Anonimizzazione.

Un data-base contiene numerosi record in cui sono stati registrati i valori (sensibili) di alcuni attributi riferiti ad alcune persone. Si tratta di attributi rappresentati da numeri razionali, normalizzati tutti tra 0 e 100. Per impedire che dai dati si possa risalire alle persone, si vuole sostituire il set dei record individuali con un numero ridotto di record che rappresentano "tipi medi" ai quali le persone classificate assomigliano. Questa operazione corrisponde a partizionare il set dei record individuali in clusters di individui simili e sostituire ogni cluster con un "tipo" che lo rappresenta.

Per preservare le proprietà dell'anonimizzazione si vuole che ogni cluster contenga almeno K individui, essendo K un parametro specificato in ingresso.

La definizione dei "tipi" deve ottimizzare una misura della loro rappresentatività, cioè della loro somiglianza con gli individui assegnati al loro cluster. Si propongono quattro criteri, tutti basati sulla distanza tra punti nello spazio Euclideo degli attributi:

1. Minimizzare la somma delle distanze tra ogni individuo ed il rappresentante del suo cluster.
2. Minimizzare la massima distanza tra un individuo ed il rappresentante del suo cluster.
3. Minimizzare la somma dei quadrati delle distanze tra ogni individuo ed il rappresentante del suo cluster.
4. Minimizzare la massima differenza in valore assoluto tra un attributo di un individuo e lo stesso attributo del rappresentante del suo cluster.

Formulare il problema, classificarlo e risolvere l'esempio descritto dai dati nel file `ANONIMIZZAZIONE.TXT`.

Dati.

Le persone sono 30. Gli attributi sono 5. K = 4.

Persone	A1	A2	A3	A4	A5
1	54	52	100	96	100
2	32	56	36	60	55
3	78	58	7	45	81
4	89	91	46	6	13
5	28	14	66	43	37
6	63	87	82	88	42
7	59	84	92	40	25
8	62	71	46	74	22
9	73	76	85	5	82
10	27	80	30	35	52
11	57	90	58	49	37
12	50	43	72	96	56
13	4	86	98	57	41
14	6	63	49	25	3
15	80	63	5	80	35
16	39	79	74	70	49
17	20	97	7	1	58
18	58	92	99	59	70
19	4	45	3	18	16
20	40	26	5	42	8
21	80	17	68	44	65
22	46	49	15	26	53
23	73	10	7	99	7
24	64	15	33	22	52
25	48	12	46	56	17
26	8	67	51	31	91
27	51	41	85	67	63
28	4	7	92	45	89
29	64	28	47	25	54
30	23	66	93	27	40

Tabella 1: Persone e loro attributi.

Soluzione.

Detto N l'insieme delle persone e n il numero delle persone, il numero di clusters non può essere maggiore di n/K, dato che ciascun cluster deve contenere almeno K persone. Tutte le funzioni obiettivo non possono che migliorare quando il numero di clusters aumenta. Sia quindi $c = \lfloor \frac{n}{K} \rfloor$ il numero di clusters e sia C l'insieme dei c clusters.

Detto T l'insieme degli attributi, sia a_{it} il valore dell'attributo $t \in T$ della persona $i \in N$ (dato) e sia x_{jt} il valore dell'attributo $t \in T$ per il rappresentante del cluster $j \in C$ (variabile continua). Le distanze sono quindi

$$d_{ij} = \sqrt{\sum_{t \in T} (x_{jt} - a_{it})^2} \ \ \forall i \in N, \forall j \in C.$$

Il problema richiede l'assegnamento delle persone ai clusters, il che può essere rappresentato da variabili binarie w_{ij}: la variabile w_{ij} vale 1 se e solo se la persona $i \in N$ è assegnata al cluster $j \in C$.

I vincoli di assegnamento richiedono che ogni persona sia assegnata ad un cluster:

$$\sum_{j \in C} w_{ij} = 1 \ \ \forall i \in N.$$

Ogni cluster, inoltre, deve contenere almeno K persone:

$$\sum_{i \in N} w_{ij} \geq K \ \ \forall j \in C.$$

Gli obiettivi, tutti da minimizzare, si formulano come segue

1. $z_1 = \sum_{i \in N, j \in C} d_{ij} w_{ij}$
2. $z_2 = \alpha$ con il vincolo $\alpha \geq d_{ij} w_{ij} \ \ \forall i \in N, \forall j \in C$
3. $z_3 = \sum_{i \in N, j \in C} d_{ij}^2 w_{ij}$
4. $z_4 = \beta$ con il vincolo $-\beta \leq (a_{it} - x_{jt}) w_{ij} \leq \beta \ \ \forall i \in N, \forall j \in C, \forall t \in T$.

Poiché i valori degli obiettivi dipendono da distanze Euclidee e il modello richiede variabili binarie, il problema è di programmazione non-lineare 0-1.

Fissati i valori delle variabili w per enumerazione implicita, ogni sottoproblema di ottimizzazione non-lineare relativo ad un singolo cluster è convesso e quindi risolvibile all'ottimo garantito.

L'unicità della soluzione ottima non è garantita, poiché potrebbero esistere partizioni diverse con lo stesso valore.