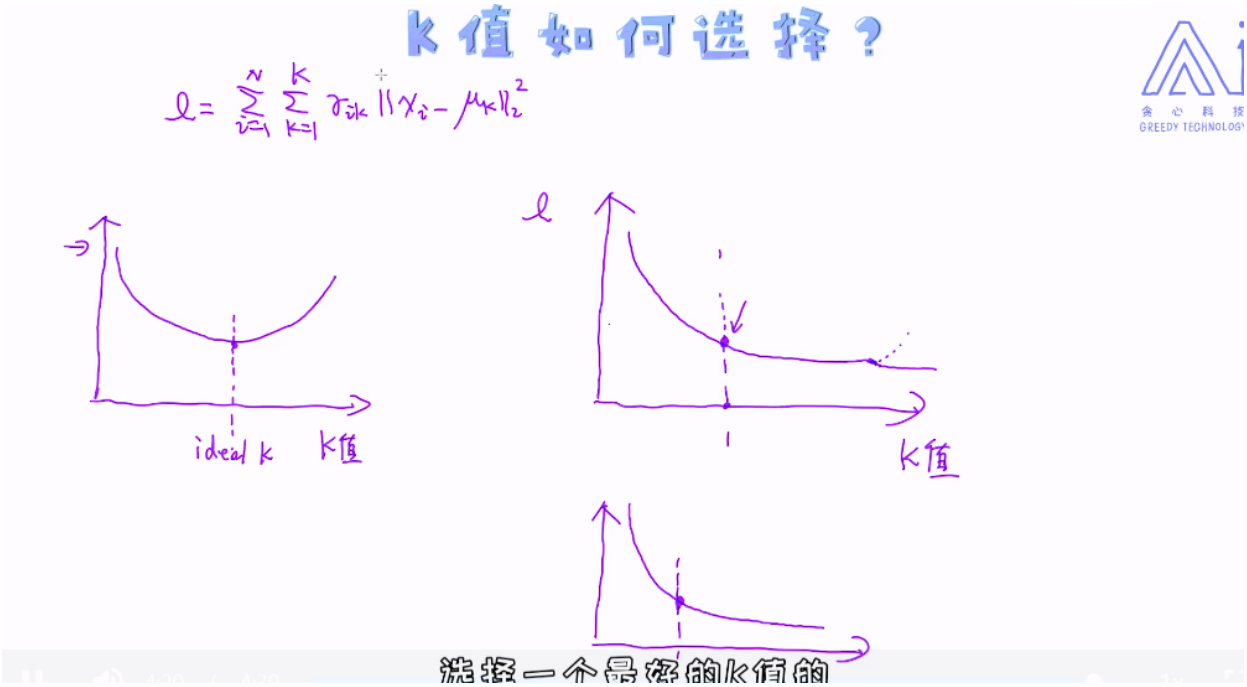
对于k-means的目标函数我们通常使用交替优化的方法。前面已经说过，目标函数里有两组不同的参数。我们可以分别对其中的每一组参数优化，但这时候相当于把另外一组参数当做是已知的就可以了。

