5.2. Cluster Analysis

- Aim: group n objects into k classes
- $k << n$, k often unknown
- Objects within classes are "close" to eachother, objects in different classes are well discriminable.
- Needed: metric to define closeness vs. discriminable.

5.2.1 k-means clustering

~~Aim~~: Given $x_1, \dots, x_n \in \mathbb{R}^p$

Aim: Partition data into clusters $S_1, \dots, S_k$,
$S_1, \dots, S_k$ a partition of $\{1, \dots, n\}$,
with centers $\mu_1, \dots, \mu_k \in \mathbb{R}^p$ as solution to

$$\min_{\substack{S_1, \dots, S_k \text{ part.} \\ \mu_1, \dots, \mu_k}} \sum_{\ell=1}^{k} \sum_{i \in S_\ell} \|x_i - \mu_\ell\|^2$$

Given the partition, the optimal centers $\bar{x}_\ell = \frac{1}{n_\ell} \sum_{i \in S_\ell} x_i$.

k-means algorithm, Lloyd's algorith

Alternate between:

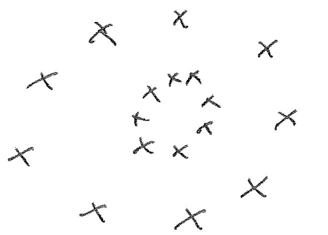
- Given centers $\mu_1, \dots, \mu_k$ assign each point x_i to cluster $\ell = \arg\min_j \|x_i - \mu_j\|$
- update the centers $\mu_\ell = \frac{1}{|S_\ell|} \sum_{i \in S_\ell} x_i$

> needs to know the no. of clusters k a-priori

> needs Euclidean space

> may end in suboptimal solutions

> solution has always convex clusters



How to cluster?

5.2.2. Spectral clustering

To overcome these difficulties.

Given $x_1, \dots, x_n$.
Construct a weighted graph $G = (V, E, W)$
Each point x_i is a vertex v_i, $i = 1, \dots, n$
Edges weights w_{ij} are $w_{ij} = K_\varepsilon(\|x_i - x_j\|)$
with Kernel K_ε, e.g., $K_\varepsilon(u) = \exp\left(-\frac{1}{2\varepsilon} u^2\right)$.

Note: $\|x_i - x_j\|$ can be substituted by any dissimilarity measure.

Consider a random walk with transition matrix

$$M = D^{-1} W$$

$$P(X_{t+1} = j \mid X_t = i) = \frac{w_{ij}}{\deg(i)} = M_{ij}$$

$$D = \operatorname{diag}(\deg(i)), \quad \deg(i) = \sum_{l=1}^{n} w_{il}.$$

Decompose $\quad M = \Phi \Lambda \Psi^T = \sum_{k=1}^{n} \lambda_k \varphi_k \psi_k^T$

$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n), \ \lambda_1 \geq \dots \geq \lambda_n$

$\Phi = (\varphi_1, \dots, \varphi_n), \ \Psi = (\psi_1, \dots, \psi_n)$

biorthonormal system, right/left eigenvectors.

Then $\quad M^t = \sum_{k=1}^{n} \lambda_k^t \varphi_k \psi_k^T$

$m_{ij}^{(t)}$ distribution of being in vertex j ~~at~~ having started in i

$$V_i \rightarrow e_i^T M^t = \sum_{k=1}^{n} \lambda_k^t e_i \varphi_k \psi_k^T = \sum_{k=1}^{n} \underbrace{\lambda_k^t \varphi_{k,i}}_{\text{coefficient}} \ \underbrace{\psi_k^T}_{\text{orth. basis}}$$

Recall: If v_i, v_j close / strongly connected, then $e_i^T M^t$ and $e_j^T M^t$ are similar.

Diffusion map truncated to d dimensions

$$\varphi_t^{(d)}(i) = \begin{pmatrix} \lambda_2^t \varphi_{2,i} \\ \vdots \\ \lambda_{d+1}^t \varphi_{d+1,i} \end{pmatrix} \qquad \text{(see 4.3)}$$

Aim: cluster vertices of the graph into k clusters.

<u>Algorithm 5.3.</u> (Spectral clustering)

Given a graph $G=(V,E,W)$, no. of clusters k, t.

