5.2 Cluster Analysis

→ The aim is to group n objects into g classes.

- $g \ll n$

- Metic : success of clustering algorithms.

This is an example of unsupervised Learning algorithm.

5.2.1. K means clustering

* How to find the number of groups g?

We assume g is given.

$x_1, \dots, x_n \in \mathbb{R}^p$: The purpose of K-means clustering is to partition the data set into clusters $C_1, \dots, C_g$ with centers $u_1, \dots, u_g$ as solutions to

$$\min_{\substack{C_1,\dots,C_g \\ u_1,\dots,u_g \in \mathbb{R}^p}} \sum_{l=1}^{g} \sum_{i \in C_l} \|x_i - u_l\|_2^2$$

When $C_1, \dots, C_g$ are given, then

$$\mu_\ell = \bar{x}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} x_i$$

$$\{x_1, \dots, x_n\} \longrightarrow \{1, \dots, n\}$$

$$\{1, \dots, n\} = \bigcup_{\ell=1}^{g} C_\ell \qquad C_\ell \cap C_{\ell'} = \phi \quad \ell \neq \ell'$$

Algorithm. (K-means clustering) – Lloyd's algorithm

[– Given centers $\mu_1, \dots, \mu_g$, each point x_i is assigned to the cluster ℓ given by

$$\ell = \arg\min_{j} \| x_i - \mu_j \|_2$$

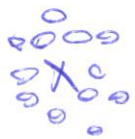
– Update the centers as $\mu_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} x_i$.]

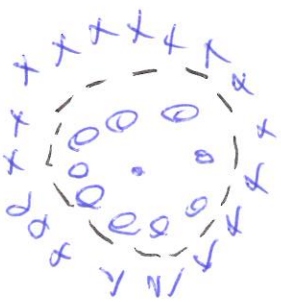
* Problem: we need to know the number of clusters g.

* We need a Euclidean space.

* the iterations might end up in a sub-optimal solution.

* You always end up with Convex clusters.






5.2.2. Spectral Clustering

$x_1, \dots, x_n \in \mathbb{R}^p$, $G = (V, E, W)$

$v_1 \quad v_n$

$w_{ij} = k_\varepsilon(\|x_i - x_j\|)$ — sym, positive, locality

e.g., $K_\varepsilon(u) = \exp\left(\frac{-1}{2\varepsilon} u^2\right)$

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

$D = \text{diag}(\deg(1), \dots, \deg(n))$

$$\deg(i) = \sum_j w_{ij}$$

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

$$\mathbb{P}(X_t = j \mid X_0 = i) = \left(M^t\right)_{ij}$$

$M = D^{-1} W \qquad S = D^{1/2} M D^{-1/2} = V \Lambda V^T$

$$\Rightarrow \quad M = \Phi \Lambda \Psi^T$$

$$M = \Phi \Lambda \Psi^T = \sum \lambda_k \Phi_k \Psi_k^T$$

(Φ, Ψ) - bionthonormal.

$$v_i \longrightarrow e_i^T M^t = \sum \lambda_k^t \Phi_{k,i} \Psi_k^T$$

Ψ_k = vectors

$$v_i \longrightarrow \begin{pmatrix} \lambda_1^t \Phi_{1,i} \\ \vdots \\ \lambda_n^t \Phi_{n,i} \end{pmatrix}$$

$$v_i = \begin{pmatrix} \lambda_2^t \Phi_{2,i} \\ \vdots \\ \lambda_{d+1}^t \Phi_{d+1,i} \end{pmatrix}$$

Spectral Clustering:

→ Use diffusion map to find the d- dimensional embedding of data.

→ Run K-means clustering on top of diffusion maps.

[Example next time for spectral clustering]

5.2.3. Hierarchical Clustering

Hierarchical clustering → agglomerative (bathamup)
→ divisive (top down)

Agglomerative clustering : assign a cluster to each point - reduce the number of clusters by iteratively merging smaller clusters into larger ones.

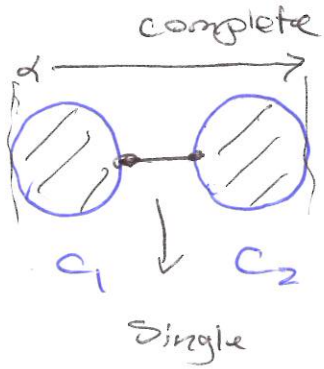
Divisive methods start with one large cluster containing all the datapoints and iteratively divide it into smaller clusters.

n objects $v_1, \dots, v_n$, pairwise dissimilarity $\Delta = (\delta_{ij})$ $i,j = 1, \dots, n$

$\delta_{ij} = \delta_{ji}$

A linkage function between two clusters C_1 and C_2 is defined as

$$d(C_1, C_2) = \begin{cases} \min_{i \in C_1, j \in C_2} \delta_{ij} & \text{single linkage} \\ \max_{i \in C_1, j \in C_2} \delta_{ij} & \text{complete linkage} \\ \frac{1}{|C_1||C_2|} \sum_{i \in C_1, j \in C_2} \delta_{ij} & \text{average linkage.} \end{cases}$$



Alg. Agglomerative Clustering

1. Initialize clusters as singletons
2. Initialize the set of availabe clusters.

$$S \leftarrow \{1, \ldots, n\}$$

Repeat.