k-means算法一定会收敛，因为圈1和圈2过程都会使目标函数下降。

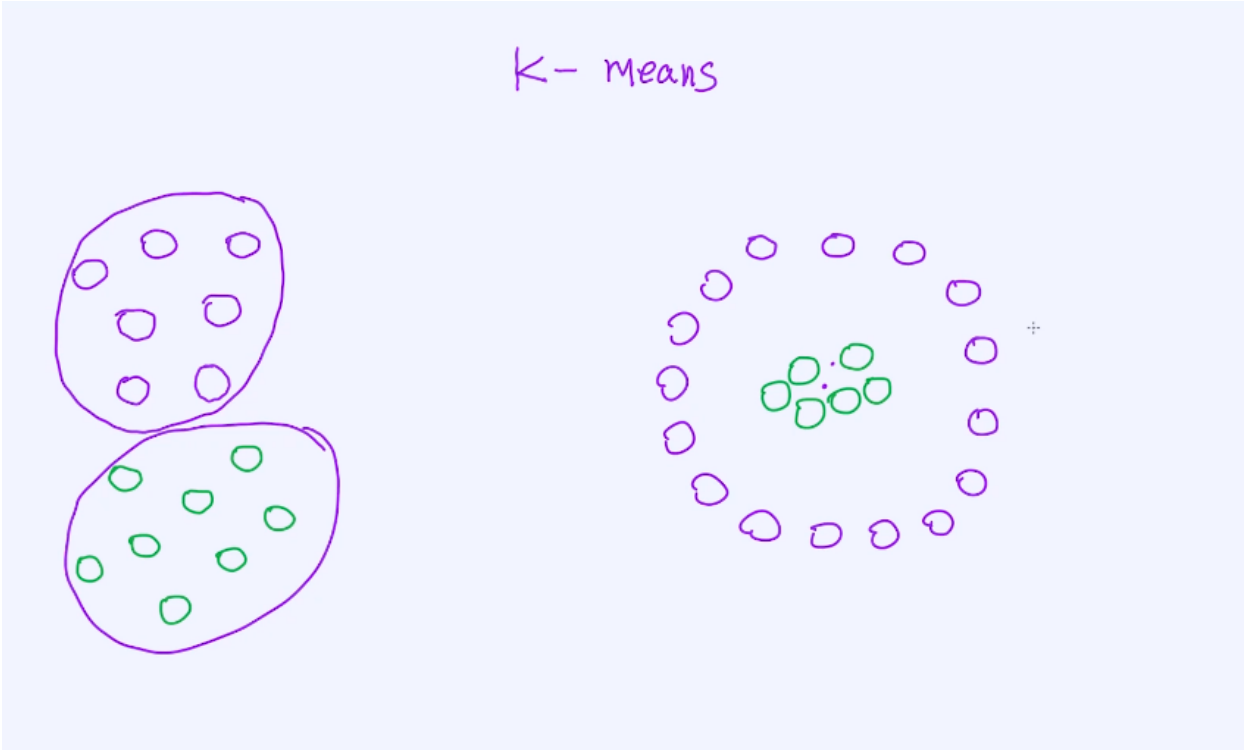
k-means算法不同初始化的值会有不同的结果，其核心是非凸函数。 如果一个目标函数是非凸函数，那我们其实不能保证或者没有办法得到全局最优解的！

