22-12-2021

$$d_j(\,),\ j=1,\dots,p$$

$$D(x_i,x_{i'}) = \sum_{j=1}^{p} w_j\, d_j(x_{ij},x_{i'j}) \ ,\quad i,i' \in \{1,\dots,N\}$$

$$\bar{D} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{i'=1}^{N}\sum_{j=1}^{p} w_j d_j(x_i,x_{i'}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{j=1}^{p} w_j d_j(x_i,x_{i'})$$

$$= \sum_{j=1}^{p} w_j \underbrace{\left[\frac{1}{N^2}\sum_{i,i'} d_j(x_i,x_{i'})\right]}_{\bar{d}_j} = \sum_{j=1}^{p} w_j \bar{d}_j = \bar{D}$$

Για να έχουν όλες οι μεταβλητές παρόμοια επίπτωση στην εκτίμηση της ανομοιότητας πρέπει $w_j \sim \dfrac{1}{\bar{d}_j}$

$X_1 \in [11,12] \qquad \bar{d}_1 < \bar{d}_2 \Rightarrow w_1 > w_2$

$X_2 \in [2,20]$

Για το $\bar{d}_j$, αν $d_j(x_i,x_{i'}) = (x_{ij}-x_{i'j})^2$

τότε $$\bar{d}_j = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}(x_{ij}-x_{i'j})^2 =$$

$$= \frac{1}{N^2}\sum_{i,i'}\left[(x_{ij}-\bar{x}_j)-(x_{i'j}-\bar{x}_j)\right]^2 =$$

$$= \frac{1}{N^2} \sum_{i} \sum_{i'} \left[(x_{ij} - \bar{x}_j)^2 + (x_{i'j} - \bar{x}_j)^2 - 2(x_{ij} - \bar{x}_j)(x_{i'j} - \bar{x}_j) \right]$$

$$= \frac{1}{N^2} \left[\underbrace{\sum_{i'} \sum_{i} (x_{ij} - \bar{x}_j)^2}_{N \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2} + \underbrace{\sum_{i} \sum_{i'} (x_{i'j} - \bar{x}_j)^2}_{N \sum_{i'=1}^{N} (x_{i'j} - \bar{x}_j)^2} - 2 \underbrace{\cancel{\sum_{i} (x_{ij} - \bar{x}_j)}}_{=0} \cdot \underbrace{\cancel{\sum_{i'} (x_{i'j} - \bar{x}_j)}}_{=0} \right]$$

$$= \frac{2}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 = \frac{2}{N} (N-1) \operatorname{var}(x_j) \approx 2 \operatorname{var}(x_j)$$

$$\boxed{w_j \sim \frac{1}{var_j}}$$

<u>Παρατήρηση</u> Αν κάνουμε κανονικοποίηση των μεταβλητών $\tilde{x}_j = \frac{x_j - \bar{x}_j}{\sqrt{var_j}}$, είναι ισοδύναμο με $w_j \sim \frac{1}{var_j}$

