Μάθημα 6

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΨΗΦΙΑΚΟ ΚΑΙ ΔΙΑΔΙΚΤΥΑΚΟ ΠΕΡΙΕΧΟΜΕΝΟ

Συσταδοποίηση Δεδομένων (Συνέχεια)

Ο Αλγόριθμος c-Means

- Οι αλγόριθμοι συσταδοποίησης είναι έτσι φτιαγμένοι, ώστε να διαμερίζουν τον χώρο των δεδομένων σε περιοχές και να αντιστοιχεί μια περιοχή σε κάθε συστάδα (δηλ. ομάδα)
- Ο αλγόριθμος c-Means ονομάζεται και αλγόριθμος Lloyd και ανήκει σε μια μεγάλη κατηγορία αλγορίθμων συσταδοποίησης που είναι γνωστοί ως αλγόριθμοι διαμέρισης (partitioning algorithms)

Βασικά Χαρακτηριστικά-Ιδιότητες του Αλγόριθμου c-Means

- Είναι επαναληπτική διαδικασία
- Χρησιμοποιεί την έννοια του **κέντρου της συστάδας** (cluster center) και στη συνέχεια κατατάσσει τα πολυδιάστατα δεδομένα (πλειάδες) ανάλογα με την απόστασή τους από τα κέντρα όλων των ομάδων
- Το **κέντρο μιας συστάδας** δεν είναι τίποτα άλλο από τη μέση τιμή των δεδομένων (πλειάδων) που ανήκουν στην ομάδα αυτή
- Ο αριθμός των ομάδων πρέπει να ορισθεί (να πάρει τιμή) από την αρχή της διαδικασίας
- Οι αρχικές τιμές των κέντρων των συστάδων (πριν την έναρξη της επαναληπτικής διαδικασίας) γίνεται με τυχαίο τρόπο

Ο Αλγόριθμος c-Means

Πλεονεκτήματα

- Είναι μία πολύ γρήγορη διαδικασία γιατί σε κάθε επανάληψη οι πράξεις που λαμβάνουν χώρα είναι απλές και γρήγορες
- Είναι πολύ εύκολη τόσο στην κατανόησή της όσο και στην υλοποίησή της
- Είναι η βάση για όλη την μοντέρνα θεωρία συσταδοποίησης

Μειονεκτήματα

- Ο εκ’ των προτέρων ορισμός του αριθμού των συστάδων αποτελεί έναν περιορισμό της μεθόδου, καθώς είτε πρέπει να τρέξουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των συστάδων είτε πρέπει με κάποιον άλλο τρόπο να έχουμε καταλήξει στον κατάλληλο αριθμό των συστάδων
- Η επιλογή τυχαίων των αρχικών τιμών των κέντρων των συστάδων οδηγεί σε επαναληπτική διαδικασία, το αποτέλεσμα της οποίας είναι πολύ ευαίσθητο στις τιμές αυτές. Αυτό γίνεται γιατί αν ξανα-εκτελέσουμε τον αλγόριθμο, οι αρχικές τιμές των κέντρων θα είναι διαφορετικές (επειδή επιλέγονται με τυχαίο τρόπο), και συνεπώς υπάρχει μεγάλη πιθανότητα και το αποτέλεσμα θα είναι διαφορετικό

Ο Αλγόριθμος c-Means

Θεωρία Συνόλων: Τομή και Ένωση Συνόλων

Αν κάποιο στοιχείο x ανήκει σε ένα σύνολο A τότε γράφουμε: $x \in A$

Αν κάποιο στοιχείο x δεν ανήκει σε ένα σύνολο A τότε γράφουμε: $x \notin A$

Έστω το σύνολο Α={2, 3, 4, 6, 10,12, 40} και το σύνολο Β={1, 3, 6, 12, 34, 50, 100}

$$10 \in A, \quad 10 \notin B$$

Ως τομή δύο συνόλων Α και Β ορίζεται το σύνολο με τα κοινά στοιχεία τους και συμβολίζεται ως: $A \cap B$

Αν τα σύνολα ΔΕΝ έχουν κανένα κοινό στοιχείο τότε γράφουμε: $A \cap B = \varnothing$