没有标签无法用交叉验证

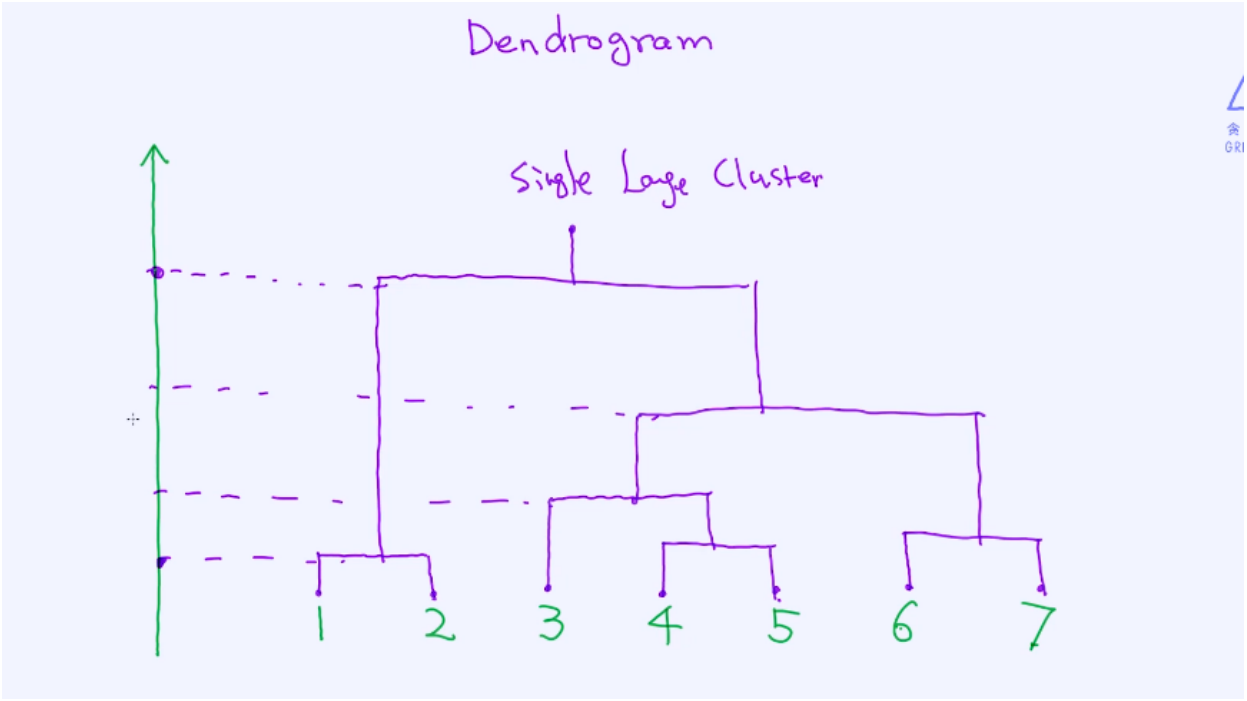
选择拐点为k值

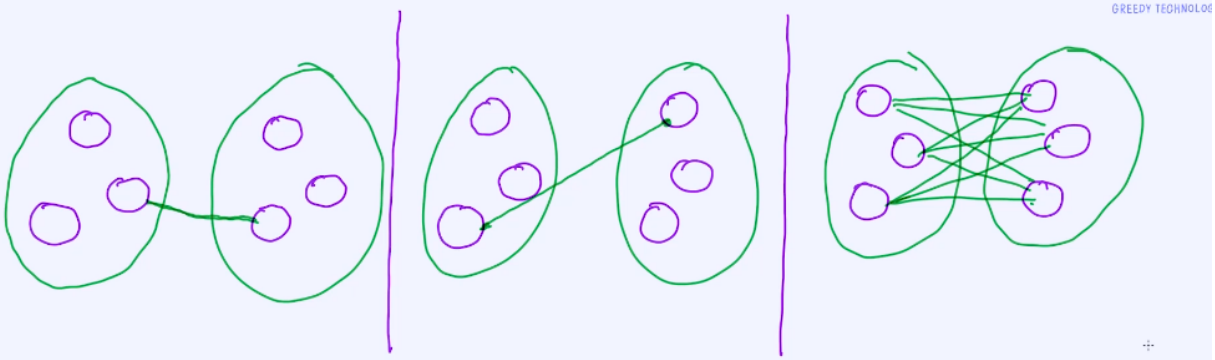


