第四次作业

1．请描述使用高斯混合模型进行数据聚类的过程。

2．对于数据：$x_1 = (4,5)^T$, $x_2 = (1,4)^T$, $x_3 = (0,1)^T$, $x_4 = (5,0)^T$, $x_5 = (4,1)^T$, $x_6 = (0,6)^T$ 现有以下三种聚类划分：

（1）$\{x_1, x_2, x_6\}$, $\{x_3, x_4, x_5\}$

（2）$\{x_1, x_4, x_5\}$, $\{x_2, x_3, x_6\}$

（3）$\{x_1, x_2, x_3, x_6\}$, $\{x_4, x_5\}$

假定我们聚类的准则是最小平方和误差，请判断上述三个划分中哪个更好？给出计算过程。

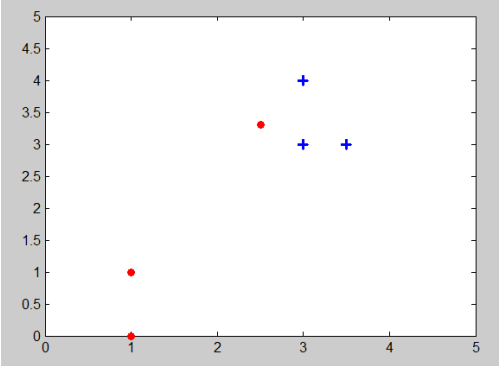
3．请阐述 K 均值聚类和模糊 K 均值聚类的异同。

4．证明：在 K 均值聚类中，在某次迭代的时候，将属于第 i 类的样本点移到第 j 类之后，属于第 i 类的样本点对应的误差平方和将变为：

$$J_i^* = J_i - \frac{n_i \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2}{n_i - 1}$$

其中 J_i 为移动前属于第 i 类样本的误差平方和

5．已知正样本点 $x_1=(1,1)^T$, $x_2=(1,0)^T$, $x_3=(2.5,3.3)^T$, 负样本点 $x_4=(3,3)^T$, $x_5=(3,4)^T$, $x_6=(3.5,3)^T$, 它们的分布如下图所示



（1）请写出线性支持向量机需要求解的原问题和对偶问题

（2）当 C 取值很大（比如 C->+∞）时，定性画出会得到的决策面，并解释原因

（3）当 C 取值很小（比如 C->0）时，定性画出会得到的决策面，并解释原因

6．结合图例，阐述线性可分支持向量机中的支持向量的概念。

7．结合图例，阐述线性可分支持向量机中的分类间隔的涵义。

8．请描述使用交叉验证对线性支持向量机的参数 C 进行设置的过程。

9．将支持向量机对应的优化问题进行对偶化之后，有什么优势？

编程题：

1、对如下的 30 个数据进行 K-均值聚类，聚类个数设置为 K=4。

（1）指出所使用的初始聚类中心，并报告在此条件下得到的最终聚类结果以及需要的迭代次数，对应的误差平方和。

（2）重新选择 3 组不同的初始聚类中心，给出对应的聚类结果和误差平方和。

编号	密度	含糖率	编号	密度	含糖率	编号	密度	含糖率
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

2、对上述数据集进行模糊 K-均值聚类，聚类个数设置为 K=4。指出使用的初始聚类中心、初始隶属度，报告在此初始化条件下的聚类结果（即：样本属于不同聚类的隶属度）以及需要的迭代次数。