Ως ένωση δύο συνόλων Α και Β ορίζεται το σύνολο με όλα τα στοιχεία τους (τα κοινά συμπεριλαμβάνονται μόνο μία φορά) και συμβολίζεται ως: $A \cup B$

$$A \cap B = \{3, 6, 12\}$$

$$A \cup B = \{1, 2, 3, 4, 6, 10, 12, 34, 40, 50, 100\}$$

Ο Αλγόριθμος c-Means

Θεωρία Συνόλων: Η έννοια της Συνάρτησης Συμμετοχής ενός συνόλου

- Αν κάποιο στοιχείο *x* ανήκει σε ένα σύνολο *A* τότε λέμε ότι αυτό συμμετέχει στο σύνολο 100%
- Αλλιώς αν δεν ανήκει στο σύνολο λέμε ότι συμμετέχει 0% στο σύνολο αυτό
- Για να συμβολίσουμε την παραπάνω πρόταση χρησιμοποιούμε την συνάρτηση συμμετοχής

$$u_A(x) = \begin{cases} 100\% & \alpha\nu \ \ x \in A \\ 0\% & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

$$u_A(x) = \begin{cases} 1 & \alpha\nu \ \ x \in A \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

- Προγραμματιστικά υπολογίζεται με ένα if-then

Ο Αλγόριθμος c-Means

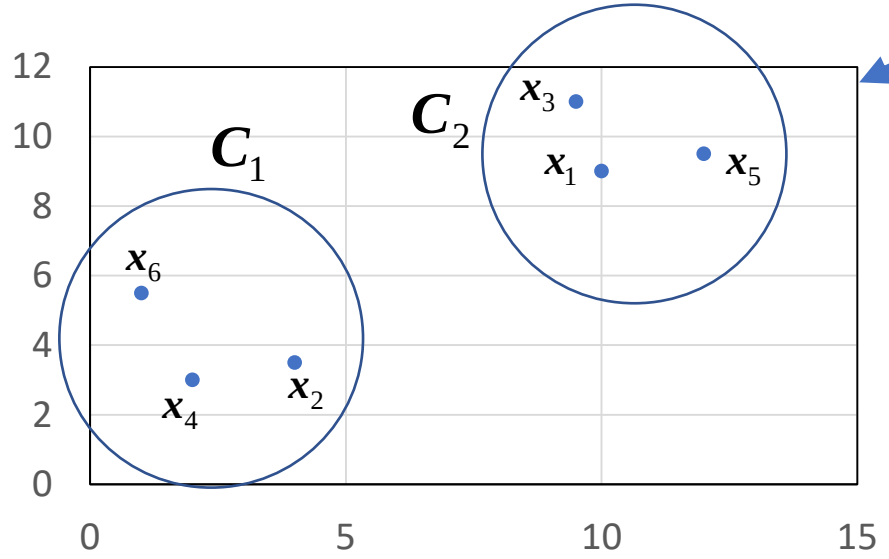
Επαναληπτικά Βήματα του Αλγόριθμου c-Means

- **Αρχικό Βήμα**
  Έχουμε γνωστό τον πίνακα δεδομένων, τον αριθμό συστάδων και μία αρχική εκτίμηση των κέντρων των συστάδων
- Βήμα 1
  Χρησιμοποιούμε τα κέντρα για υπολογίσουμε τις συναρτήσεις συμμετοχής των δεδομένων στις συστάδες
- Βήμα 2
  Χρησιμοποιούμε τις συναρτήσεις συμμετοχής που υπολογίστηκαν στο προηγούμενο βήμα για υπολογίσουμε τα νέα κέντρα των συστάδων.
  Αν τα νέα κέντρα είναι πολύ κοντά στα παλιά τότε ο αλγόριθμος σταματέι. Αλλιώς πηγαίνει στο Βήμα 1.
  Τέλος

Ο Αλγόριθμος c-Means

Α. Η χρήση της Συνάρτησης Συμμετοχής στον Προσδιορισμό των Κέντρων των Συστάδων

	x_1	x_2
x_1	10	9
x_2	4	3.5
x_3	9.5	11
x_4	2	3
x_5	12	9.5
x_6	1	5.5



Διαμερισμός

$N = 6$

$c = 2$

$p = 2$

	C_1	C_2
x_1	0	1
x_2	1	0
x_3	0	1
x_4	1	0
x_5	0	1
x_6	1	0

$$u_1(x_1)=0 \quad u_2(x_1)=1$$
$$u_1(x_2)=1 \quad u_2(x_2)=0$$
$$u_1(x_3)=0 \quad u_2(x_3)=1$$
$$u_1(x_4)=1 \quad u_2(x_4)=0$$
$$u_1(x_5)=0 \quad u_2(x_5)=1$$
$$u_1(x_6)=1 \quad u_2(x_6)=0$$

$$u_{11}=0 \quad u_{12}=1$$
$$u_{21}=1 \quad u_{22}=0$$
$$u_{31}=0 \quad u_{32}=1$$
$$u_{41}=1 \quad u_{42}=0$$
$$u_{51}=0 \quad u_{52}=1$$
$$u_{61}=1 \quad u_{62}=0$$