层次聚类不用选择k值，kmeans算法不能捕捉样本层次关系



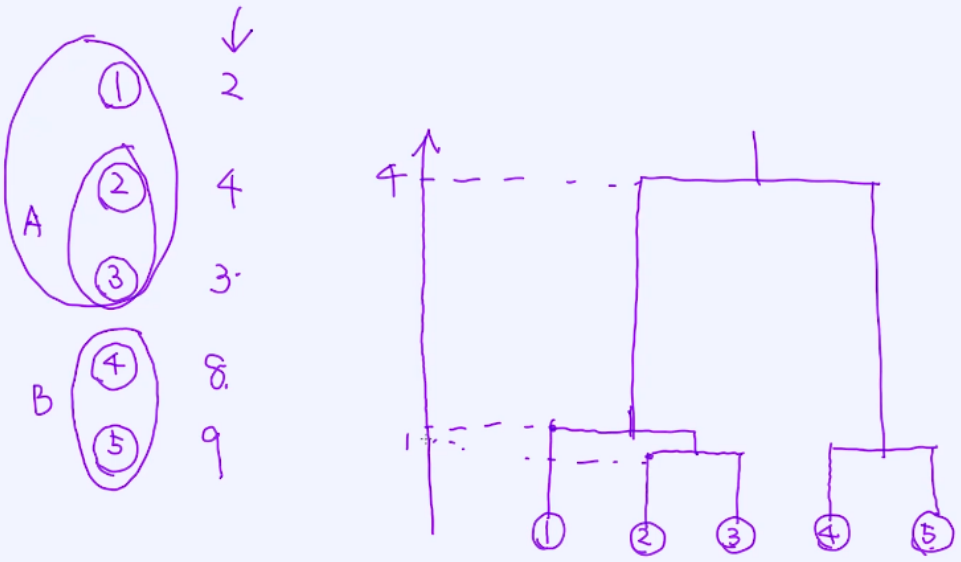
合并越顶层cluster之间的距离越大

