SI 221
Annales corrigées en classification automatique
Octobre 2007

Sujet

On considère 4 points A, B, C, D dans le plan, déterminés par leurs coordonnées cartésiennes :

$$A = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 5 \end{pmatrix} \quad C = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \quad D = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

On cherche à calculer la “matrice d'écart” $M = (m_{ij})$ (i.e. telle que m_{ij} représenta la distance entre le point i et le point j) dans différentes normes, et à utiliser cette matrice pour en déduire un arbre de classification
1) En se dotant de la métrique fondée sur la distance L^1, calculer la matrice d'écart et proposer un arbre T_1 de classification fondé sur l'ultramétrique sous dominante.
2) Avec la distance L^∞, calculer la matrice d'écart et proposer un arbre T_2 de classification.
3) On envisage une classification automatique en deux classes. Quelles sont les deux classes envisageables si l'on s'inspire de l'arbre T_2 ? L'algorithme des k-moyennes donne-t-il un résultat satisfaisant ?
4) Même question avec l'arbre T_1 (on choisira un cas où les deux classes n'ayant pas le même nombre de représentants).
5) Si l'on cherche un classifieur minimisant le critère de l'erreur quadratique, lequel de ces deux classifieurs fondés sur l'algorithme des k-moyennes choisirez vous ?

Corrigé

On considère 4 points A, B, C, D dans le plan, déterminés par leurs coordonnées cartésiennes :

$$A = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ 5 \end{pmatrix} \quad C = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \quad D = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

On cherche à calculer la “matrice d'écart” $M = (m_{ij})$ (i.e. telle que m_{ij} représenta la distance entre le point i et le point j) dans différentes normes, et à utiliser cette matrice pour en déduire un arbre de classification fondé sur l'ultramétrique sous dominante.

1. En se dotant de la métrique fondée sur la distance L^1, calculer la matrice d'écart et proposer un arbre T_1 de classification.

	A	B	C
B	5		
C	7	4	
D	6	11	7

L'arbre se construit comme suit :

– *on regroupe B et C (ecart minimum 4) : $\{B, C\}$.*
– *on recalcule la matrice d'écart.*
– *on regroupe A et $\{B, C\}$*
– *on regroupe tout.*

Tracez vous même l'arbre correspondant.

2. Avec la distance L^∞, calculer la matrice d'écart et proposer un arbre T_2 de classification.

	A	B	C
B	5		
C	4	3	
D	3	8	7

L'arbre se construit comme suit :

– *on regroupe B et C (ecart minimum 3) : $\{B, C\}$.*
– *on regroupe AUSSI A et D (ecart minimum 3) : $\{A, D\}$.*
– *on regroupe tout.*

Tracez vous même l'arbre correspondant.

3. On envisage une classification automatique en deux classes. Quelles sont les deux classes envisageables si l'on s'inspire de l'arbre T_2 ?

   On peut prendre comme classe C_1 *l'ensemble* $\{A,D\}$, *de centre de gravité* G_1, *et comme classe* C_2 *l'ensemble* $\{B,C\}$, *de centre de gravité* G_2

$$G_1 = \begin{pmatrix} 1.5 \\ -1.5 \end{pmatrix} \quad G_2 = \begin{pmatrix} 1.5 \\ 4.5 \end{pmatrix}$$

   L'algorithme des k-moyennes donne-t-il un résultat satisfaisant ?

   On vérifie que l'on ne change pas les étiquettes en faisant tourner l'algorithme (attention, il faut bien entendu utiliser la norme L^2 *dans l'algorithme, sinon cela n'a pas de sens) . La classification est stable.*

4. Même question avec l'arbre T_1 (on choisira un cas où les deux classes n'ayant pas le même nombre de représentants).

   On peut prendre comme classe C_1 *l'ensemble* $\{A,B,C\}$, *de centre de gravité* G_1, *et comme classe* C_2 *l'ensemble* $\{D\}$, *de centre de gravité* G_2

$$G_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad G_2 = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

   On vérifie que l'on ne change pas les étiquettes en faisant tourner l'algorithme. La classification est stable.

5. Si l'on cherche un classifieur minimisant le critère de l'erreur quadratique, lequel de ces deux classifieurs fondés sur l'algorithme des k-moyennes choisirez vous ?

   Les erreurs quadratiques (dispersion intra-classe) donnent dans le cas T_2 *14 et dans le cas* T_1 *20. On préfèrera donc l'arbre donné par la distance* L^∞. *Veifiez d'ailleurs au passage que l'on maximise aussi la dispersion inter-classes (démonstration dans le cours...).*