Πίνακας συναρτήσεων συμμετοχής του διαμερισμού U

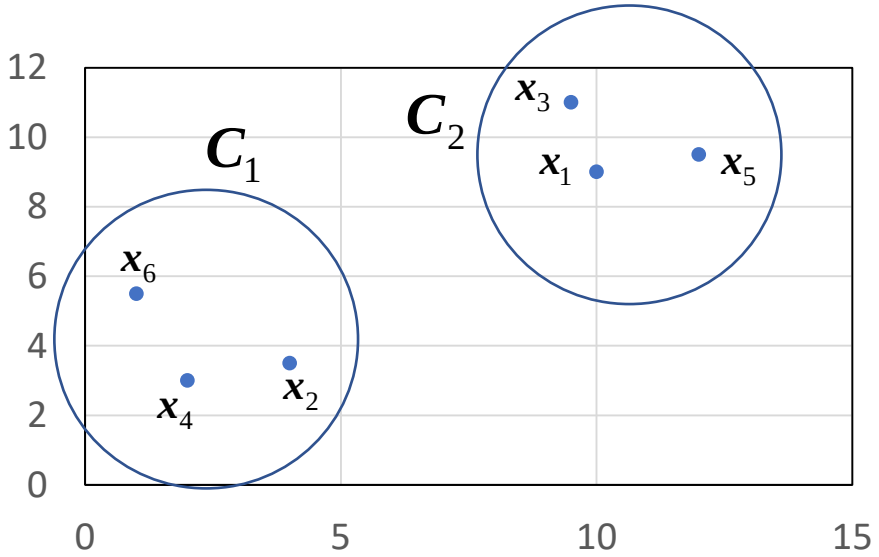
0	1
1	0
0	1
1	0
0	1
1	0

$$U = [u_{ki}]$$
$$k = 1, 2, \ldots, N$$
$$i = 1, 2, \ldots, c$$

Ο Αλγόριθμος c-Means

Α. Η χρήση της Συνάρτησης Συμμετοχής στον Προσδιορισμό των Κέντρων των Συστάδων

	x_1	x_2
x_1	10	9
x_2	4	3.5
x_3	9.5	11
x_4	2	3
x_5	12	9.5
x_6	1	5.5



$$N = 6 \qquad k = 1, 2, \ldots, \mathrm{N}$$
$$c = 2 \qquad i = 1, 2, \ldots, c$$
$$p = 2 \qquad j = 1, 2, \ldots, p$$

Τα κέντρα των ομάδων συμβολίζονται ως:

$$\boldsymbol{v}_1 = [v_{11}\ v_{12}] \qquad \boldsymbol{v}_2 = [v_{21}\ v_{22}]$$

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \\ u_{61} & u_{62} \end{bmatrix} = \boldsymbol{U} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$v_{11} = \frac{x_{21} + x_{41} + x_{61}}{3} = \frac{4 + 2 + 1}{3} = 2.3$$

$$v_{11} = \frac{u_{21}x_{21} + u_{41}x_{41} + u_{61}x_{61}}{3} = \frac{1 \cdot 4 + 1 \cdot 2 + 1 \cdot 1}{3} = 2.3$$

$$v_{11} = \frac{u_{11}x_{11} + u_{21}x_{21} + u_{31}x_{31} + u_{41}x_{41} + u_{51}x_{51} + u_{61}x_{61}}{u_{11} + u_{21} + u_{31} + u_{41} + u_{51} + u_{61}} =$$

