

机器学习 K-means 作业

姓名：张泽群 学号：19049100002 班级：1班

Similarity-1

答： $D_{M1} = 5$ $D_{M2} = \sqrt{\frac{52}{3}}$

Similarity - 1

(1) $\text{mean} = [0, 0]^T$ $\text{cov} = \text{I}(e)$

$x = [3, 4]^T$ $\text{dist}_M = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$

$= \sqrt{[3, 4] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix}} = \sqrt{25} = 5$

(2) $\text{cov} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\text{dist}_M = \sqrt{[3, 4] \begin{bmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix}} = \sqrt{\frac{52}{3}}$

答：会发生变化。

Similarity-2

答：不是统计独立。其皮尔逊相关系数为 $P_x = \frac{7}{\sqrt{65}} \approx 0.868243$

$$(2) \quad \therefore \text{均值 } m_{S_1}, m_{S_2} = \int_{-\infty}^{\infty} S_1 \cdot p(S_1) dS_1 \\ = \int_{-\frac{1}{3}}^{\frac{1}{3}} \frac{S_1}{2\sqrt{3}} dS_1 = \frac{S_1^2}{4\sqrt{3}} \Big|_{-\frac{1}{3}}^{\frac{1}{3}} = 0$$

$$P_s = \frac{\text{COV}(S_1, S_2)}{\sqrt{\text{Var}(S_1)\text{Var}(S_2)}} = \frac{E[(S_1 - m_{S_1})(S_2 - m_{S_2})]}{\sqrt{DS_1 \cdot DS_2}}$$

$$= \frac{E[S_1 S_2]}{\sqrt{DS_1 \cdot DS_2}} \quad \because S_1, S_2 \text{ 相互独立}$$

$$\therefore E[S_1 S_2] = E[S_1] E[S_2] = m_{S_1} \cdot m_{S_2} = 0$$

$$m_{X_1} = 2m_{S_1} + 3m_{S_2} = 0$$

$$m_{X_2} = 2m_{S_1} + m_{S_2} = 0$$

$$P_s = 0$$

$$\rightarrow P_x = \frac{\text{COV}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}} = \frac{E[(X_1 - m_{X_1})(X_2 - m_{X_2})]}{\sqrt{DX_1 \cdot DX_2}}$$

$$= \frac{E[X_1 X_2]}{\sqrt{[EX_1^2 - (EX_1)^2] \cdot [EX_2^2 - (EX_2)^2]}}$$

$$\begin{matrix} \uparrow & \uparrow \\ m_{X_1}^2 = 0 & m_{X_2}^2 = 0 \end{matrix}$$

$$= \frac{E[(2S_1 + 3S_2)(2S_1 + S_2)]}{\sqrt{E[(2S_1 + 3S_2)^2] \cdot E[(2S_1 + S_2)^2]}}$$

$$= \frac{E[4S_1^2 + 3S_2^2]}{\sqrt{E[4S_1^2 + 9S_2^2] \cdot E[4S_1^2 + S_2^2]}} = \frac{7E(S_1^2)}{\sqrt{65} \cdot E(S_1^2)} = \frac{7}{\sqrt{65}} \approx 0.868243$$

SOM-3

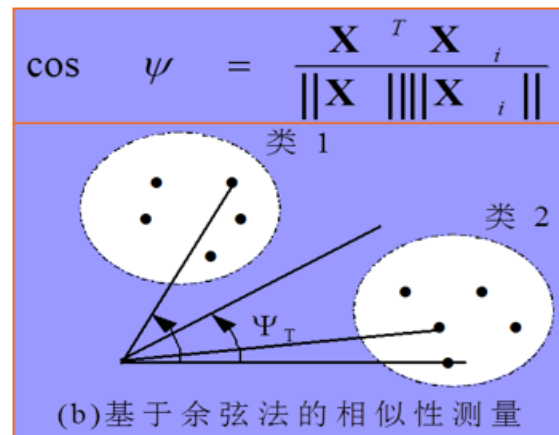
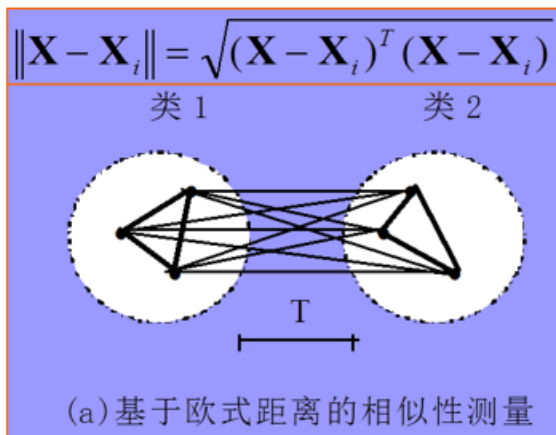
答：最为匹配的含义：当获取一个样本 \mathbf{x}_i 后，遍历竞争层中每一个节点：计算 \mathbf{x}_i 与节点之间的**相似度** (通常使用欧式距离) 选取**距离最小**的节点作为**优胜节点**(winner node),

