1 Cluster Analysis

Scopo: dividere in classi (cluster) un insieme di elementi, secondo certi criteri di somiglianza.

Punti fondamentali della cluster analysis:

- concetto di distanza tra due elementi
- distanza tra due cluster (serve ad es. per la fase di linkage)
- raggruppamento o divisione? Divisione in k gruppi diretta invece che iterativa?
- albero completo o numero di cluster prefissato?
- indici di buona suddivisione? Indici di omogeneità di un cluster?

1.1 Distanza tra due elementi

Se i due elementi sono descritti da stringhe numeriche

$$x = (x_1, ..., x_n)\,, \quad y = (y_1, ..., y_n)$$

si può misurare la loro distanza in vari modi:

- euclidea:
$$d_2(x, y) = \sqrt{\sum (x_i - y_i)^2}$$
- Manhattan:
$$d_1(x, y) = \sum |x_i - y_i|$$
- caso peggiore:
$$d_\infty(x, y) = \max |x_i - y_i|$$
- e tanti altri. Non c'è limite alla fantasia.

1.2 Distanza tra due cluster

Dati due cluster C_1 e C_2:

- si può misurare la *minima* distanza tra i loro elementi
- si può misurare la *massima* distanza tra i loro elementi
- si può calcolare la *media* delle distanze tra i loro elementi (o lo scarto quadratico medio)
- si possono calcolare i *baricentri* e poi le loro distanze
- varianti pesate

Serve ad es. per la fase di linkage (raggruppamento).

1.3 Raggruppamento o divisione? Albero completo o numero di cluster prefissato?

Scelti questi due criteri di distanza, di procede iterativamente in due modi possibili:

- o partendo da tanti cluster quanti sono gli elementi in gioco, e raggruppando i cluster secondo una certa logica quantitativamente basata sulle distanze scelte
- oppure partendo da un unico cluster e suddividendolo.

Esempio di metodo di raggruppamento: si cercano i due elementi più vicini tra loro, si fondono in un unico cluster e si prende il baricentro come sostituto dei due elementi; l'insieme di n elementi è diventato un insieme di $n-1$ elementi. Si ripete. Da un certo punto in poi il numero di cluster può diminuire per più di una unità. Ci si ferma se si è raggiunto il numero di cluster prefissato, altrimenti si prosegue ad albero fino ad avere un solo cluster (es. *AGNES*).

Per esemplificare il metodo di suddivisione, servono due indici I_1 e I_2 che descriveremo. Si parte da un unico cluster, lo si spezza in due cluster in tutti i modi possibili calcolando, per ciascun modo, l'indice I_1 (che sarà un indice di buona suddivisione basato sulle distanze prescelte); si sceglie la suddivisione in due cluster che massimizza I_1. Se si cerca un albero completo, si

ripete la suddivisione per entrambi i cluster ottenuti, e così via fino ad avere cluster formati da singoli elementi (*DIANA*). Se invece abbiamo prefissato il numero di cluster, dividiamo solo il cluster che ha il più alto valore dell'indice I_2 (indice che misura il grado di disomogeneità del cluster); ottenuti iterativamente k cluster, si divide sempre il peggiore secondo I_2, fino a quando non si raggiunge il numero di cluster desiderato.

Esistono poi metodi che svolgono una suddivisione diretta, non iterativa, come *PAM* di *R*. Fissato il numero k di cluster che vogliamo ottenre, per ogni possibile scelta di k punti dello spazio dei dati, punti detti *medoids*, il meotdo aggrega i dati in k clusters attorno ai medoids, col criterio di minima distanza. Il metodo poi giudica la bontà della clusterizzazione ottenuta tramite un indicatore I_k (simile ad I_1). Il metodo cerca quindi i medoids che massimizzano I_k.