- Pick the two most similar clusters to merge

$$(j, k) = \underset{j,k \in S}{\operatorname{argmin}}\ d(C_j, C_k)$$

- Merge C_k into C_j as $C_j \leftarrow C_j \cup C_k$
- Mark k as unavailable $S \leftarrow S - \{k\}$.
- Update dissimilarities $\delta(C_i, C_j)$
- continue until no more cluster is available.

6 Support Vector Machine.

Given a training set $\{(x_1, y_1), \dots (x_n, y_n)\}$

$x_i \in \mathbb{R}^p \quad y_i \in \{-1, +1\}$.

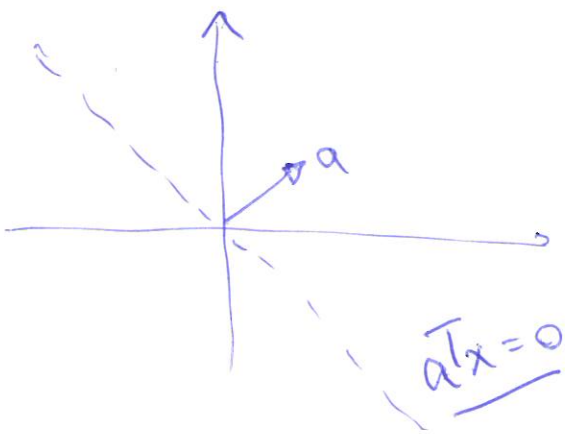
To select a hyperplane that separates the points into two classes and maximize the margins, i.e., the distance between the hyperplane and the closest points of the training set from each class.

6.1 Hyperplanes and Margins

$a \in \mathbb{R}^p, b \in \mathbb{R}$.

a) Given $a \in \mathbb{R}^p \quad \{x \in \mathbb{R}^p \mid a^T x = 0\}$

is a $(p-1)$-dimensional linear subspace orthogonal to a.



b) Given $a \in \mathbb{R}^p$, $b \in \mathbb{R}$

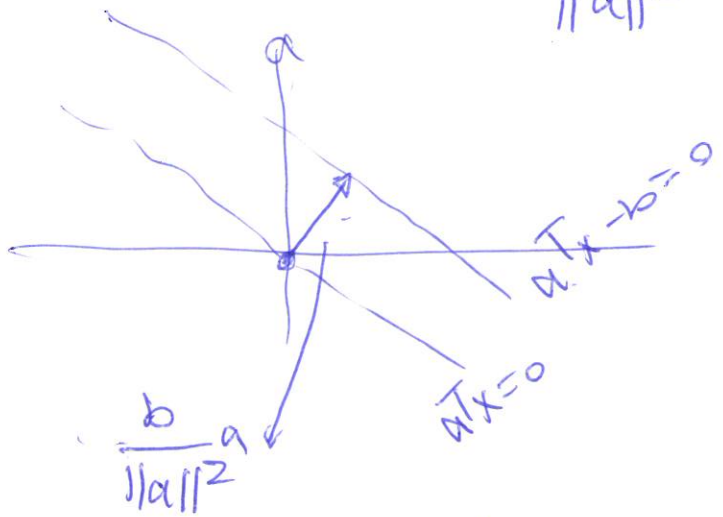
$$\{x \in \mathbb{R}^p \mid a^T x - b = 0\}$$

the linear space shifted by the vector $\frac{b}{\|a\|^2} a$.

