Preparazione dei dati

Introduzione

- Le analisi di business intelligence dipendono fortemente dalla qualità dei dati di ingresso.
- I dati estratti da varie fonti e raccolti nei data mart possono presentare anomalie che vanno identificate e corrette.
- Vedremo quali sono le principali tecniche per identificare e rimuove le anomalie e delle semplici soluzioni per migliorare l'accuratezza e l'efficienza degli algoritmi di apprendimento.

Validazione

In un dataset i dati possono essere affetti da:

- Incompletezza
  - Attributi o record mancanti
  - Record senza valori di certi attributi
  - Valori disponibili solo in forma aggregata
- Rumore
  - Errori di misurazione
  - Record e valori anomali
  - Dati duplicati
- Inconsistenza
  - Contraddizioni fra valori o fra record

Dati incompleti

- In un record manca il valore di un attributo.
- Misure possibili:
  - Ignorare la tupla.
  - Inserire il valore manualmente.
  - Inserire una costante globale (p.e. 0, ∞, ⊥).
  - Inserire il valore medio dell'attributo.
  - Inserire il valore medio dell'attributo per la classe di tuple cui appartiene questa.
  - Inserire il valore più verosimile, calcolato con un opportuno metodo (p.,e. regressione).

Dati soggetti a rumore

- Si indica con rumore una variazione casuale presente nei valori di un attributo numerico tale a originare anomalie.
- Occorre in primo luogo identificare gli outlier e procedere quindi alla loro rimozione o regolarizzazione.
- Se si suppone una distribuzione normale di ciascun attributo a_j del campione, la tecnica più semplice è considerare outlier i valori all'esterno dell'intervallo:

$$\left(\mu_j - 2z_{\alpha/2}\sigma_j, \mu_j + 2z_{\alpha/2}\sigma_j\right)$$

dove si è indicato con $z_{\alpha/2}$ il quantile di ordine $\alpha/2$ della distribuzione normale standard.

Eliminazione degli outlier

- Una tecnica alternativa si basa sulla distanza tra le osservazioni e sull'utilizzo di tecniche che permettono di raggruppare osservazioni simili (*clustering*).
- Si raggruppano le osservazioni in gruppi in maniera tale la distanza reciproca tra gli elementi di un gruppo sia minore di quella con elementi di altri gruppi
- Si considera outlier un elemento non compreso in nessun raggruppamento.
- Un semplice algoritmo di clustering è quella delle k-medie (*kmeans*):

Algoritmo delle k-medie

- Tecnica semplice ma efficace:
  1. Scegliere un valore di k, il numero di cluster da generare.
  2. Scegliere in modo casuale k osservazioni nel dataset.
     - Questi saranno i centri dei cluster.
  3. Collocare ogni altra osservazione nel cluster con il centro più vicino a essa.
  4. Utilizzare le osservazioni in ogni cluster per calcolarne il nuovo centro.
  5. Se i cluster non si sono modificati allora terminare, altrimenti ripetere il ciclo.
- Occorre definire un concetto di distanza (vicinanza).
- La distanza più comune fra osservazioni numeriche è quella euclidea della somma dei quadrati degli scarti.

