Datasta Tietoon, Autumn 2007

SOLUTIONS TO EXERCISES 4

H4 / Problem 1.
Case $n = 2$. There are two vectors $\{1, 2\}$. Only one possibility,

$$C_1 = \{1\}, C_2 = \{2\}$$

Case $n = 3$. There are three vectors $\{1, 2, 3\}$. There are three possible groupings,

$$\begin{aligned} C_1 = \{1\}, C_2 = \{2, 3\}, \text{ or} \\ C_1 = \{2\}, C_2 = \{1, 3\}, \text{ or} \\ C_1 = \{3\}, C_2 = \{1, 2\}. \end{aligned}$$

Case $n = 4$. There are four vectors $\{1, 2, 3, 4\}$. There are seven possible groupings,

$$\begin{aligned} C_1 = \{1\}, C_2 = \{2, 3, 4\}, \text{ or} \\ C_1 = \{2\}, C_2 = \{1, 3, 4\}, \text{ or} \\ C_1 = \{3\}, C_2 = \{1, 2, 4\}, \text{ or} \\ C_1 = \{4\}, C_2 = \{1, 2, 3\}, \text{ or} \\ C_1 = \{1, 2\}, C_2 = \{3, 4\}, \text{ or} \\ C_1 = \{1, 3\}, C_2 = \{2, 4\}, \text{ or} \\ C_1 = \{1, 4\}, C_2 = \{2, 3\}. \end{aligned}$$

For $n = 5$ there are $5 + 4 + 3 + 2 + 1 = 15$ possible groupings. It seems that the number of groupings for n points is $2^{n-1} - 1$.

Let us prove that the number is $2^{n-1} - 1$. Take a binary vector of length n such that its i-th element

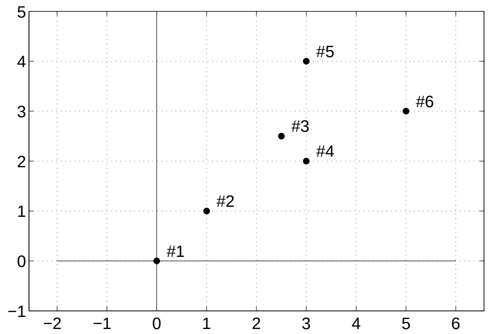
$$b_i = \begin{cases} 0, & \text{if } i\text{-th point is first cluster} \\ 1, & \text{if } i\text{-th point is second cluster} \end{cases}$$

All possible combinations are allowed except $b_i = 0$ for all i, $b_i = 1$ for all i, because then there is only one cluster. Thus the number is $2^n - 2$ (there are 2^n different binary vectors of length n).

But one half are equivalent to the other half because "first" and "second" cluster can be changed (consider case $n = 2$).

The final number is $\frac{1}{2}(2^n - 2) = 2^{n-1} - 1$.

H4 / Problem 2.



Hierarchical clustering and dendrogram (ryhmittelypuu). Combine clusters using the nearest distance ("single linkage"). In the beginning there are six clusters

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$$

Items 3 and 4 are nearest and combined

$$\{1\}, \{2\}, \{3, 4\}, \{5\}, \{6\}$$

Then the nearest clusters are 1 and 2

$$\{1, 2\}, \{3, 4\}, \{5\}, \{6\}$$

Next, 5 is connected to the cluster $\{3, 4\}$, because the distance from 5 to 3 (nearest) is smallest

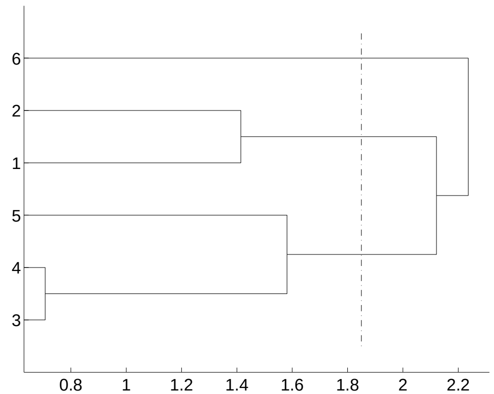
$$\{1, 2\}, \{3, 4, 5\}, \{6\}$$

Note that distance between 2 and 3 is smaller that of 6 to 4 or 5, and therefore

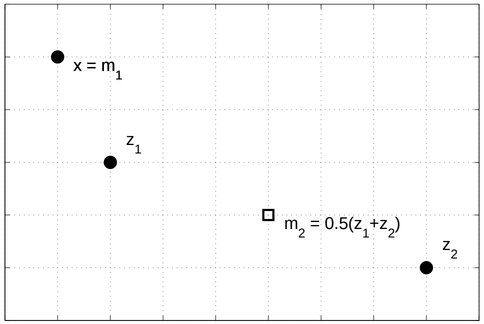
$$\{1, 2, 3, 4, 5\}, \{6\}$$

The algorithm ends when all points/clusters are combined to one big cluster.

