10

Le distanze e gli indici di similarità

In questo capitolo analizziamo le "prossimità" tra unità statistiche, alle quali corrispondono i vettori riga x_i', $(i = 1, \ldots, n)$ nella matrice dei dati. Un indice di prossimità tra due generiche righe unità i e j è definito come funzione dei rispettivi vettori riga della matrice dei dati:

$$f(x_i', x_j')$$

Vedremo di seguito una serie di esempi dove si specifica la natura di f.

Osservazione: con il termine prossimità ci si riferisce sia al concetto di rassomiglianza tra e unità sia a quello antitetico di diversità dato che è equivalente affermare che due unità sono molto simili oppure poco diverse.

Le informazioni fornite tra gli indici di prossimità tra coppie di elementi costituiscono la premesse per l'individuazione di gruppi di unità omogenee. La formazione di gruppi omogenei di unità (che sarà oggetto del capitolo sulla *cluster analysis*) può interpretarsi come una riduzione delle dimensioni dallo spazio $\Re^n$, poiché si riuniscono le unità in k sottoinsiemi (tipicamente con $k \ll n$).

10.1 Definizione di distanze

Si dice distanza tra due punti corrispondenti ai vettori x e $y \in \Re^p$, una funzione che gode delle seguenti proprietà:

1. non negatività
$$d(x, y) \geq 0 \qquad \forall x, y \in \Re^p$$

2. identità
$$d(x, y) = 0 \qquad \text{se e solo se } x = y$$

3. simmetria
$$d(x, y) = d(y, x) \qquad \forall x, y \in \Re^p$$