经典机器学习三-无监督半监督主动学习

无监督学习是从未标记标签的数据中该学习到数据的分布特点，此类学习任务中研究最多应用最广的是聚类任务，一般分为原型聚类、密度聚类和层次聚类三大部分。无监督学习的种类很多，比较有代表性的是老师讲解的KMeans和DBScan。

KMeans算法的过程如下：

输入：样本集以及聚类簇数k.

过程：从D中随机选择k个样本作为初始均值向量

Repeat

令

For j = 1, 2, …, m do

计算样本与各个均值向量的距离

根据距离最近的均值向量确定的簇标记；

将样本划入相应的簇

End for

For i = 1, 2, …, k do

计算新均值向量：

If then

将当前均值向量更新为

Else

保持当前均值向量不变

End if

End for

Until 当前均值向量均未更新

输出：簇划分

KMeans 的优点在于简单快捷，但是这个K值不太容易决定，而且在有些数据上起始表现并不好，比如一些例如moon数据集这样需要通过曲线分簇的数据就比较困难，KMeans的算法本身限制了其只能适用于线性可分的区域。

DBSCAN是“Density-Based Spatial Clustering of Applications with Noise”的简称，属于一种基于密度的聚类，此类算法假设聚类结构能够通过样本分布的紧密程度确定。在通常的情况下，密度聚类算法从样本的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇以获得最终的聚类结果。

DBSCAN是基于一组“邻域”(neighborhood)参数来刻画样本分布的紧密程度，其关键的参数有邻域、核心对象、密度直达、密度可达和密度相连的概念。DBSCAN的簇定义为：由密度可达关系到处的最大密度相连样本集合。

层次聚类试图在不同层次对数据集进行划分，从而形成树形的聚类结构。数据集的划分采用“自底向上”的聚合策略，也可采用自顶向下的分拆策略。

另外，本课程讲解了Mean Shift算法以及EM算法。

EM算法是针对现实应用中经常出现的不完整训练样本而设计的，对于一些存在“未观测”变量的情形，通常将其作为隐变量进行处理，对这些因变量计算期望来最大化已观测数据的对数边际似然。