The result can be visualized using the dendrogram, see the figure below. The best choice for three clusters is $\{1, 2\}$, $\{3, 4, 5\}$, $\{6\}$.



H4 / Problem 3.



Now $||\mathbf{z}_1 - \mathbf{m}_1|| < ||\mathbf{z}_1 - \mathbf{m}_2||$ and so $\mathbf{z}$ moves together with $\mathbf{x}$. New centers are:

$$\mathbf{m}_1' = 0.5(\mathbf{x} + \mathbf{z}_1), \qquad \mathbf{m}_2' = \mathbf{z}_2$$

$$\begin{aligned}
J^{OLD} &= ||\mathbf{z}_1 - \mathbf{m}_2||^2 + ||\mathbf{z}_2 - \mathbf{m}_2||^2 + \underbrace{||\mathbf{x} - \mathbf{m}_1||^2}_{0} \\
&= ||\mathbf{z}_1 - 0.5(\mathbf{z}_1 + \mathbf{z}_2)||^2 + ||\mathbf{z}_2 - 0.5(\mathbf{z}_1 + \mathbf{z}_2)||^2 \\
&= 0.25||\mathbf{z}_1 - \mathbf{z}_2||^2 + 0.25||\mathbf{z}_1 - \mathbf{z}_2||^2 \\
&= 0.5||\mathbf{z}_1 - \mathbf{z}_2||^2 \\
J^{NEW} &= ||\mathbf{z}_1 - \mathbf{m}_1'||^2 + \underbrace{||\mathbf{z}_2 - \mathbf{m}_2'||^2}_{0} + ||\mathbf{x} - \mathbf{m}_1'||^2 \\
&= 0.5||\mathbf{x} - \mathbf{z}_1||^2
\end{aligned}$$

Now we remember that $||\mathbf{z}_1 - \mathbf{m}_1||^2 < ||\mathbf{z}_1 - \mathbf{m}_2||^2$ (that is why $\mathbf{z}_1$ moved to the other cluster).

$$\Rightarrow ||\mathbf{z}_1 - \underbrace{\mathbf{x}}_{\mathbf{m}_1}||^2 < ||\mathbf{z}_1 - \underbrace{0.5(\mathbf{z}_1 + \mathbf{z}_2)}_{\mathbf{m}_2}||^2 = 0.25||\mathbf{z}_1 - \mathbf{z}_2||^2$$

So,

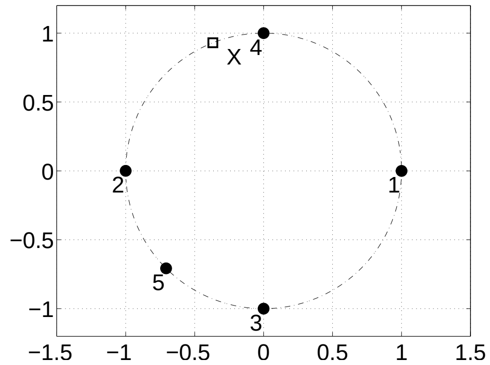
$$J^{NEW} = 0.5||\mathbf{x} - \mathbf{z}_1||^2 < 0.5 \cdot 0.25||\mathbf{z}_1 - \mathbf{z}_2||^2 < 0.5||\mathbf{z}_1 - \mathbf{z}_2||^2 = J^{OLD}$$

H4 / Problem 4.
Number or neurons is N^2. For each neuron j, whe have to compute

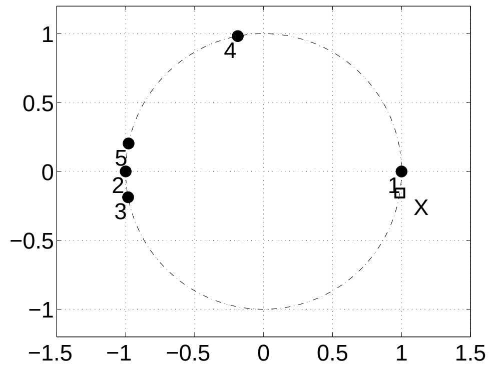
$$\sum_{i=1}^{d} (x_i - m_{ij})^2$$

which takes d subtractions, d multiplications, $d-1$ additions. This means totally $N^2(2d-1)$ additions (subtraction and addition are usually equivalent) and $N^2 d$ multiplications.

H4 / Problem 5.
Choose $\mathbf{x}$ so that its angle is a little less than 135°.



Now best matching unit (BMU): 4, neighbours: 5 and 3. They move on the circle half-way towards $\mathbf{x}$.



Now choose x so that its angle is very small negative. BMU: 1, neighbours: 5 and 2. They are moving closer to $\mathbf{x}$ along unit circle. 5 jumps over 4, and 2 jumps over 3. Now 1D SOM is in order: 1, 2, 3, 4, 5.