



Analisi dei Dati

Lezione 9 - Preprocessing dei dati

Motivazioni

- I dati nel mondo reale sono sporchi
 - incompleti: mancano valori per gli attributi, mancano attributi importanti, solo valori aggregati
 - rumorosi: contengono errori e/o outliers
 - inconsistenti: contengono codici o nomi diversi per gli stessi dati
- Senza dati di qualità non c'è analisi di qualità
 - decisioni di qualità` debbono essere basate su dati di qualità
 - il data warehouse richiede una integrazione consistente di dati di qualità

Attività principali di pre-processing

- pulizia dei dati (data cleaning)
 - aggiunta di valori mancanti, aggiustamento dei dati rumorosi, identificazione e eliminazione degli outliers, soluzione delle inconsistenze
- Integrazione dei dati
 - integrazione di database, cubi e files
- Trasformazione dei dati
 - normalizzazione e aggregazione
- Riduzione dei dati
 - riduzione del volume dei dati mantenendo la qualità dell'analisi
- Discretizzazione dei dati

Pulizia dei dati (data cleaning)

- aggiunta dei dati mancanti
- identificazione degli outliers e riduzione dei dati rumorosi
- correzione dei dati inconsistenti

Dati mancanti

- I dati non sono sempre disponibili
 - molte tuple non hanno valori registrati per alcuni attributi, p.e. il reddito dei clienti nei dati delle vendite
- La mancanza dei dati può essere dovuta a:
 - malfunzionamento dei sistemi di acquisizione
 - cancellazione dovuta a inconsistenza con dati già registrati
 - dati non inseriti per incomprensione
 - certi dati possono non essere considerati importanti al momento dell'inserimento
 - mancanza di registrazione dei cambiamenti nei dati
- Ci può essere necessità di inferire i dati mancanti

Trattamento dei dati mancanti

- ignorare la tupla;
- aggiungere il valore mancante manualmente
- usare globalmente una costante per i valori mancanti: p.e. “non disponibile”
- usare il valor medio dell’attributo
- usare il valore più probabile dopo aver applicato una tecnica di inferenza (Bayesiana o albero di decisione)

Dati con rumore

- Rumore: errore o varianza random sui valori di una variabile
- Valori scorretti di un attributo possono essere dovuti a:
 - strumenti difettosi di raccolta dati
 - problemi di immissione dei dati
 - problemi di trasmissione dei dati
 - limitazioni tecnologiche
 - inconsistenze nelle convenzioni di rappresentazione
- Ulteriori problemi che richiedono pulizia dei dati
 - record duplicati
 - dati incompleti
 - dati inconsistenti

Trattamento del rumore dei dati

- **Binning (partizionamento):**
 - si ordinano i dati e si partizionano in bins (gruppi) di uguale dimensione)
 - si riducono le differenze (smoothing) all'interno dei bins o per valori medi, o per valori mediani, o per i valori min e max, ecc.
- **Clustering**
 - con algoritmi di clustering si individuano e rimuovono gli outliers
- **Combinazione di ispezione automatica e manuale**
 - determinare automaticamente i valori sospetti e farli controllare da un esperto
- **Analisi di regressione**
 - determina gli outliers e consente di avvicinarli alla curva (fitting sulla curva)

Esempio di Binning

- supponiamo di avere la seguente lista di prezzi:
4,8,9,15,21,21,24,25,26,28,29,34
- Partizionamento in bins di uguale dimensione:
 - Bin 1: 4,8,9,15
 - Bin 2: 21,21,24,25
 - Bin 3: 26,28,29,34
- Smoothing usando la media:
 - Bin 1: 9,9,9,9
 - Bin 2: 23,23,23,23
 - Bin 3: 29,29,29,29
- Smoothing usando gli estremi dell'intervallo
 - Bin 1: 4,4,4,15
 - Bin 2: 21,21,25,25
 - Bin 3: 26,26,26,34

Funzione SE (o IF)

- Specifica un test logico da eseguire e ritorna uno dei valori in base al risultato del test
- Sintassi: SE(Test; ThenValue; ElseValue)
 - **Test** è un valore o un'espressione qualsiasi che può dare come risultato VERO o FALSO.
 - **ThenValue** (facoltativo) è il valore restituito se il test logico è VERO.
 - **ElseValue** (facoltativo) è il valore restituito se il test logico è FALSO.