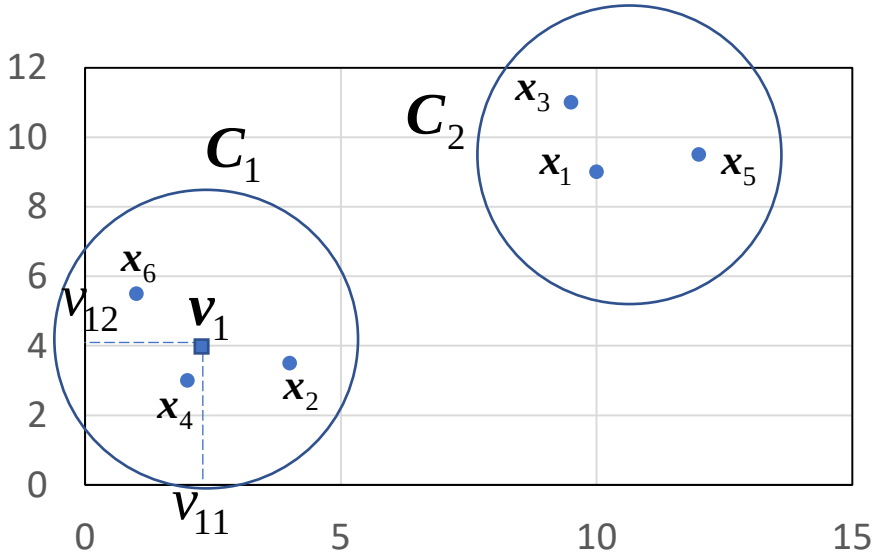
$$= \frac{0 \cdot 10 + 1 \cdot 4 + 0 \cdot 9.5 + 1 \cdot 2 + 0 \cdot 12 + 1 \cdot 1}{0 + 1 + 0 + 1 + 0 + 1} = \frac{1 \cdot 4 + 1 \cdot 2 + 1 \cdot 1}{3} = 2.3$$

$$v_{11} = \frac{\sum_{k=1}^{6} u_{ki}\, x_{k1}}{\sum_{k=1}^{6} u_{ki}}$$

Ο Αλγόριθμος c-Means

Α. Η χρήση της Συνάρτησης Συμμετοχής στον Προσδιορισμό των Κέντρων των Συστάδων

	x_1	x_2
x_1	10	9
x_2	4	3.5
x_3	9.5	11
x_4	2	3
x_5	12	9.5
x_6	1	5.5



$$N = 6 \qquad k = 1, 2, \ldots, \mathrm{N}$$
$$c = 2 \qquad i = 1, 2, \ldots, c$$
$$p = 2 \qquad j = 1, 2, \ldots, p$$

Τα κέντρα των ομάδων συμβολίζονται ως:

$$\boldsymbol{v}_1 = \begin{bmatrix} v_{11} & v_{12} \end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix} v_{21} & v_{22} \end{bmatrix}$$

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \\ u_{61} & u_{62} \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$v_{12} = \frac{x_{22} + x_{42} + x_{62}}{3} = \frac{3.5 + 3 + 5.5}{3} = 4$$

$$v_{12} = \frac{u_{21}x_{22} + u_{41}x_{42} + u_{61}x_{62}}{3} = \frac{1 \cdot 3.5 + 1 \cdot 3 + 1 \cdot 5.5}{3} = 4$$

$$v_{12} = \frac{u_{11}x_{12} + u_{21}x_{22} + u_{31}x_{32} + u_{41}x_{42} + u_{51}x_{52} + u_{61}x_{62}}{u_{11} + u_{21} + u_{31} + u_{41} + u_{51} + u_{61}} =$$

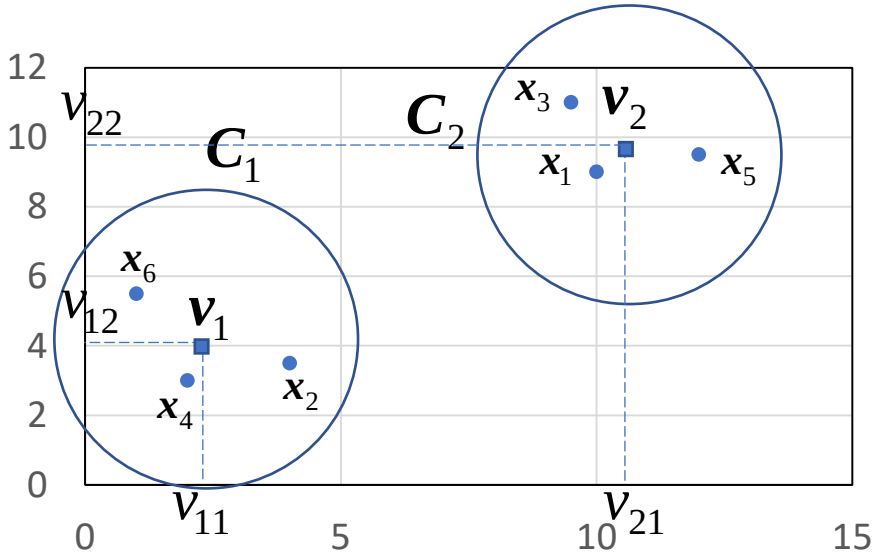
$$= \frac{0 \cdot 9 + 1 \cdot 3.5 + 0 \cdot 11 + 1 \cdot 3 + 0 \cdot 9.5 + 1 \cdot 5.5}{0 + 1 + 0 + 1 + 0 + 1} = \frac{1 \cdot 3.5 + 1 \cdot 3 + 1 \cdot 5.5}{3} = 4$$

