Pattern Recognition
Lab week 6

Useful matlab functions:

`pdist, linkage, cluster, clusterdata, dendrogram`

Assignment 1: *clustering, distance-based connected components, distance threshold.*

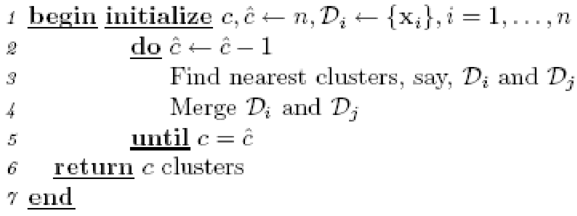
Given is the data set in `cluster_data.mat`, which contains 2-dimensional data points. The dissimilarity between 2 points, a and b, is given by the *Minkowski metric*, see equation (1).

$$d_M(a,b) = \left(\sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k} \tag{1}$$

1. Take $k = 2$. What does the Minkowski metric with $k = 2$ remind you of?
2. Make a plot of all the points in the data set and connect by a line any two points if their dissimilarity is smaller than a threshold t, resulting in a clustering defined by the connected components.
3. Study the influence of the threshold value ($t = \{0.05, 0.1, 0.15, 0.2, 0.25\}$) on the final result of clustering. Which of these values for t do you consider to be best suited for clustering this particular dataset? Why?

Assignment 2: *agglomerative hierarchical clustering, distance function, dendrogram.*

The *agglomerative hierarchical clustering* algorithm joins points (or groups of points) together based on their similarity. This can be done until only c groups of points remain:



```
begin initialize $c, \hat{c} \leftarrow n, \mathcal{D}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \ldots, n$
        do $\hat{c} \leftarrow \hat{c} - 1$
            Find nearest clusters, say, $\mathcal{D}_i$ and $\mathcal{D}_j$
            Merge $\mathcal{D}_i$ and $\mathcal{D}_j$
        until $c = \hat{c}$
    return $c$ clusters
end
```

1. Using the data set given (`cluster_data.mat`), use *agglomerative hierarchical clustering* to cluster the points in $c = 4$ groups. Compute the cluster centroids (as means of all points in a cluster) and plot them together with the data for each of the following four

distance functions: *min*, *max*, *average*, *mean*, given by equations 2 to 5.

$$d_{\min}(D_i, D_j) = \min_{x \in D_i, x' \in D_j} \|x - x'\| \tag{2}$$

$$d_{\max}(D_i, D_j) = \max_{x \in D_i, x' \in D_j} \|x - x'\| \tag{3}$$

$$d_{\text{avg}}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{x' \in D_j} \|x - x'\| \tag{4}$$

$$d_{\text{mean}}(D_i, D_j) = \|m_i - m_j\| \tag{5}$$

$$\text{with } m_i = \frac{1}{n_i} \sum_{x \in D_i} x \tag{6}$$

hint: Call the `help` function on the 'Useful Matlab functions' listed on the first page of this document.

2. Plot the *dendrograms* for the four previous solutions. Describe and explain the differences between the dendrograms.

Assignment 3: *criterion functions for clustering, minimum variance partitioning.*

Let $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $x_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, $x_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$, $x_4 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$, $x_5 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$;
and consider the following five possible clusterings:

$$\{\{x_1, x_2, x_3\}, \{x_4, x_5\}\} \tag{7}$$

$$\{\{x_2, x_3, x_5\}, \{x_1, x_4\}\} \tag{8}$$

$$\{\{x_4\}, \{x_1, x_2, x_3, x_5\}\} \tag{9}$$

$$\{\{x_4, x_5\}, \{x_1, x_2, x_3\}\} \tag{10}$$

$$\{\{x_3, x_5\}, \{x_1, x_2, x_4\}\} \tag{11}$$

1. Calculate for each of the given clusterings (7-11) the sum-of-squared error criterion J_e (with m_i as defined in eq. 6).

$$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} \|x - m_i\|^2$$

2. Which of the clusterings minimizes the sum-of-squared error?

Assignment 4: *Application of hierarchical clustering.*

The Netherlands has 12 provinces. Some statistical data is given below. This data was taken from the site `http://en.wikipedia.org/wiki/Ranked_list_of_Dutch_provinces` as of October 2012.

Province	Population (2004 estimate)	Area (km^2)	Density	GDP (2003, PPS in mil. €)	GDP per cap. (2003, in €)
South Holland	3,453,000	2,860	1,207.3	95,868	27,825
North Holland	2,583,900	2,660	971.4	65,295	27,169
Utrecht	1,159,200	1,356	854.9	38,355	33,148
Limburg	1,143,000	2,167	527.5	28,038	24,585
North Brabant	2,406,900	4,938	487.4	65,295	27,169
Gelderland	1,967,600	4,995	393.9	45,043	22,942
Overijssel	1,105,800	3,337	331.4	25,854	23,441
Flevoland	356,400	1,426	249.9	6,915	19,439
Groningen	575,900	2,344	245.7	18,496	32,245
Zeeland	378,300	1,792	211.1	9,354	24,706
Friesland	642,500	3,361	191.2	13,989	21,830
Drenthe	482,300	2,652	181.9	10,323	21,427

Requirement:

Write a short report that contains (at least):

- Preprocessing. Apply the z-transform to every feature individually and work further with the transformed data.
- Compute a dissimilarity matrix using Euclidean distance.
- Use hierarchical clustering with single linkage to create a dendrogram. Don't forget to label the provinces in the dendrogram.
- Comment on the result. Which provinces are most similar and which are most dissimilar? Which are the main factors (features) that determine the result?