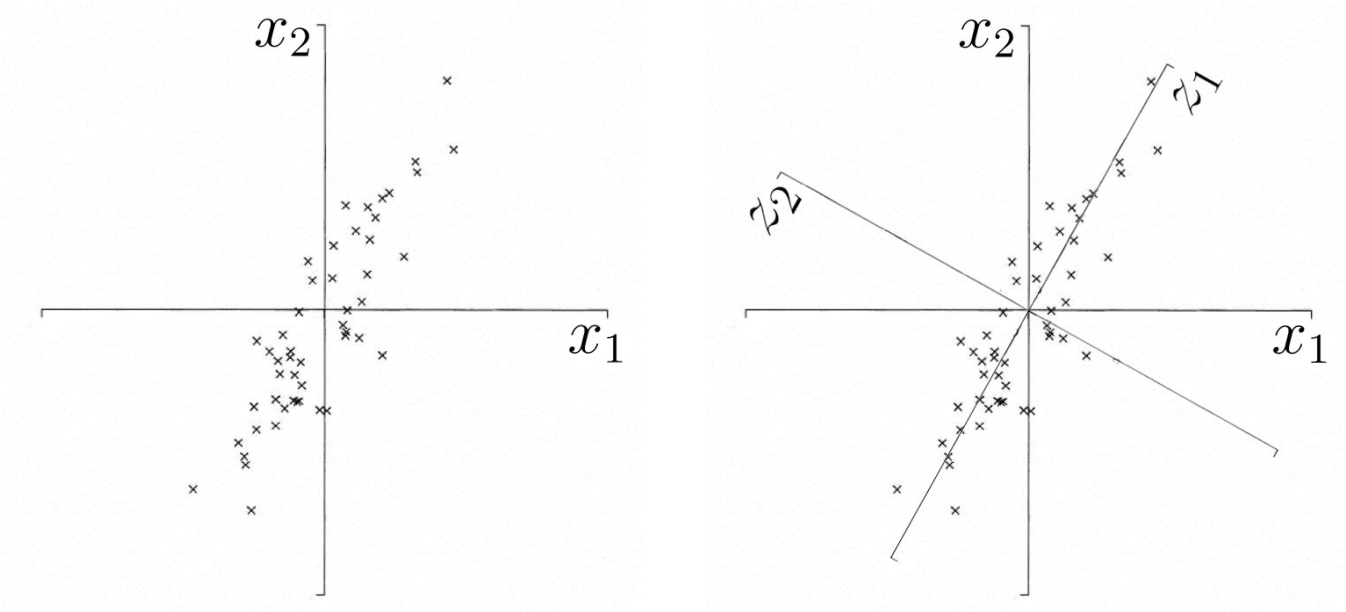
Riduzione degli attributi

- La riduzione degli attributi mira ad eliminare un sottoinsieme di variabili non rilevanti per l'analisi.
- Uno degli aspetti più critici in un processo di apprendimento è la determinazione di una combinazione di variabili predittive in grado di descrivere il fenomeno.
- Alcuni metodi selezionano sottoinsiemi di variabili, mentre altri scelgono loro opportune combinazioni.
- In quest'ultimo caso, si ottengono sottospazi in cui eseguire l'analisi.

Analisi alle componenti principali

- Si propone di ricavare una trasformazione che sostituisce un sottoinsieme degli attributi originali con un numero inferiore di nuovi attributi, ottenuti come loro combinazione lineare.
- Lo scopo è di avere la massima varianza in un campione in un numero minimo di variabili.

Rappresentazione geometrica



- La prima componente z_1 è la retta di minima distanza dai dati nello spazio di partenza.
- La seconda componente z_2 è la retta di minima distanza dai dati nello spazio ortogonale a z_1

Definizione algebrica

- Dato un insieme di n osservazioni su un insieme p di variabili $x = (x_1, x_2, \cdots, x_n)$, la prima componente principale del campione è definita dalla trasformazione lineare

$$z_1 = a_1^T x = \sum_{i=1,p} a_{i1} x_i$$

dove il vettore $a_1 = (a_{11}, a_{21}, \cdots, a_{p1})$ è scelto in modo da massimizzare la varianza di z_1

Definizione algebrica

- In maniera analoga, la k-esima componente è definita da:

$$z_k = a_k^T x \quad k=1,\ldots,p$$

dove il vettore $a_k=(a_{1k}, a_{2k}, \cdots, a_{pk})$ è scelto in modo da massimizzare la varianza di z_k soggetto al vincolo:

$$Cov[z_k,z_l]=0 \; per \; k > l > 0 \quad e \quad a_k^T a_k = 1$$

- Detta $V=X^TX$ la matrice di covarianza dei campioni, a_i è l'i-esimo autovettore di V corrispondente all'autovalore λ_i

Esercizi

- Determinare mediante *kmeans* gli eventuali outlier di ese1.mat (load –ascii ese1.mat)
- Visualizzare nello spazio delle prime due componenti principali il dataset ese2.mat (gscatter)
- L'indice

$$I_q = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}$$

misura la percentuale di varianza dispiegata dalle prime q componenti principali e fornisce un'indicazione della quantità di informazioni preservate da q attributi. Determinare quanti attributi di ese2 conservano il 98.5% dell'informazione complessiva.

Esercizi

- Eseguire il kmeans sul dataset contenuto in ese2.mat e sulle sue prime due componenti principali e commentare le differenze.
- Costruire un dataset composto da 4 classi in cui i primi due attributi ritengono il 98.5% dell'informazione.