Compute the $(k-1)$-dim. diff. map

$$\phi_t^{(k-1)}(i) = \begin{pmatrix} \lambda_2^t \varphi_{1,i} \\ \vdots \\ \lambda_k^t \varphi_{k,i} \end{pmatrix}$$

and cluster $\phi_t^{(k-1)}(1), \dots, \phi_t^{(k-1)}(n) \in \mathbb{R}^{k-1}$

using, e.g., k-means clustering.

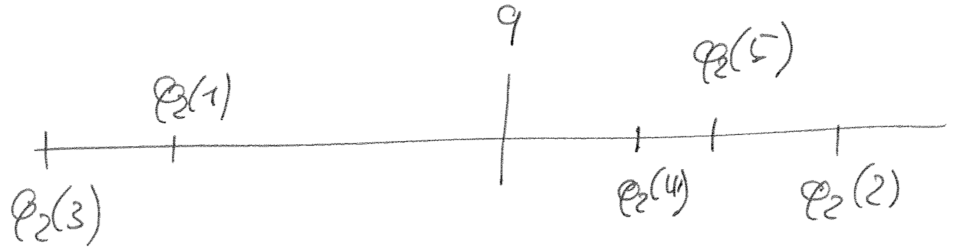
Particularly for two clusters S, S^c:

$$\phi_t^{(1)}(i) \in \mathbb{R}^1, \quad i=1,\dots,n$$

will be on a line. Natural way of clustering on a line is to define a threshold q such that

$$v_i \in S \text{ if } \phi_t^{(1)}(i) \le q.$$

e.g. for 5 points



5.2.3 Hierarchical clustering

Given n objects $v_1, \dots, v_n$ and pairwise dissimilarities

$$\delta_{ij} = \delta_{ij} \quad , \quad \Delta = (\delta_{ij})_{i,j=1,\dots,n}$$

Define a linkage function between clusters C_1, C_2

$$d(C_1, C_2) = \begin{cases} \min\limits_{i \in C_1, j \in C_2} \delta_{ij} & \text{single linkage} \\ \max\limits_{i \in C_1, j \in C_2} \delta_{ij} & \text{Complete linkage} \\ \frac{1}{|C_1||C_2|} \sum\limits_{i \in C_1, j \in C_2} \delta_{ij} & \text{average linkage} \end{cases}$$

Algorithm (Agglomerative Clustering)

```
Initialize clusters as singletons : for $i = 1$ to $n$ do $C_i \leftarrow \{i\}$
Initializ set of clusters available for merging : $S \leftarrow \{1, \dots, n\}$
repeat
    Pick 2 most similar cluster to merge : $(j,k) \leftarrow \arg\min_{j,k \in S} d(C_j, C_k)$
    Create new cluster $C_\ell \leftarrow C_j \cup C_k$
    Mark j and k as unavailable : $S \leftarrow S \setminus \{j, k\}$
    if $C_\ell \neq \{1, \dots, n\}$ then
        Mark $\ell$ as available , $S \leftarrow S \cup \{\ell\}$
    for each $i \in S$ do
        Update dissimilarities $d(C_i, C_\ell)$
until no more clusters are available for merging
```

Table 13.2.3 Mahalanobis distances D_{ij} between 10 island races of white-toothed shrews (from Delany and Healy, 1966; Gower and Ross, 1969)

		Scilly Islands					Channel Islands				France
		1	2	3	4	5	6	7	8	9	10
Scilly Islands	1. Tresco	0									
	2. Bryher	1.61	0								
	3. St Agnes	1.97	2.02	0							
	4. St Martin's	1.97	2.51	2.88	0						
	5. St Mary's	1.40	1.70	1.35	2.21	0					
Channel Islands	6. Sark	2.45	3.49	3.34	3.83	3.19	0				
	7. Jersey	2.83	3.94	3.64	2.89	3.01	3.00	0			
	8. Alderney	9.58	9.59	10.05	8.78	9.30	9.74	9.23	0		
	9. Guernsey	7.79	7.82	8.43	7.08	7.76	7.86	7.76	2.64	0	
French mainland	10. Cap Gris Nez	7.86	7.92	8.36	7.44	7.79	7.90	8.26	3.38	2.56	0