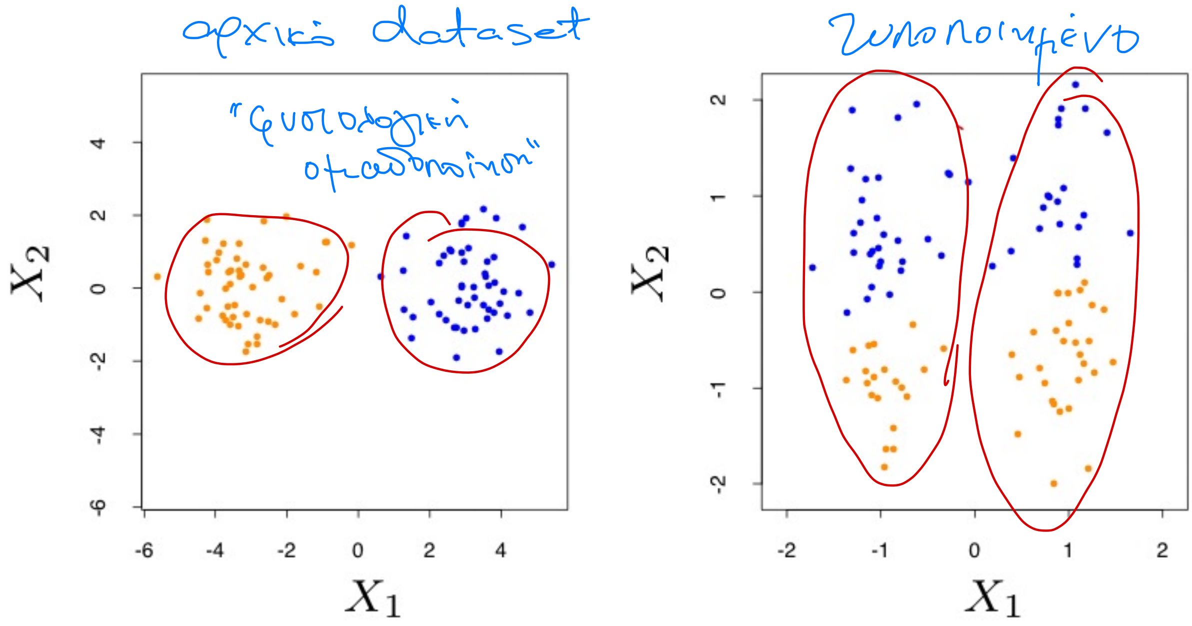


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with K=2) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights* $1/[2 \cdot \mathrm{var}(X_j)]$. *The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

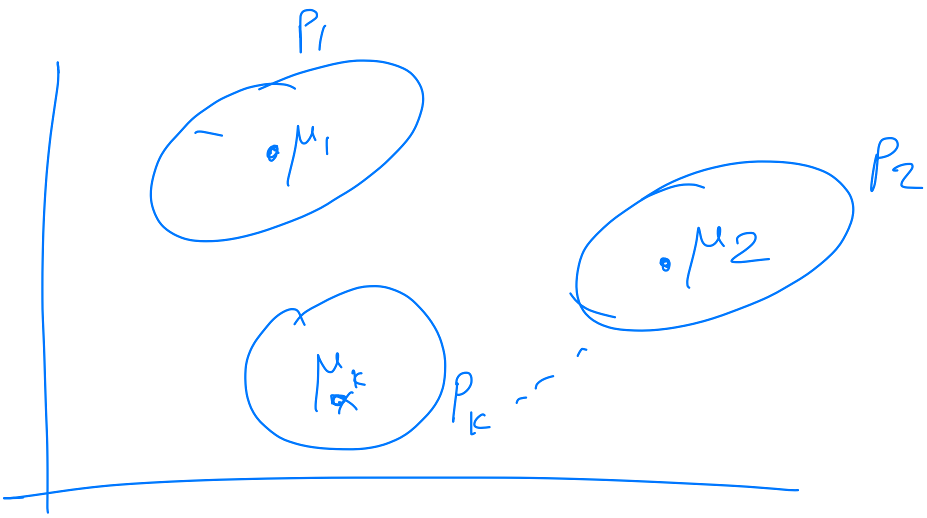
Clustering Algorithms

Χωρίζουν το dataset σε ομάδες (συστάδες - clusters) έτσι ώστε η ανομοιότητα μεταξύ παρατηρήσεων στην ίδια ομάδα να είναι μικρή κ' σε διαφορετικές ομάδες μεγάλη

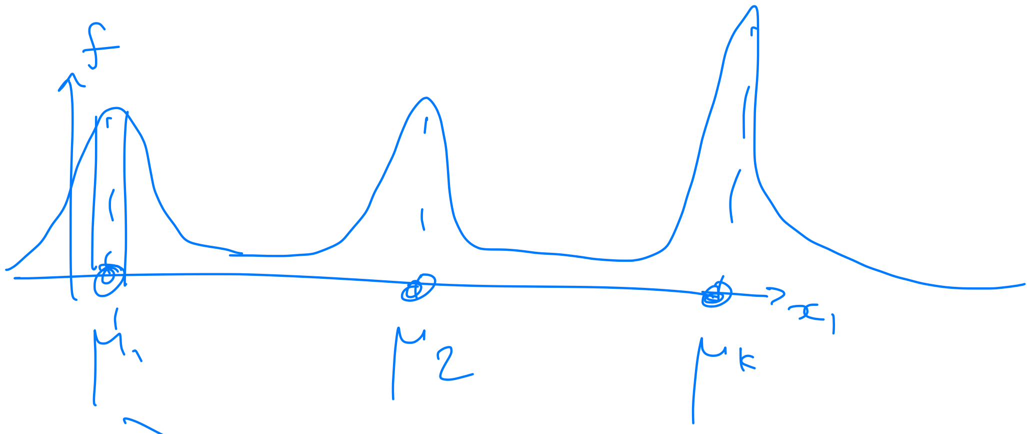
Συνδυαστικοί αλγόριθμοι — Δεν κάνουν υπόθεση πιθανοθεωρητικού μοντέλου

Μείξης — Υπόθεση ότι οι μεταβλητές ακολουθούν από κοινού κατανομή (μείξη επί μέρους κατανομών)

$$\underline{X}=(X_1,\dots,X_p) \to \begin{cases} F_1 & \text{με πιθ } p_1 \\ F_2 & \text{" " } p_2 \\ \vdots & \vdots \\ F_k & \text{" " } p_k \end{cases}$$



