POLITECNICO DI TORINO
ESAME DI STATO PER L'ABILITAZIONE ALLA PROFESSIONE DI INGEGNERE
RAMO INFORMATICA
VECCHIO ORDINAMENTO

I SESSIONE 2011

Prova scritta del 15 giugno 2011

Premessa.
Molti fenomeni fisici vengono rappresentati mediante un insieme di misure di alcune caratteristiche distintive (*features*), organizzate in un vettore. Si pensi, ad esempio, ai fenomeni atmosferici, descritti dall'insieme di temperature, pressione, umidità, ecc. , o ad un segnale elettrico stazionario, descritto dai punti dello spettro, e così via. Questi vettori possono essere considerati dei punti in uno spazio euclideo multidimensionale.
In molte applicazioni (pattern-recognition, data-base, ecc.), è richiesto di analizzare come questi dati si aggregano (clustering). A tal fine la letteratura propone algoritmi come k-Means e il più sofisticato ISODATA, qui di seguito descritto.

Algoritmo Isodata

L'algoritmo Isodata (Iterative Self-Organizing Data Analysis Technique A) è simile alla procedura K-Means: si aggiungono in più delle procedure euristiche (importante, tra le altre, quella detta di *split and merge*).

All'inizio si fissa un numero di cluster iniziali N_c (non necessariamente quello che si desidera come definitivo) e i loro centri (arbitrari) $\bar{z}_1,\bar{z}_2,...,\bar{z}_{N_c}$. Ad esempio, si possono scegliere come centri di cluster i primi N_c vettori.

Passo 1) Si specificano i seguenti parametri di processo:
K: numero dei centri di cluster desiderati o auspicati (in tutti i casi, si può prefissare il numero massimo di cluster)
ϑ_N : numero di campioni con cui confrontare il numero di campioni in un cluster
ϑ_S : parametro di deviazione standard
ϑ_C : parametro di aggregazione di un blocco
L: massimo numeri di coppie di centri di cluster che possono essere aggregate
I: numero di iterazioni permesse

Passo 2) Si distribuiscono gli *N* campioni $\{\bar{x}_1,...,\bar{x}_N\}$ tra gli attuali centri di cluster in base alla relazione: $\bar{x} \in S_J$ se $\left\|\bar{x}-\bar{z}_j\right\| < \left\|\bar{x}-\bar{z}_i\right\|$, con $i=1,2,...,N_c; i \neq j$
In pratica, si attribuisce un vettore al centro di cluster più vicino.

Passo 3) Si scaricano i cluster con meno di ϑ_N campioni; formalmente:
Se per un *j* è $N_j < \vartheta_N$, allora si elimina S_j e si riduce N_c di 1.

Passo 4) Si ricalcolano i centri dei cluster $\bar{z}_j$, $j = 1,2,...,N_c$, ponendoli uguali alla media dei campioni in ciascun cluster, $\bar{z}_j = \frac{1}{N_j}\sum_{\bar{x}\in S_j}\bar{x}$, con $j = 1,2,...,N_c$

Passo 5) Si calcola la distanza media $\overline{D}_j = \frac{1}{N_j}\sum_{\bar{x}\in S_j}\left\|\bar{x}-\bar{z}_j\right\|$, con $j = 1,2,...,N_c$

Passo 6) Si calcola la media pesata delle distanze medie, cioè $\overline{D} = \frac{1}{N}\sum_{j=1}^{N_c} N_j \overline{D}_j$

Passo 7) *a*) Se questa è l'ultima iterazione porre $\vartheta_C = 0$ e andare al **passo 11**).
b) Se $N_c \le k/2$ (pochi cluster rispetto al previsto), andare al **passo 8**) (fase di *split*).
c) Se questa è un'iterazione pari oppure $N_c \ge 2k$ (troppi cluster), andare al **passo 11**) (fase di *merge*); altrimenti continuare (*Passo 14*).

Passo 8) Calcolare il vettore di deviazioni standard $\bar{\sigma}_j = (\sigma_{1j},\sigma_{2j},...,\sigma_{nj})$ per ciascun cluster, secondo la relazione $\sigma_{ij} = \sqrt{\frac{1}{N_j}\sum_{\bar{x}\in S_j}(x_{ik}-z_{ij})^2}$, $i = 1,2,...,n$,

$j = 1,2,...,N_c$ dove :

- n è la dimensionalità dei vettori $\bar{x}$
- x_{ik} è la i-esima componente del k-esimo campione di S_j
- z_{ij} è la i-esima componente del j-esimo centro di cluster

In pratica, il vettore $\bar{\sigma}_j$ rappresenta la deviazione standard dei campioni in S_j lungo gli assi principali delle coordinate.

