作业7

1 GMM算法与k均值聚类的异同

1.1 相同点

- 它们都是可以用于聚类的算法
- 都需要指定K值，即类别数
- 都使用EM算法来求解
- 往往都只能收敛到局部最优

1.2 不同点

高斯混合模型可以给出一个样本属于某类的概率是多少；且高斯混合模型为生成模型，可以用于生成新的样本点

2 meanshift算法阅读报告

Meanshift算法建立在核密度估计的基础之上，它假设数据点集符合某一个概率分布，是沿着密度上升方向寻找同属于一簇数据点的迭代算法。核密度估计，是从数据点估计概率分布的非参数估计算法。

2.1 基本的Mean Shift算法

给定d维空间R^d的n个样本点, 在空间中任选一个点x, 定义mean shift（漂移向量）向量的基本形式为：

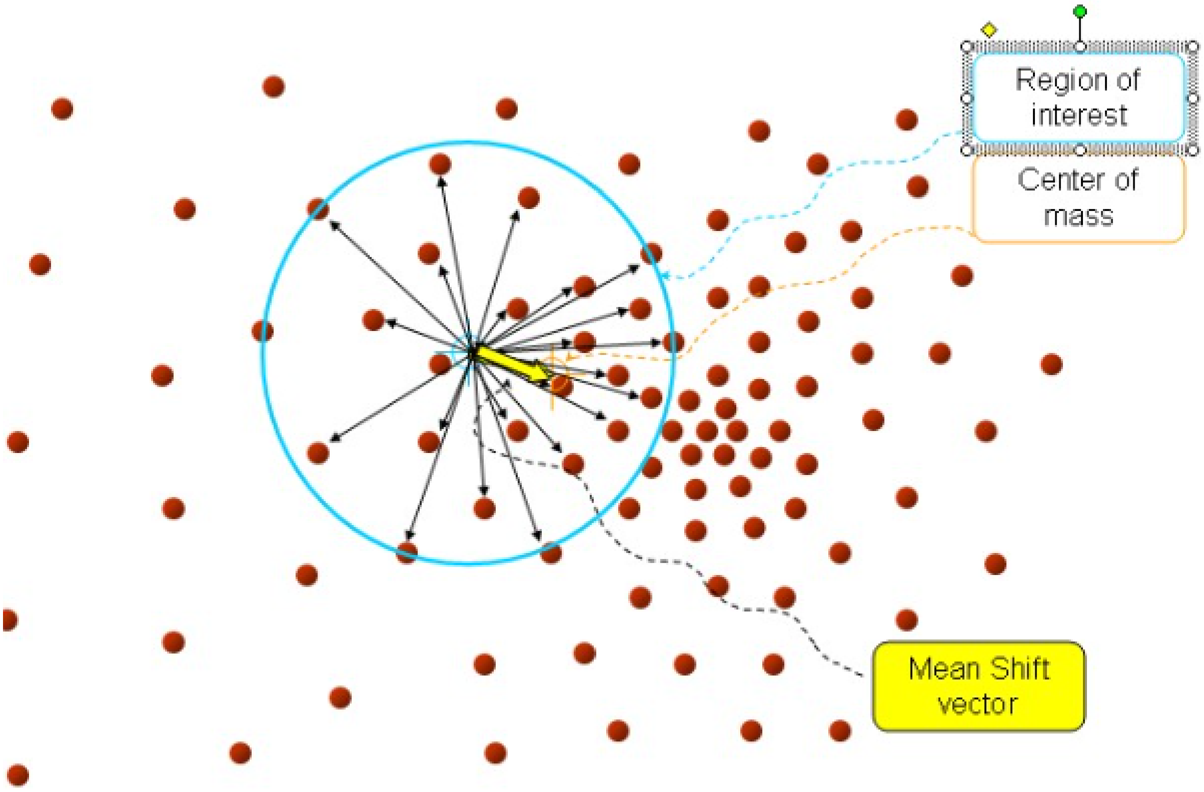
$$M_h = 1/K \sum_{x_i \in S_k} (x_i - x)$$

S_k是一个半径为h的超球域， 满足以下关系：

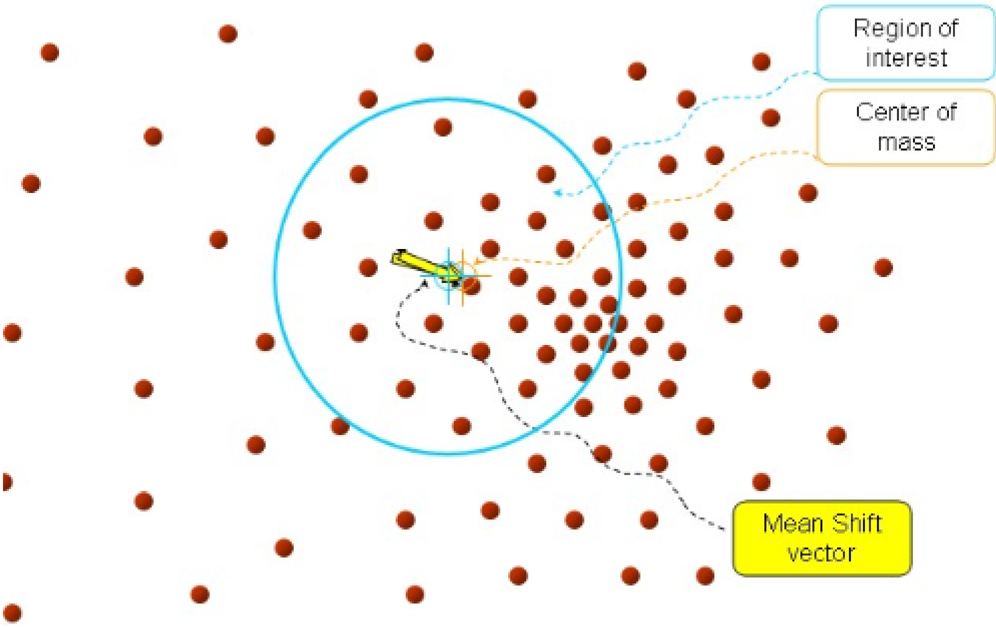
$$S_k(x) = \{y : (y - x_i)^T (y - x_i) < h^2\}$$