$$v_{12} = \frac{\sum_{k=1}^{6} u_{k1} x_{k2}}{\sum_{k=1}^{6} u_{k1}}$$

Ο Αλγόριθμος c-Means

Α. Η χρήση της Συνάρτησης Συμμετοχής στον Προσδιορισμό των Κέντρων των Συστάδων

	x_1	x_2
x_1	10	9
x_2	4	3.5
x_3	9.5	11
x_4	2	3
x_5	12	9.5
x_6	1	5.5



$N = 6$ $\quad k = 1, 2, \ldots, \mathrm{N}$

$c = 2$ $\quad i = 1, 2, \ldots, c$

$p = 2$ $\quad j = 1, 2, \ldots, p$

	x_1	x_2
v_1	2.3	4
v_2	10.5	9.83

Τα κέντρα των ομάδων συμβολίζονται ως:

$$\boldsymbol{v}_1 = [v_{11} \;\; v_{12}] \qquad \boldsymbol{v}_2 = [v_{21} \;\; v_{22}]$$

$\boldsymbol{U}$

u_{11}	u_{12}
u_{21}	u_{22}
u_{31}	u_{32}
u_{41}	u_{42}
u_{51}	u_{52}
u_{61}	u_{62}

$=$

$\boldsymbol{U}$

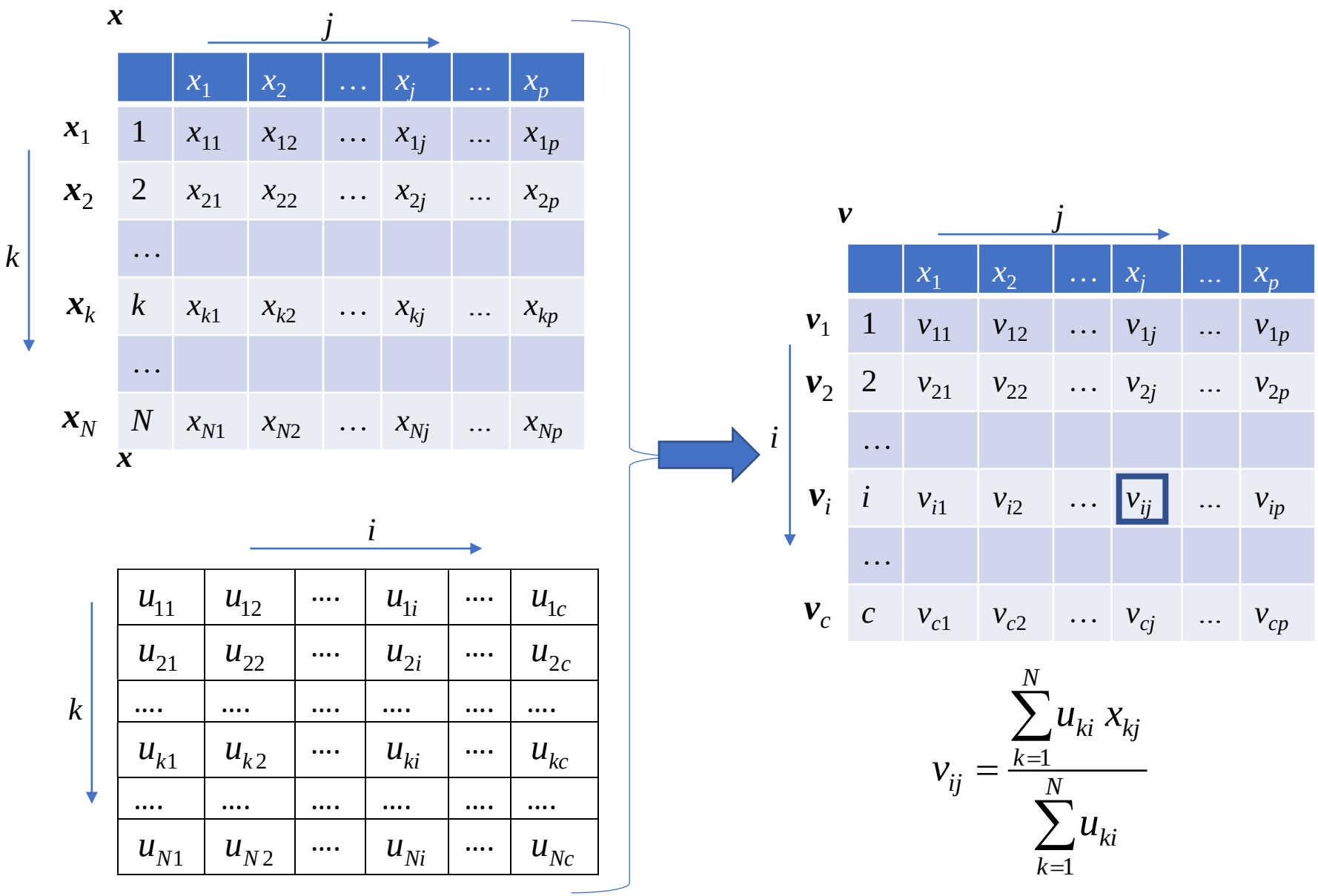
0	1
1	0
0	1
1	0
0	1
1	0

$$v_{21} = \frac{\sum_{k=1}^{6} u_{k2}\, x_{k1}}{\sum_{k=1}^{6} u_{k2}} = \frac{1\cdot 10 + 0\cdot 4 + 1\cdot 9.5 + 0\cdot 2 + 1\cdot 12 + 0\cdot 1}{1+0+1+0+1+0} = \frac{10 + 9.5 + 12}{3} = 10.5$$

$$v_{22} = \frac{\sum_{k=1}^{6} u_{k2}\, x_{k2}}{\sum_{k=1}^{6} u_{k2}} = \frac{1\cdot 9 + 0\cdot 3.5 + 1\cdot 11 + 0\cdot 3 + 1\cdot 9.5 + 0\cdot 5.5}{1+0+1+0+1+0} = \frac{9 + 11 + 9.5}{3} = 9.83$$

