Datasta Tietoon, Autumn 2007

EXERCISE PROBLEMS 4 [Nov 30th 2007, Dec 5th 2007]

H4 / 1. (Cluster analysis)

We are given n vectors. In how many ways can we divide them into two clusters (groups)? Solve at least the cases $n = 2, 3, 4, 5$.

H4 / 2. (Cluster analysis)

We are given the following data matrix:

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 2.5 & 3 & 3 & 5 \\ 0 & 1 & 2.5 & 2 & 4 & 3 \end{bmatrix}$$

a) Plot the data vectors in the coordinate plane.

b) Perform a hierarchical clustering based on your plot. As the distance between two clusters, use the smallest distance between two vectors belonging to the two clusters. Plot the clustering tree. What is the best clustering into three clusters ?

H4 / 3. (Cluster analysis)

We are given three vectors $\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2$. In the beginning $C_1 = \{\mathbf{x}\}$, $C_2 = \{\mathbf{z}_1, \mathbf{z}_2\}$.

a) Derive the means $\mathbf{m}_1$, $\mathbf{m}_2$ of the two clusters.

b) It turns out that $\|\mathbf{z}_1 - \mathbf{m}_1\| < \|\mathbf{z}_1 - \mathbf{m}_2\|$ and thus in the c-means-algorithm the vector $\mathbf{z}_1$ is moved from cluster C_2 into the cluster C_1. Denote the new clusters by $C_1' = \{\mathbf{x}, \mathbf{z}_1\}$, $C_2' = \{\mathbf{z}_2\}$. Derive the new means $\mathbf{m}_1'$, $\mathbf{m}_2'$.

c) Prove that

$$\sum_{\mathbf{x} \in C_1} \|\mathbf{x} - \mathbf{m}_1\|^2 + \sum_{\mathbf{x} \in C_2} \|\mathbf{x} - \mathbf{m}_2\|^2 > \sum_{\mathbf{x} \in C_1'} \|\mathbf{x} - \mathbf{m}_1'\|^2 + \sum_{\mathbf{x} \in C_2'} \|\mathbf{x} - \mathbf{m}_2'\|^2$$

meaning that the criterion used in the c-means clustering is decreasing.

H4 / 4. (SOM)

Let us consider the computational complexity of the SOM algorithm. Assume that the size of the map is $N \times N$ units (neurons), and the dimension of the input and weight vectors is d. How many additions and multiplications are needed when the winner neuron is found for an input vector $\mathbf{x}$, when the Euclidean distance is used ?

H4 / 5. (SOM)

Let us assume that the weight vectors $\mathbf{m}_i$ and input vectors $\mathbf{x}$ of the SOM are on the unit circle (they are 2-dimensional unit vectors). The map is a 1-dimensional grid of 5 units whose weight vectors are initially as shown in Fig. 1. The neighborhood is defined cyclically so that the neighbors of units $b = 2, 3, 4$ are $b - 1, b + 1$, those of unit 5 are 4 and 1, and those of unit 1 are 5 and 2.

In the training, the coefficient $\alpha = 0.5$, meaning that at each step the weight vectors of the winner unit and its neighbors move along the unit circle halfway towards the input $\mathbf{x}$. You are free to choose your inputs freely from the unit circle. Choose a sequence of training vectors $\mathbf{x}$ so that the weight vectors become ordered.