Passo 9) Si cerca la componente massima di ciascun $\bar{\sigma}_j$, con $j = 1,2,...,N_c$ e si denota come $\sigma_{j\,max}$

Passo 10) Se per ogni $\sigma_{j\,max}$ ci si trova di fronte a questa situazione:

$$\sigma_{j\,max} > \vartheta_S \text{ e } \begin{cases} a)\,\overline{D}_j > \overline{D} \text{ e } N_j > 2(\vartheta_N + 1) \\ \qquad\qquad oppure \\ b)\,N_C \le K/2 \end{cases}$$

si spezza $\bar{z}_j$ in due centri di cluster $\bar{z}_j^{+}$ e $\bar{z}_j^{-}$.

Il nuovo centro di cluster $\bar{z}_j^{+}$ è ricavato da $\bar{z}_j$ sommando una data quantità γ_j alla componente di $\bar{z}_j$ che corrisponde alla massima componente di $\bar{\sigma}_j$;
per $\bar{z}_j^{-}$ invece γ_j viene sottratto. γ_j in pratica viene calcolato come $\gamma_j = k\sigma_{j\,max}$ con $0 \le k \le 1$.

L'obiettivo è quello di creare una perturbazione, cioè una differenza sensibile nella distanza tra un generico campione e i due nuovi centri di cluster, senza tuttavia modificare la disposizione degli altri cluster in modo apprezzabile. Se avviene lo "splitting", si va al passo 2 altrimenti si continua;

(è evidente che la $\bar{z}_j$ originaria scompare, N_C aumenta di 1)

Nota: per semplicità, i passi 8) e 9) e la condizione su σ del passo 10 possono essere ignorati. Se sono soddisfatte le condizioni *a)* e *b)* indicate nel passo 10, la perturbazione può essere creata semplicemente sommando (per $\bar{z}_j^{+}$) e sottraendo (per $\bar{z}_j^{-}$) una quantità γ a ciascuna coordinata del vettore $\bar{z}_j$. γ può essere calcolata come una frazione minima di $\overline{D}_j$.

Passo 10bis) Vai a 14)

Passo 11) Si calcolano le distanze a coppie D_{ij} tra i centri dei cluster:

$D_{ij} = \left\| \bar{z}_i - \bar{z}_j \right\|$, $i = 1,2,...,N_C - 1$ e $j = 1,2,...,N_C$

Passo 12) Si confrontano le distanze D_{ij} con il parametro ϑ_C. Si ordinano le *L* più piccole distanze che sono minori di ϑ_C in ordine crescente: $[\, D_{i_1 j_1}, D_{i_2 j_2}, ..., D_{i_L j_L} \,]$ (si ricordi che *L* è il numero massimo di centri di cluster che si possono aggregare).

Passo 13) A partire dalla distanza più piccola $D_{i_e j_e}$ si aggregano i centri a coppie $\bar{z}_{ie}$ e $\bar{z}_{je}$ (a condizione che $\bar{z}_{ie}$ e $\bar{z}_{je}$ non siano già stati usati per una aggregazione).

In pratica da $\bar{z}_{ie}$ e $\bar{z}_{je}$ si crea un nuovo centro come somma pesata dei due (i pesi sono il numero di campioni in ogni cluster)

$$\bar{z}_e^{\,*} = \frac{1}{N_{i_e} + N_{j_e}} \left[N_{i_e} \cdot \left(\bar{z}_{i_e} \right) + N_{j_e} \cdot \left(\bar{z}_{j_e} \right) \right]$$

(Praticamente $\bar{z}^*$ è il baricentro)

Si cancellano quindi $\bar{z}_{ie}$ e $\bar{z}_{je}$, riducendo N_C di 1. Si osservi che, poiché ogni centro di cluster è usato al più una volta, non è detto che si riescano a fare *L* aggregazioni.

Passo 14) Se questa è l'ultima iterazione, fine. Altrimenti andare al **passo 2**) (è previsto che si possa andare anche al **passo 1**) per modificare qualche parametro, se l'utente lo ritiene opportuno per una migliore adattatività: qui non è richiesto).

Tema.

In un file di tipo testo sono memorizzati un insieme di vettori da utilizzare per la clusterizzazione. Nella prima riga un numero intero indica la dimensionalità dei vettori. Nelle righe successive ci sono gli elementi dei vettori espressi come numeri reali, separati da almeno uno spazio (blank) o da un RETURN (<CR>). Il numero di vettori può superare le migliaia.