S_k的k表示在n个样本点中x_i中，有k个点落入S_k区域中。

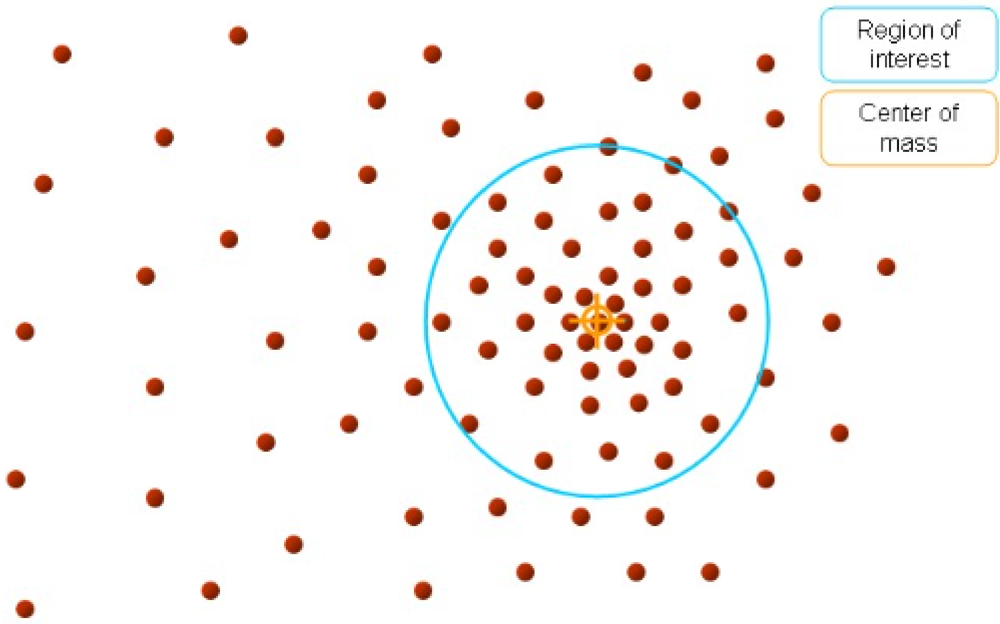
通俗理解便是：在R^n维空间中， 任意选择一个点，以这个点为圆心， h为半径做一个超球体， 落在这个球内部的所有点都会和圆心产生一个向量，向量以圆心为起点，以超球体内的任意一点为终点，将所有的向量相加（矢量加），就得到M_h, 即漂移向量，如下图：



再以当前漂移向量的终点为圆心， 重复上述过程，做一个超球体，再得到一个漂移向量。



如此重复，mean shift会收敛到概率密度最大的地方。



2.2 引入核函数的Mean Shift算法

引入核函数的mean shift算法变形为:

$$\hat{f}_{h,K}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} K(||\frac{x-x_i}{h}||^2)$$

上式中，K为核函数，h为半径，$\frac{c_{k,d}}{nh^d}$为单位密度，想要让上式达到最大，对上式求导。

$$f' = \frac{2c_{k,d}}{nh^d} \sum_{i=1}^{n} (x-x_i)K'(||\frac{x-x_i}{h}||^2)$$

令: $g(x) = -k'(x)$，K(x)叫做g(x)的影子核，即求到的负方向，将求导后的式子进行替换:

$$f' = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x-x_i)K'(||\frac{x-x_i}{h}||^2) = \frac{2c_{k,d}}{nh^d}[\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)][\frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x]$$

对上式，采用高斯核:

$$\frac{2c_{k,d}}{nh^{d+2}}[\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)][\frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x] =$$
$$\frac{c_{g,d}}{nh^d} \cdot \frac{2c_{k,d}}{c_{g,d}h^2}[\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)][\frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x]$$

其中:

$$\hat{f}_{h,G}(x) = \frac{c_{g,d}}{nh^d}[\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)]$$
$$m_{h,G}(x) = \frac{\sum_{i=1}^{n} x_i g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x$$

$m_{h,G}(x)$相当于meanshift向量的式子

2.3 算法步骤

给定：R^d空间中有n个样本点。

1 在未被标记的数据点中随机选择一个作为中心点x

2 以中心点为圆心，以h为半径，做一个超球体，将被超球体包含在内的点记做集合M，以圆心为起点，M内的点为终点，做向量，并求和，得到shift 向量。

3 以当前shift向量的终点为新的圆心，即x=x+shift, 圆心方向向shift的方向移动，移动距离为d = ||shift||。

4 重复2,3步骤，直到d ϵ(人为设定)，退出迭代。

5 将迭代过程中超球体内的点都归到一簇。

6 若当前簇圆心被包含在另一已存在的簇内，合并两簇。