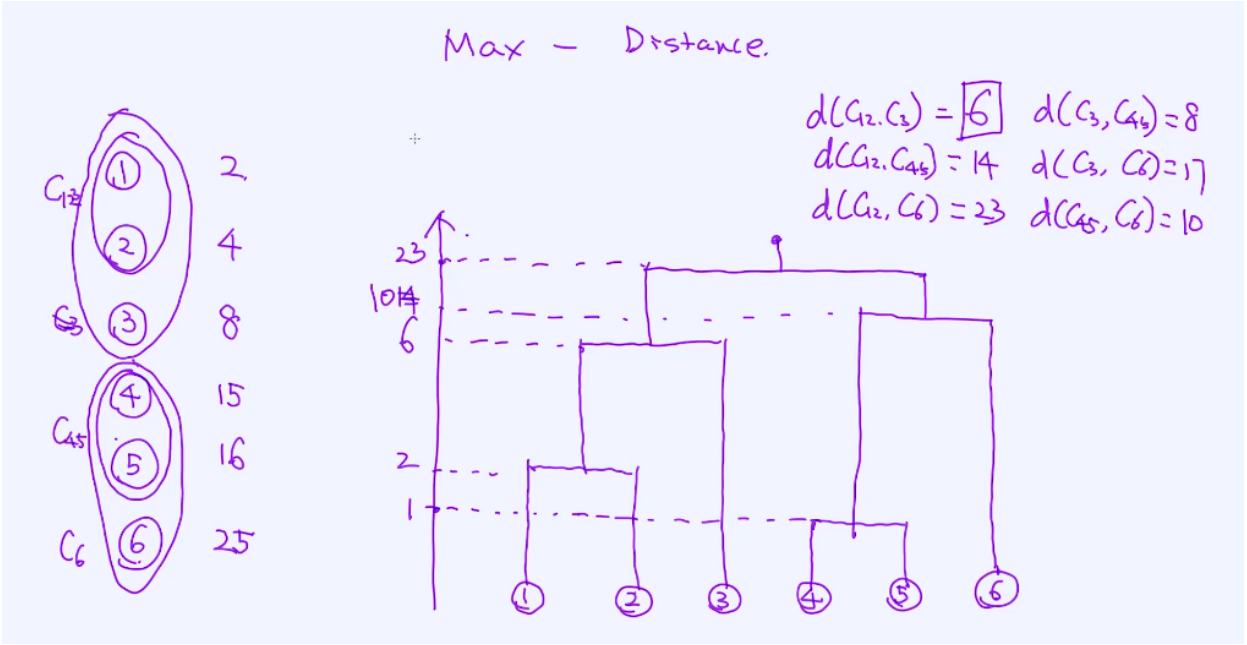




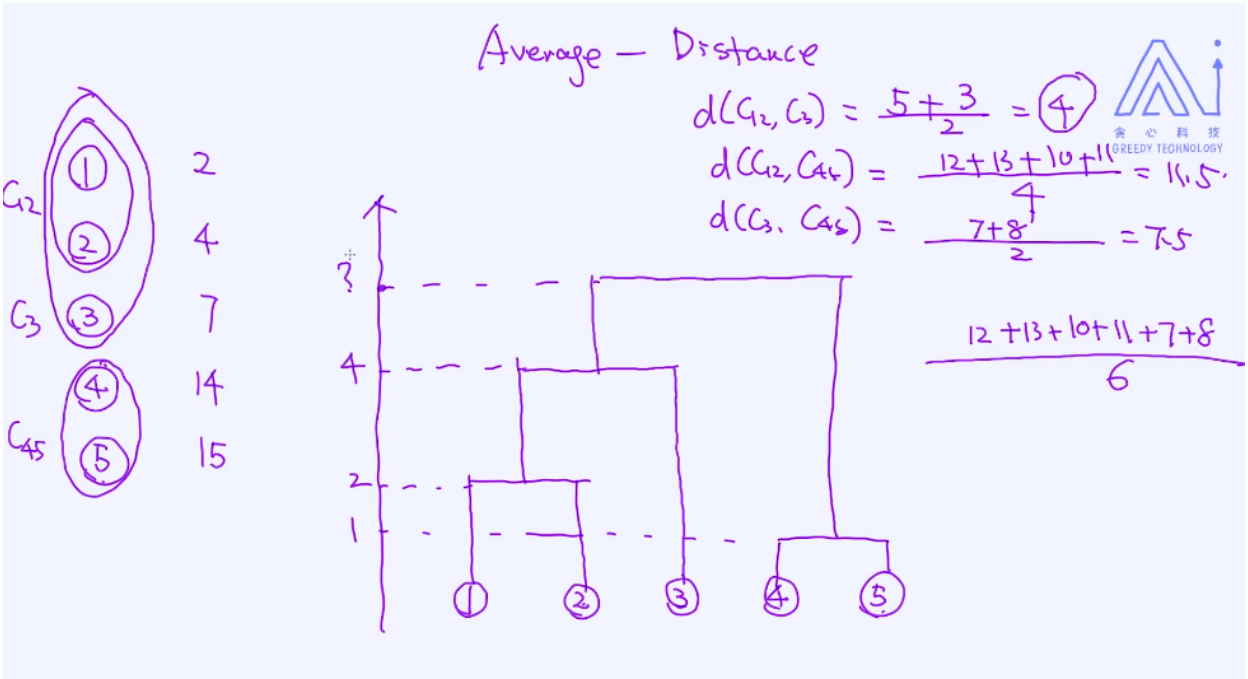
Min　　Max　　Average (平均)