Mardia, Kent, Bibby : Multivariate Analysis

Table 13.3.1 Single linkage procedure

Order	Distances (ordered)	Clusters
1	$d_{35} = 1.35$	(1), (2), (3, 5), (4), (6), (7), (8), (9), (10)
2	$d_{15} = 1.40$	(1, 3, 5), (2), (4), (6), (7), (8), (9), (10)
3	$d_{12} = 1.61$	(1, 2, 3, 5), (4), (6), (7), (8), (9), (10)
4	$d_{25} = 1.70$	(1, 2, 3, 5), (4), (6), (7), (8), (9), (10)†
5‡	$d_{14} = 1.969$	(1, 2, 3, 4, 5), (6), (7), (8), (9), (10)
6‡	$d_{13} = 1.972$	(1, 2, 3, 4, 5), (6), (7), (8), (9), (10)†
7	$d_{23} = 2.02$	(1, 2, 3, 4, 5), (6), (7), (8), (9), (10)†
8	$d_{45} = 2.21$	(1, 2, 3, 4, 5), (6), (7), (8), (9), (10)†
9	$d_{16} = 2.45$	(1, 2, 3, 4, 5, 6), (7), (8), (9), (10)
10	$d_{24} = 2.51$	(1, 2, 3, 4, 5, 6), (7), (8), (9), (10)†
11	$d_{9,10} = 2.56$	(1, 2, 3, 4, 5, 6), (7), (8), (9, 10)
12	$d_{89} = 2.64$	(1, 2, 3, 4, 5, 6), (7), (8, 9, 10)
13	$d_{17} = 2.84$	(1, 2, 3, 4, 5, 6, 7), (8, 9, 10)
14	$d_{34} = 2.88$	(1, 2, 3, 4, 5, 6, 7), (8, 9, 10)†
15	$d_{47} = 2.89$	(1, 2, 3, 4, 5, 6, 7), (8, 9, 10)†
16	$d_{67} = 3.00$	(1, 2, 3, 4, 5, 6, 7), (8, 9, 10)†
.	.	.
.	.	.
.	.	.
45	$d_{49} = 7.08$	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

† No new clusters.

‡ More accurate values of distances to break the tie.

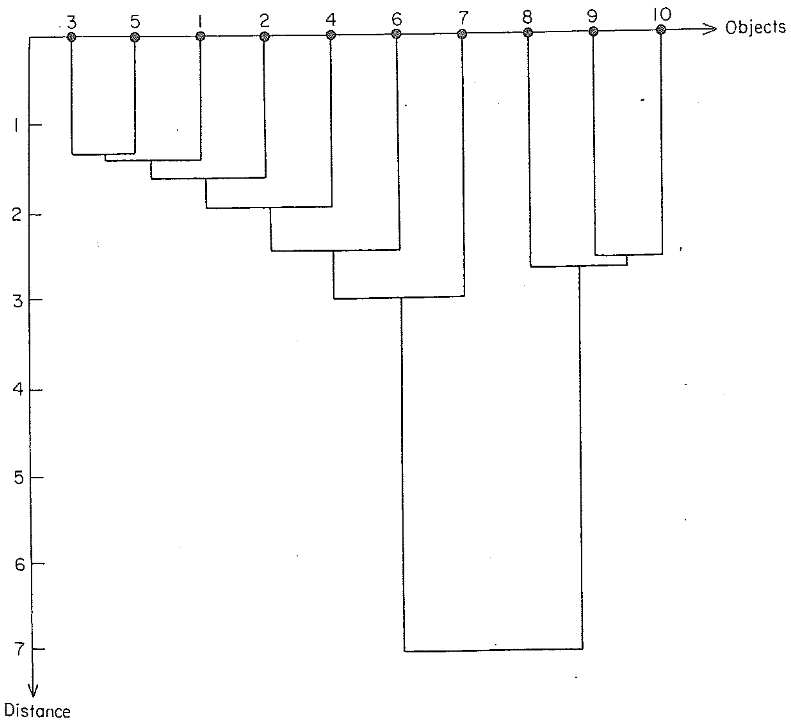


Figure 13.3.1 Dendrogram for shrew data (single linkage).