$$a^T x - b = 0 \qquad a^T(x - x_0) = a^T x - b$$

$$x_0 = \frac{ba}{\|a\|^2} \leftarrow \text{vector}$$

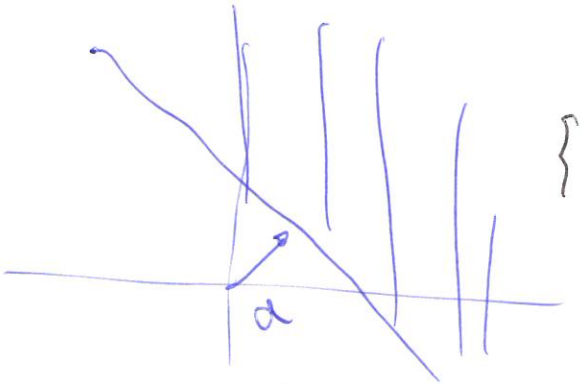
Hence, the linear space $\{a^T x = 0\}$ shifted by $\frac{b}{\|a\|^2} a$ is given by $\{a^T x - b = 0\}$ and is a hyperplane of distance $\frac{b}{\|a\|^2}$ from $\{a^T x = 0\}$



c) A half space of points lying on one side of the hyperplane $\{a^T x - b = 0\}$ is given by

$$\{x \in \mathbb{R}^p \mid a^T x \geq b\}.$$

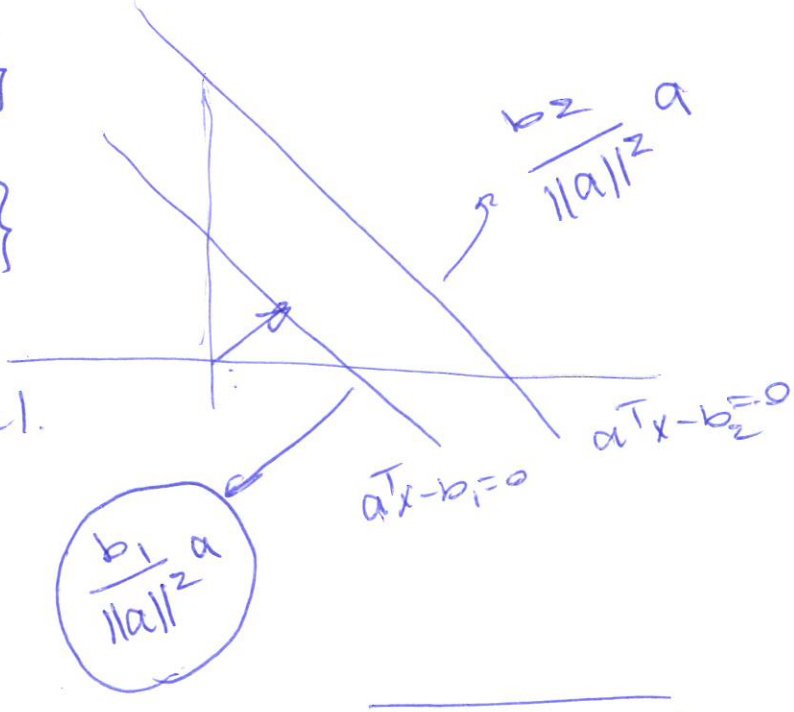
$\{a^T x \geq b\}$



$\{a^T x \geq b\}$

d) $H_1 = \{a^T x - b_1\}$

$H_2 = \{a^T x - b_2\}$

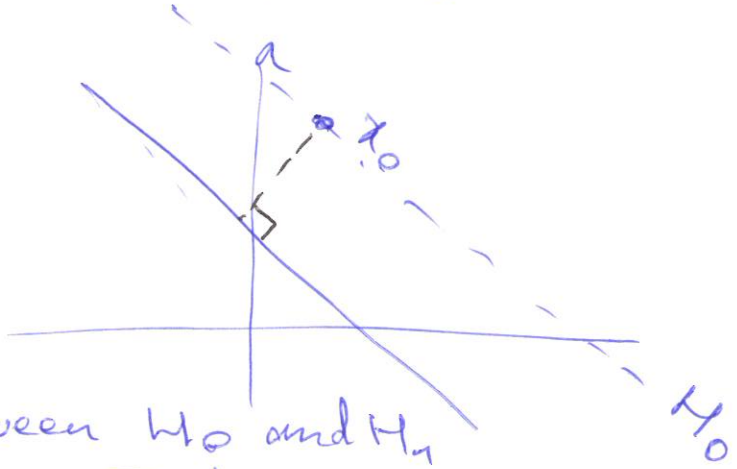


H_1, H_2 are parallel.

The distance between two hyperplanes is given by

$$\left\| \frac{b_1}{\|a\|^2} a - \frac{b_2}{\|a\|^2} a \right\| = \frac{|b_1 - b_2|}{\|a\|}$$

e) $H_1 = \{x \mid a^T x - b = 0\}$ $x_0 \in \mathbb{R}^p$



$H_0 = \{x \mid a^T x - b_0 = 0\}$

$a^T x_0 - b_0 = 0$

The distance between H_0 and H_1

$$\frac{|b_1 - b_0|}{\|a\|} = \frac{|b_1 - a^T x_0|}{\|a\|}$$