Min – Distance

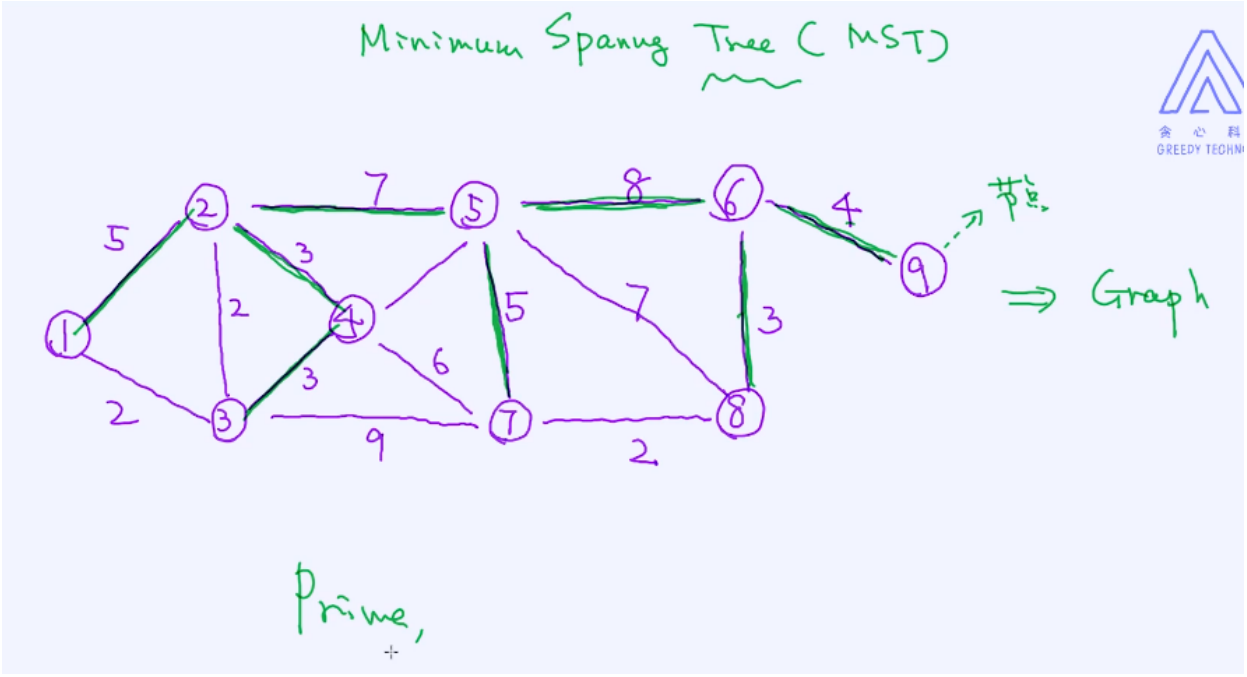




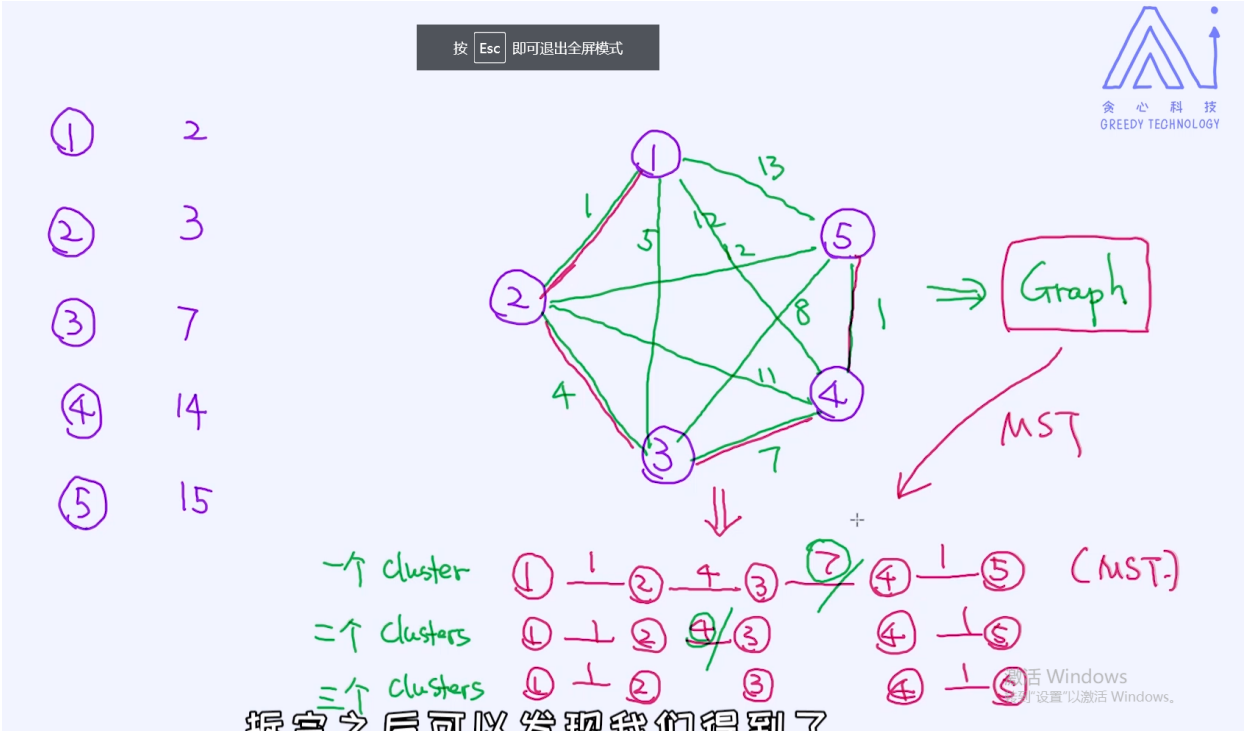
这三种方法里最稳定的方法



MST



自上而下聚类：最小生成树



在这里其实可以用类似于Elbow的方法。当我们合并完某两个clusters之后，就会发现距离会迅速变大，这个节点其实就是我们需要的！