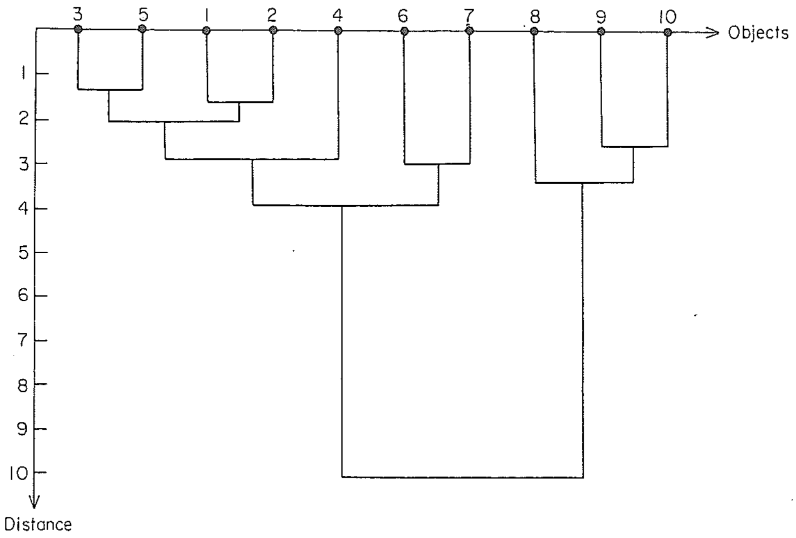


Figure 13.3.2 Dendrogram for the shrew data (complete linkage).

Visualization by <u>dendrograms</u> or <u>binary trees</u>.

6. Support Vector Machines (SVM)

- SVM learning alg. among the best "off-the-shelf" supervised learning algorithm.
- Application of kernels makes SVM very flexible.

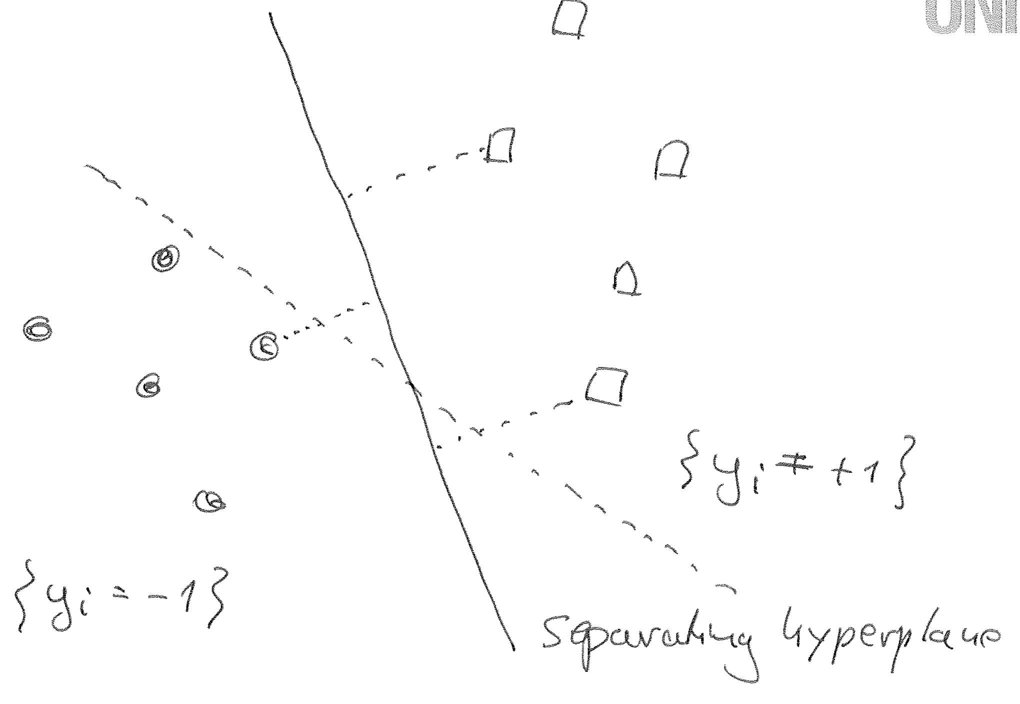
Advantages:

- effective in high-dim. spaces
- also effective if no. of dim. $\geq$ no. of samples
- uses only a subset of points in the decision fct. (called support vectors), also memory efficient.

Given a ~~t~~ training set $(x_1, y_1), \dots, (x_n, y_n)$,

$x_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$.
data points class membership

Key idea: select a particular hyperplane that separates the points into two classes and maximizes the margin, i.e., the distance between the hyperplane and the closest point of the training set.


$\{y_i = +1\}$
$\{y_i = -1\}$
Separating hyperplane