Ο Αλγόριθμος c-Means

Α. Η χρήση της Συνάρτησης Συμμετοχής στον Προσδιορισμό των Συστάδων Πολυδιάστατα Δεδομένα



x		x_1	x_2	…	x_j	…	x_p
x_1	1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
x_2	2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
	…						
x_k	k	x_{k1}	x_{k2}	…	x_{kj}	…	x_{kp}
	…						
x_N	N	x_{N1}	x_{N2}	…	x_{Nj}	…	x_{Np}

u_{11}	u_{12}	….	u_{1i}	….	u_{1c}
u_{21}	u_{22}	….	u_{2i}	….	u_{2c}
….	….	….	….	….	….
u_{k1}	u_{k2}	….	u_{ki}	….	u_{kc}
….	….	….	….	….	….
u_{N1}	u_{N2}	….	u_{Ni}	….	u_{Nc}

v		x_1	x_2	…	x_j	…	x_p
v_1	1	v_{11}	v_{12}	…	v_{1j}	…	v_{1p}
v_2	2	v_{21}	v_{22}	…	v_{2j}	…	v_{2p}
	…						
v_i	i	v_{i1}	v_{i2}	…	v_{ij}	…	v_{ip}
	…						
v_c	c	v_{c1}	v_{c2}	…	v_{cj}	…	v_{cp}

$$v_{ij} = \frac{\sum_{k=1}^{N} u_{ki}\, x_{kj}}{\sum_{k=1}^{N} u_{ki}}$$

Συμπέρασμα

Αν ξέρουμε τον πίνακα των συναρτήσεων συμμετοχής τότε τα κέντρα των συστάδων υπολογίζονται με βάση τον παρακάτω κώδικα

```
for i=1:c
    for j=1:p
        s1=0;
        s2=0;
        for k=1:N
            s1=s1+u(k,i)*x(k,j);
            s2=s2+u(k,i);
        endfor
    endfor
    v(i,j)=s1/s2;
endfor
```

ΚΑΛΟ ΑΠΟΓΕΥΜΑ