即 $i(\mathbf{x}) = \arg \min \|\mathbf{x} - \mathbf{w}_j\|$

$$d_j(\mathbf{x}) = \sum_{i=1}^D (x_i - w_{ji})^2$$

距离的判别函数为：

可以用相似性/距离测度来表示。神经网络的输入模式向量的相似性测量可用向量之间的距离来衡量。常用的方法有欧氏距离法和余弦法两种。



SOM-C-1

```

clear all
clc

% 1. 导入数据
data = [-7.82, -4.58, -3.97;
        -6.68, 3.16, 2.71;
        4.36, -2.19, 2.09;
        6.72, 0.88, 2.80;
        -8.64, 3.06, 3.50;
        -6.87, 0.57, -5.45;
        4.47, -2.62, 5.76;
        6.73, -2.01, 4.18;
        -7.71, 2.34, -6.33;
        -6.91, -0.49, -5.68;
        6.18, 2.81, 5.82;
        6.72, -0.93, -4.04;
        -6.25, -0.26, 0.56;
        -6.94, -1.22, 1.13;
        8.09, 0.20, 2.25;
        6.81, 0.17, -4.15;
        -5.19, 4.24, 4.04;
        -6.38, -1.74, 1.43;
        4.08, 1.30, 5.33;
        6.27, 0.93, -2.78];

data = data';
% 2. 数据归一化
data = mapminmax(data);

plot3(data(1,:), data(2,:), data(3,:), '+r')           %在3D坐标上画这20个点
title('初始随机样本点分布'); %

% 3. 创建网络
net = newsom([0 1; 0 1; 0 1], [5, 8]); %控制竞争神经元的个数

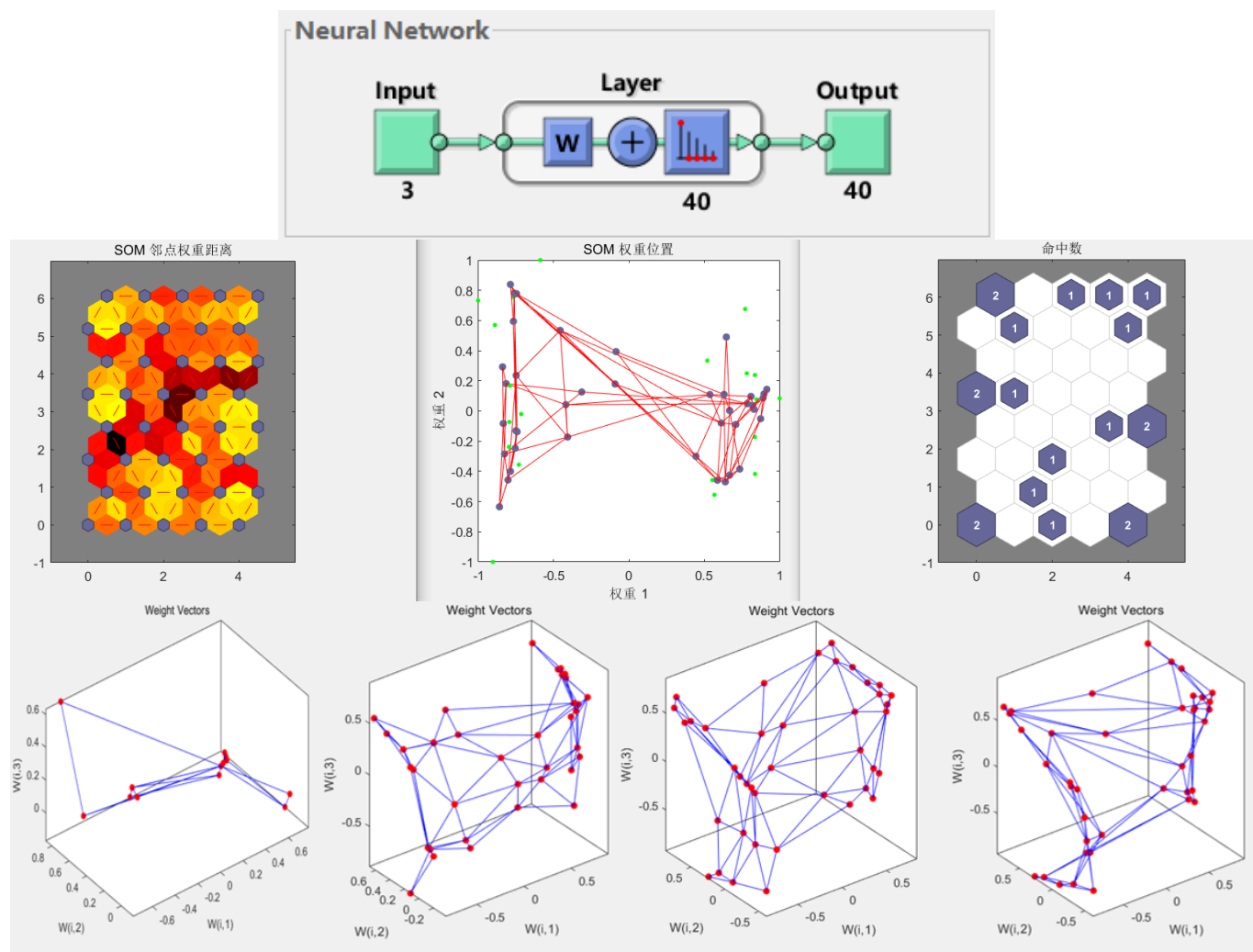
w1_init = net.iw{1, 1};
%绘制出初始权值分布图
figure;
plotsom(w1_init, net.layers{1}.distances)

% 4. 训练网络

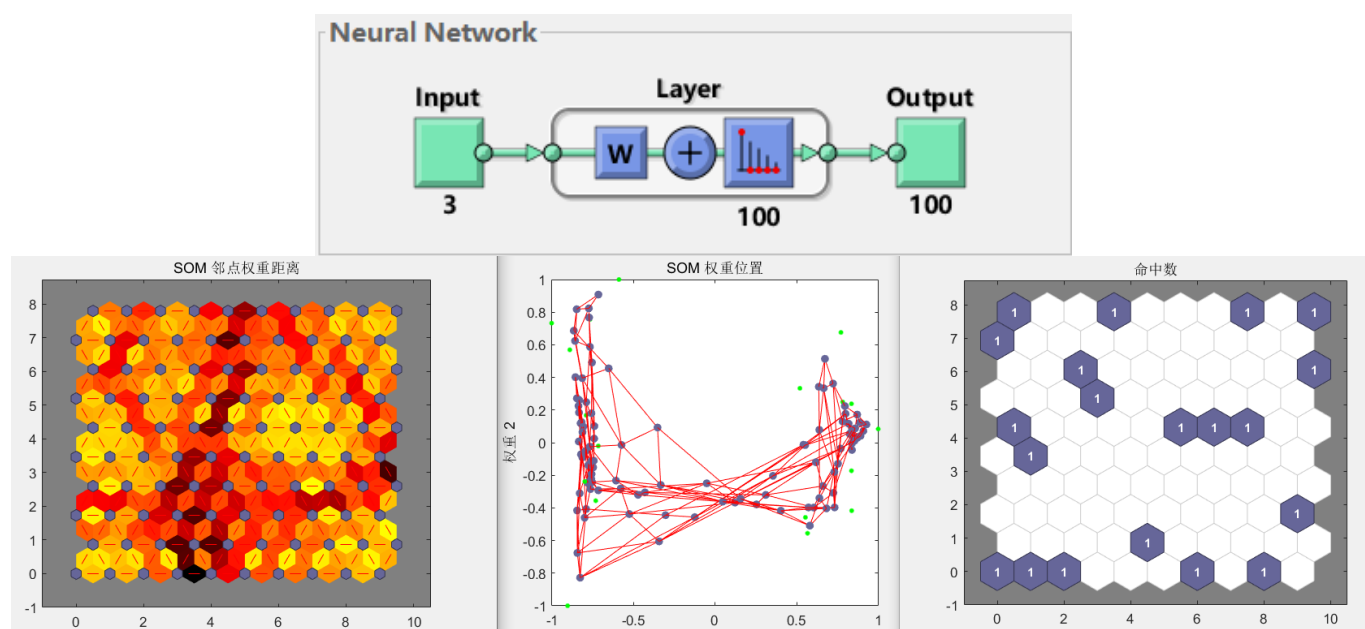