Realizzare un programma in linguaggio di alto livello (C, C++, Java) che attui l'algoritmo ISODATA sopra descritto.
Il nome del file dei dati deve essere introdotto da tastiera o passato come parametro.
Descrivere nel dettaglio ma in modo sintetico:

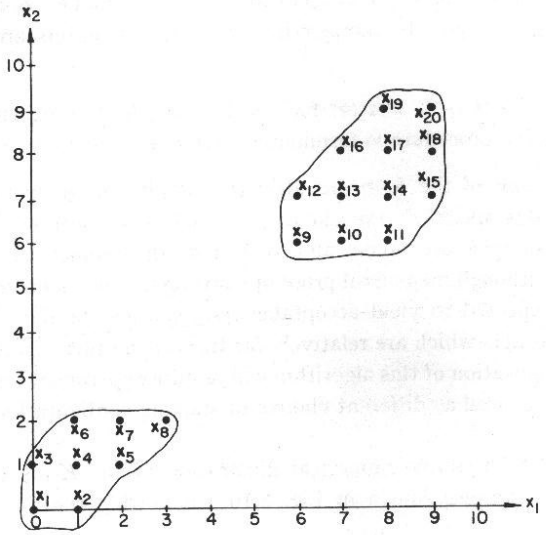
- le strutture dati (si tenga conto che il numero dei vettori può essere molto grande, il numero di cluster è limitato).

- gli algoritmi utilizzati nella manipolazione dei dati per realizzare il progetto.

Realizzare nel linguaggio prescelto il programma main e le funzioni principali (si possono trascurare le funzioni minori o quelle riconducibili ad algoritmi standard, come l'ordinamento ecc.).
Quanto non stabilito espressamente deve essere scelto dall'esaminando e sarà oggetto di valutazione.

Esempio numerico

Siano dati i campioni in figura:



1. Si ipotizza inizialmente k=2. Si sceglie $\bar{z}_1(1) = \bar{x}_1 = (0,0)'$, $\bar{z}_2(1) = \bar{x}_2 = (1,0)'$

2. Poiché $\|\bar{x}_1 - \bar{z}_1(1)\| < \|\bar{x}_1 - \bar{z}_i(1)\|$, con $i = 2$, $\bar{x}_1$ appartiene a $S_1(1)$. Idem per $\bar{x}_3$. Gli altri campioni sono più prossimi a $\bar{z}_2(1)$.

   Quindi:

   $S_1(1) = \{\bar{x}_1, \bar{x}_3\}$, $S_2(1) = \{\bar{x}_2, \bar{x}_4, ..., \bar{x}_{20}\}$

3. Si aggiorna il valore dei centri dei cluster

   $$\bar{z}_1(2) = \frac{1}{N_1} \sum_{\bar{x} \in S_1(1)} \bar{x} = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) = \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix}$$

   $$\bar{z}_2(2) = \frac{1}{N_2} \sum_{\bar{x} \in S_2(1)} \bar{x} = \frac{1}{18}(\bar{x}_2 + \bar{x}_4 + ... + \bar{x}_{20}) =$$

   $$\begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}$$

4. Non si sono esaurite le iterazioni, si va al passo 2

5. Ricalcolando le distanze rispetto ai nuovi centri si conclude che:

$$S_1(2)=\{\bar{x}_1,\bar{x}_2,...,\bar{x}_8\},$$
$$S_2(2)=\{\bar{x}_9,\bar{x}_{10},...,\bar{x}_{20}\}$$

6. Si calcolano di nuovo i centri dei cluster

$$\bar{z}_1(3)=\frac{1}{8}\{\bar{x}_1+\bar{x}_2+...+\bar{x}_8\}=\begin{pmatrix}1.25\\1.13\end{pmatrix},$$
$$\bar{z}_2(3)=\frac{1}{12}\{\bar{x}_9+\bar{x}_{10}+...+\bar{x}_{20}\}=\begin{pmatrix}7.67\\7.33\end{pmatrix}$$

7. Si supponga che a questo punto il secondo cluster risulti troppo grande e occorra effettuare uno split. Si procede sommando e sottraendo a $\bar{z}_2(3)$ un valore γ (pari ad es. a 0.001):

8. Si formano due nuovi centri di cluster:

$$\bar{z}_2^{+}(4)=\begin{pmatrix}7.671\\7.331\end{pmatrix}$$
$$\bar{z}_2^{-}(4)=\begin{pmatrix}7.669\\7.329\end{pmatrix}$$

9. Si rifanno le assegnazioni (su tutti i centri di cluster attualmente presenti), ecc.