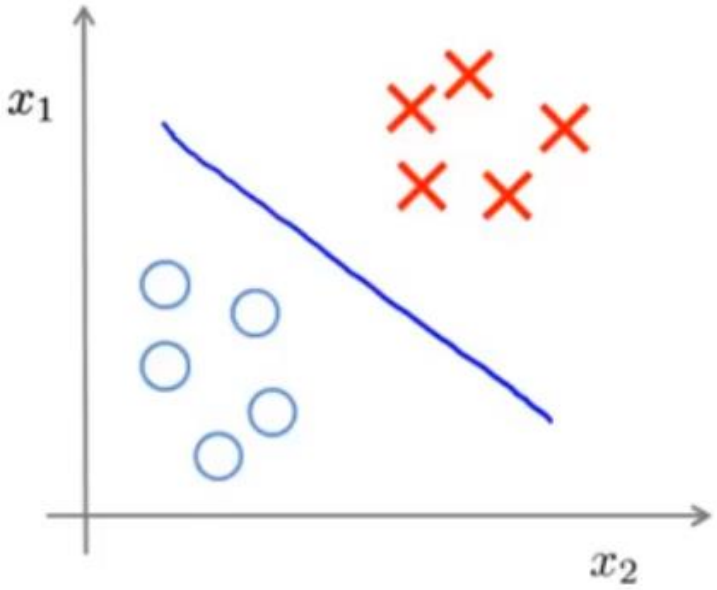
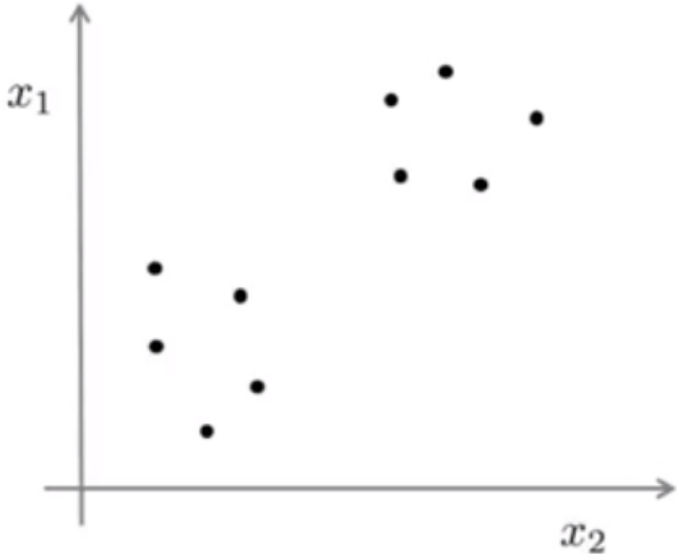
1、前言