for i = 100:300:1000           %每循环一次，加300，共训练4次，分别是100,400,700,1000
net.trainParam.epochs = i;
net = train(net, data);
figure;
plotsom(net.iw{1, 1}, net.layers{1}.distances) %逐步绘制出权值分布
end

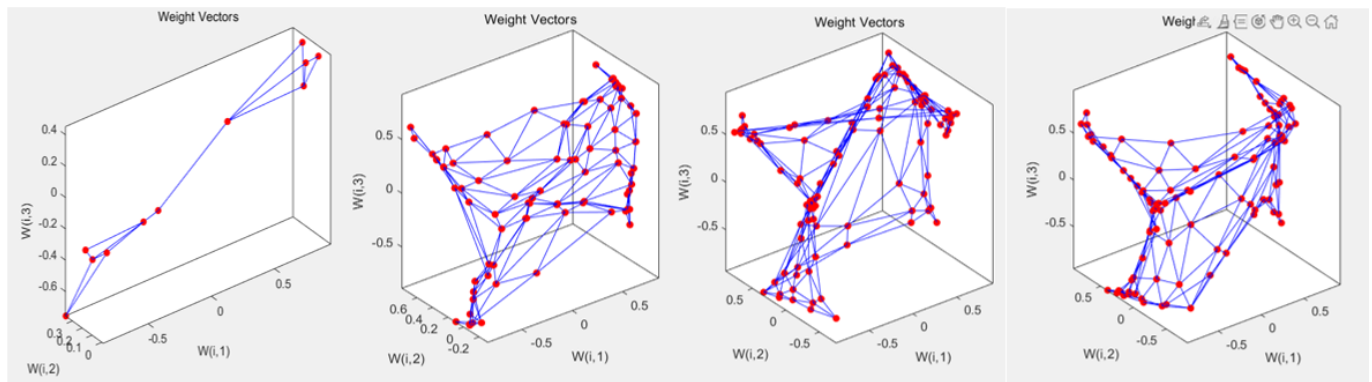
```

运行结果：



同时通过改变 `newsom()` 中第二参数，可以控制神经元的个数，从而对比聚类的效果。





结论： SOM网络结构比较简单，只有输入层和输出层（输出层我们通常也称为竞争层），输入层神经元的数量是由输入向量的维度决定的，一个神经元对应一个特征。

SOM网络结构的区别主要在竞争层：可以的有一维和二维的（竞争层也可以有更高的维度。不过出于可视化的目的，高维竞争层用的比较少）。其中，二维平面有2种平面结构：

①Rectangular；②Hexagonal。

输出层神经元数量设定和训练集样本的类别数相关，如果神经元节点数少于类别数，则不足以区分全部模式，训练的结果势必将相近的模式类合并为一类；相反，如果神经元节点数多于类别数，则有可能分的过细，或者是出现“死节点”，即在训练过程中，某个节点从未获胜过且远离其他获胜节点，因此它们的权值从未得到过更新。