1.4 Indici di buona suddivisione? Indici di omogeneità di un cluster?

Qui si entra più nel tecnico e rinviamo alla letteratura specializzata. Un'idea che lasciamo a livello un po' intuitivo è basata sulla filosofia ANOVA: vogliamo che la varianza all'interno di ciascun cluster sia nettamente minore della varianza tra i cluster, oppure della varianza originaria. Ad esempio, per fare un caso semplice, se abbiamo n elementi

$$X_1, ..., X_n$$

e li suddividiamo in due cluster

$$C_1 = \left\{X'_1, ..., X'_{k_1}\right\}$$
$$C_1 = \left\{X''_1, ..., X''_{k_2}\right\}$$

(con $k_1 + k_2 = n$), formiamo la *varianza intercluster globale*:

$$\sigma^2_{inter} := \frac{1}{k_1} \sum d\left(X'_i, \overline{X}'\right)^2 + \frac{1}{k_2} \sum \left(X''_i - \overline{X}''\right)^2$$

e la solita *varianza globale*

$$\sigma^2 := \frac{1}{n} \sum \left(X_i - \overline{X}\right)^2.$$

Un indice di buona suddivisione è il rapporto

$$I_1 := \frac{\sigma^2}{\sigma^2_{inter}}.$$

Dato un cluster C possiamo associargli l'indice I_2 dato dal miglior valore di I_1 in C. Un cluster con I_2 elevato è un cluster in cui si spiega molta varianza suddvididendolo. E' quindi un cluster poco omogeneo. Se si deve scegliere un cluster da suddividere, conviene quello col valore più grande possibile di I_2:

$$I_2 = \max I_1$$

dove il massimo è calcolato al variare di tutte le suddivisioni possibili di quel cluster.

2 Classification and regression trees

Ricordiamo che il problema generale della regressione può essere formulato così: abbiamo una tabella di dati in cui le colonne $V_1, V_2, ...$ corrispondono a diverse variabili, le righe a diversi casi; si considerano alcune colonne come predittori, fattori, variabili indipendenti (indichiamole con $X_1, X_2, ...$) mentre altre come variabili da predire, output, variabili dipendenti (indichiamole $Y_1, Y_2, ...$). Si vuole trovare una relazione tra esse, genericamente parlando della forma

$$Y_j = f_j\left(X_1, X_2, ...\right) + \text{errore}.$$

La tecnica chiamata *regression trees* fa una cosa simile, operando nel seguente modo. Si sceglie una delle X, diciamo X_1. Scelta poi una soglia λ, si prendono tutti i casi che hanno $X_1 < \lambda$ e quelli con $X_1 \geq \lambda$. Questa è una suddivisione dell'insieme globale dei casi. Si valuta, relativamente alla sola variabile Y, l'indicatore I_1 e si prende il valore di λ che massimizza I_1. In realtà si ottimizza anche rispetto ai predittori X, salvo che si imponga una sequenza a priori dei predittori da usare.

Dopo queste operazioni il cluster globale è stato suddiviso in due. Da qui in poi si ripete, o fino all'albero completo oppure fino a k cluster, come nei metodi di suddivisione illustrati sopra.

In definitiva, rispetto agli altri metodi di suddivisione, la differenza è questa: c'è sempre un indicatore I_1 che serve a decidere quale sia la suddivisione migliore, ma non tutte le suddivisioni sono ammesse. Sono ammesse

solo le suddivisioni ottenute tramite disuguaglianze del tipo

$$X_i < \lambda.$$

Se quindi la differenza è che ci sono limitazioni alle possibilità di suddivisione, che vantaggi ci sono? Il vantaggio è che l'albero così ottenuto può essere usato per fare previsioni o classificazioni di ulteriori casi, in linea con la regressione che si può usare per previsioni della Y relativamente a nuovi valori di X. Si opera così. L'albero ottenuto col metodo regression tree ha nei suoi punti di biforcazione l'indicazione della disuguaglianza $X_i < \lambda$ utilizzata in quella biforcazione. Ad esempio, dal vertice principale (cluster globale iniziale) parte una biforcazione con scritto $X_2 < 35.2$ a destra, $X_2 \geq 35.2$ a sinistra. E così via ad ogni biforcazione successiva. Preso un nuovo caso di cui si conoscano i valori dei predittori X_i, si parte dal vertice principale ed a ogni biforcazione si segue la direzione che corrisponde alla disuguaglianza vera per quel caso. Alla fine il caso viene classificato in un cluster.