Αναζήτηση κορυφών (mode-seeking)



(modes) τοπικά μέγιστα της f

αναζήτηση modes της από κοινού κατανομής χωρίς παραμετρική υπόθεση για το κέντρο μείξης

Συνδυαστικοί αλγόριθμοι

Θεωρούμε K = αριθμός ομάδων (εξ αρχής σταθερό)

Συνάρτηση κωδικοποιητή (encoder function)

$C(i)$ = ομάδα στην οποία καταχωρείται η παρατ. i

$i = 1, \dots, N$

$$C : \{1, \dots, N\} \to \{1, \dots, K\}$$

"Loss function" : κριτήριο

$$W(C) = \frac{1}{2} \overbrace{\sum_{k=1}^{K}}\; \underbrace{\sum_{i : C(i)=k} \; \sum_{i' : C(i')=k} \overbrace{d(x_i, x_{i'})}^{d_{ii'}}}_{\text{άθροισμα αποστάσεων ανά δύο των παρατ. που ανατέθηκαν στην ομάδα } k.}$$

= within cluster point scatter

Θέλουμε το $W(C)$ μικρό

$B(C)$: between cluster scatter

$$= \frac{1}{2} \sum_{k=1}^{K} \; \sum_{i : C(i)=k} \; \sum_{i' : C(i') \neq k} d(x_i, x_{i'})$$

$$T(C) = W(C) + B(C) = \frac{1}{2} \sum_{k=1}^{K} \; \sum_{i : C(i)=k} \left[\underbrace{\sum_{i' : C(i')=k} d_{ii'} + \sum_{i' : C(i') \neq k} d_{ii'}}_{= \sum_{i'=1}^{N} d_{ii'}} \right]$$

$$\Rightarrow T(C) = \boxed{T = \frac{1}{2} \sum_{i} \sum_{i'} d_{ii'}} = \underbrace{\text{const w.r.t. } C}_{!!}$$

$$\Rightarrow \boxed{W(C) + \underline{B(C)} = T} \quad \forall C$$

$$\boxed{\min W(C) \iff \max B(C)}$$

K-means clustering algorithm

Επαναληπτικός αλγόριθμος

$\left[d_j(x_i, x_{i'}) = (x_{ij} - x_{i'j})^2 \quad \forall j\right]$

$$\boxed{d_{ii'} = \|x_i - x_{i'}\|^2}$$

Θέλουμε C, και τα κέντρα κάθε ομάδας

$m_1, \dots, m_K$.

[Μπορούμε να δείξουμε ότι για κάθε C

$$W(C) = \sum_{k=1}^{K} N_k \sum_{i: C(i)=k} \|x_i - \bar{x}_k\|^2 \Bigg]$$

centroids

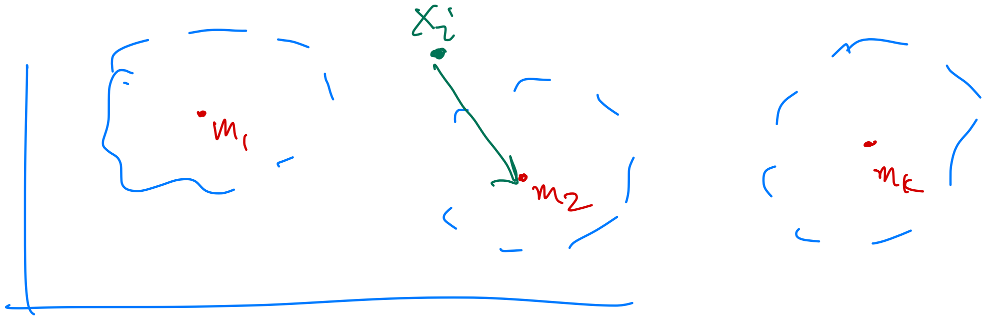
N_k : αρ. παρατ. που $C(i)=k$

---

(1) Αυθαίρετο C : αυθαίρετη κατάταξη σε ομάδες.