先来回顾一下监督学习算法，如下图，这是一个监督学习的分类回归算法，给定数据以及它们的标签，通过拟合输入与输出，最后得出它们之间的映射关系。而在无监督学



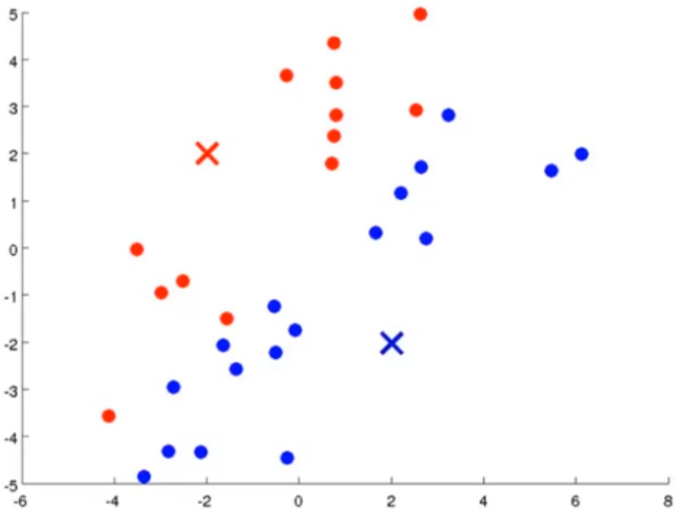
习算法中，我们得到的数据是不带标签的，如这幅图，为此，如果要将它们进行分类预测



的话，那么就要引出我们的第一个无监督学习算——聚类算法。聚类算法的应用很多，譬如：服务器集群分布、客户群体分类、天体分析等。

2、聚类算法

步骤 1：



根据数据分布选取适当个数的聚点，该图选择的聚点个数为2，然后进行随机选取聚点位置。

步骤 2：

将样本点与聚点远近作为标准进行蔟划分。

步骤 3：

对同一蔟类样本点求均值，得到的均值点位该类聚点下一次移动的位置。

重复步骤 2、3 若干次。

3、优化方式

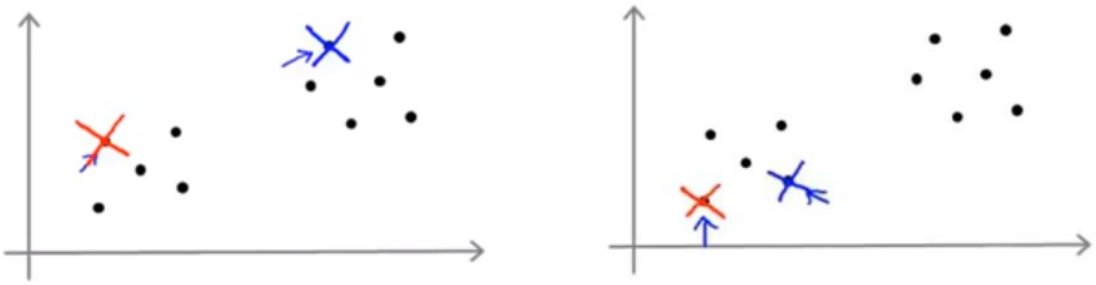
同样的，聚类算法也有相应的代价函数用于优化该分类器。代价函数如下：

$$\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-\mu_{c^{(i)}}||^2$$

于是，上方所讲步骤 2 和 3 都是用来最小化该代价函数的过程。

4、问题

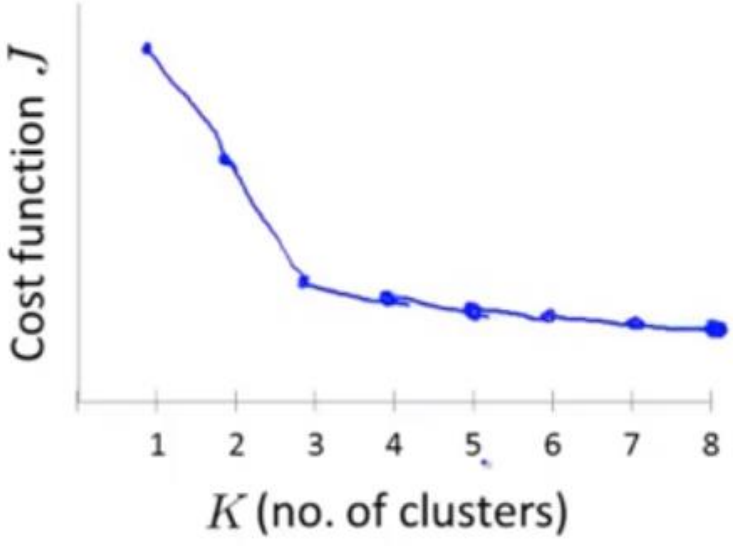
聚类算法最容易产生的几个问题其中之一就是局部最优问题，如下图所示：



左图表示的聚类算法达到了全局最优，但右图所见，可能只是局部最优解。

针对此问题，一个好的做法是，选取多次随机产生的聚点进行代价值计算，取代价值较小的那一个，可能就达到了全全局最优，此方法适用于聚点数量较少（通常<10）

还有一个问题就是聚点数量的选择，一般而言在数据量很大的情况下，难以初次就决定出正确的聚点个数，为此，引入了“肘部原则”。即从聚点数量为 1 开始，逐步增



加，看拐点对应的个数，即为要求的聚点数量。另一种选择原则是根据应用的后续目的，比如在商家根据身高决定衣服尺寸大小，如果想要服务得更好，那么好的做法是选择 5 个尺码大小的分类，而不是 3 个。