(2) Δεδομ. του C : θέτουμε τα $m_1, \dots, m_K$: $W(C)$ ελάχιστο

προφανώς $m_k = \bar{x}_k$, $k=1,\dots,K$

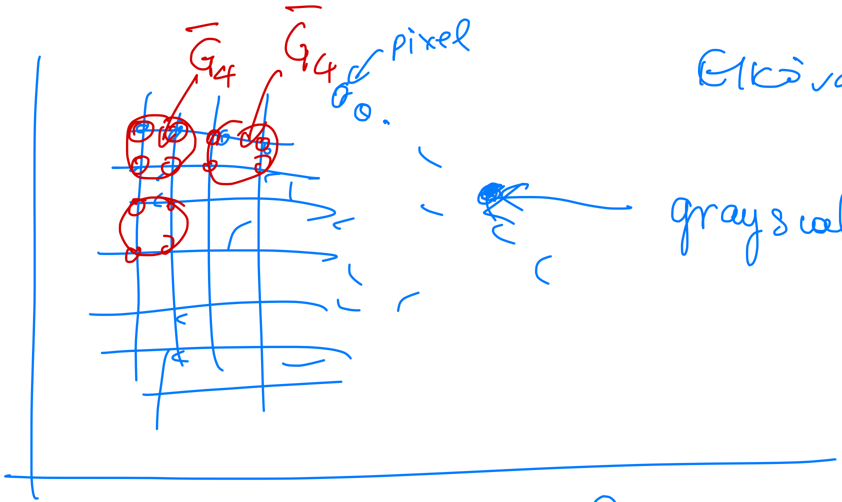


(3) Επιστρέφουμε στο 1: Με δεδομένα $m_1, \dots, m_K$
νέο C': $\forall i=1,\dots,N$: $C'(i)=k$ αν

$$\|x_i - m_k\|^2 = \min_{k'} \|x_i - m_{k'}\|^2$$

Εφαρμογή

Vector quantization για συμπίεση φωτογραφιών.



Εικόνα

grayscale 1 – 256
1 byte (8 bit)

NxN pixels $\Longrightarrow$ $\left[\begin{array}{c} G \\ \hline \vdots \\ \hline \end{array}\right]_{N^2}$ $\Longleftrightarrow$ $\left[\ \boxed{\ }^{\,G}\ \ \ \right]_{N \times N}$

$\Downarrow$

$\begin{bmatrix} \bar{G}_4 & \bar{G}_4 \end{bmatrix}_{\frac{N}{2} \times \frac{N}{2}}$ $\quad \frac{N^2}{4} \cdot 8$ bit

μέγεθος $N^2 \cdot 8$ bits.

$$\begin{array}{c} \\ N^2/4 \end{array}\begin{bmatrix} G \\ \vdots \\ G \end{bmatrix} \leftarrow \text{8 bit (256 τιμές)}$$

grayscales of pixel blocks (2x2)

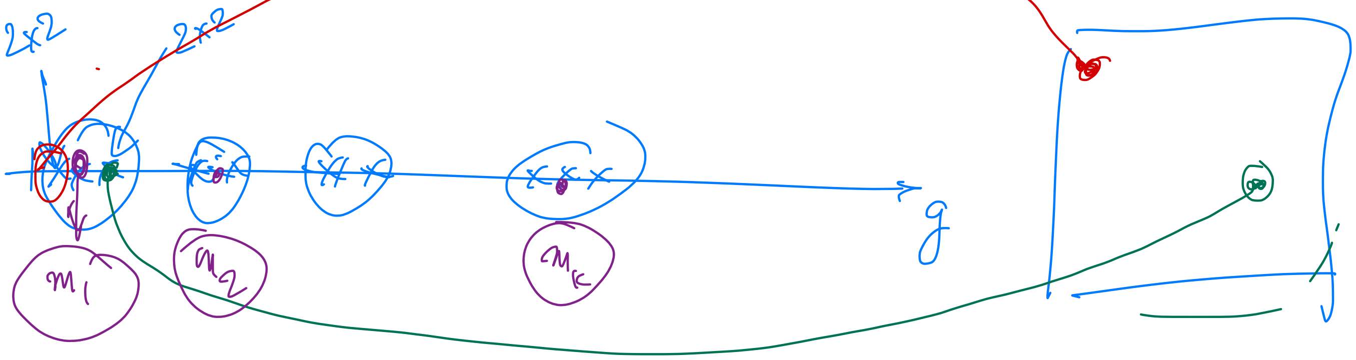
data set (p=1)

⇓

cluster analysis (k-means)

k: παράμετρος

Ομαδοποιούμε τις φωτεινότητες σε k ομάδες



$$N^2/4 \begin{bmatrix} 1 \\ 1 \\ 2 \\ k \\ 3 \\ 1 \end{bmatrix} \leftarrow \boxed{\log_2 k \text{ bit}}$$

$\rightarrow$ (k τιμές)

$K = 64$

$\log_2 = 6$

4/2/21 εξετάσεις

Μέχρι να σταματήσει ο αλγόριθμος σε κάθε βήμα $W(c)$ ελαττώνεται.

Σταματάει σε τοπικό ελάχιστο $W(C)$, ως προς C.

Επαναλαμβάνουμε 2,3 έως ότου

σε δύο διαδοχικά βήματα 3 δεν υπάρχει αλλαγή

στις ομάδες