Esercitazione (I)

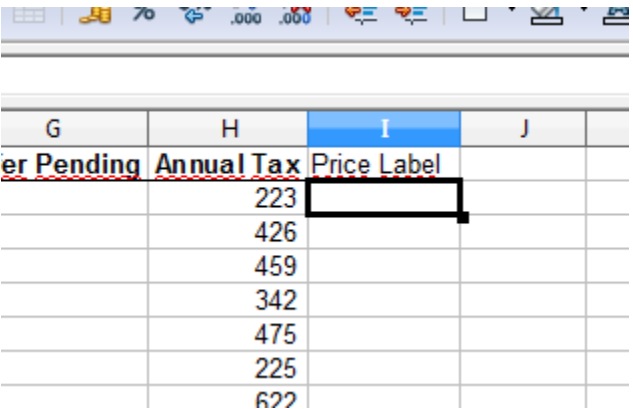
- Usando il file HOMEDATA
 - Modificare i valori dei prezzi delle case in base a tre valori di soglia scelti in modo tale che ogni intervallo abbia la stessa ampiezza
 - Le etichette da utilizzare per i tre intervalli sono:
 - price_low
 - price_medium
 - price_high

Esercitazione (2)

- Ordinare i valori di Price, determinare il minimo (54000) e il massimo (215000) valore e suddividere il range totale (161000) in tre parti (53666)
- I valori soglia per i tre intervalli sono dunque:
 - 107666
 - 161333
 - 215000

Esercitazione (3)

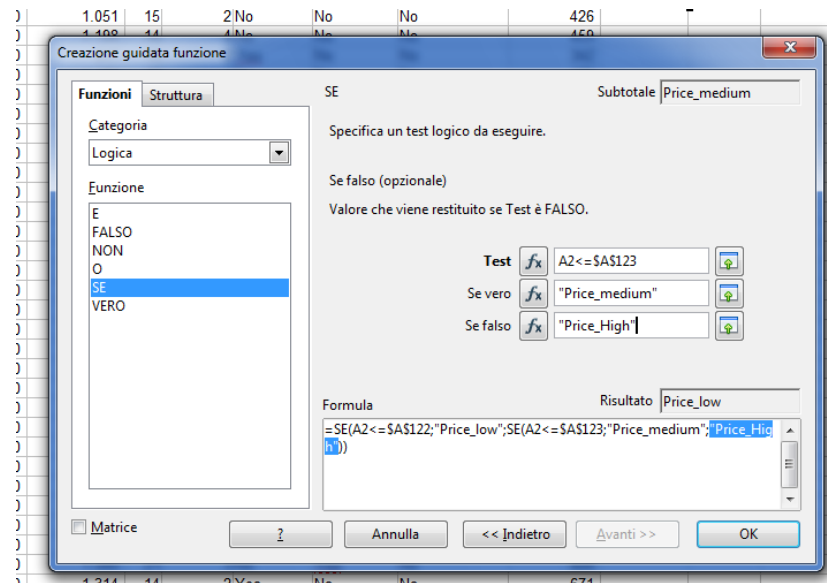
- Aggiungiamo una colonna a destra e mettiamo come titolo, ad esempio, “Price label”
- I valori di questa nuova colonna saranno le tre label stabilite per i tre intervalli
- Confrontiamo ogni valore di Price con le tre soglie e, utilizzando la funzione SE (IF) inseriamo il valore opportuno



G	H	I	J
er Pending	Annual Tax	Price Label	
	223		
	426		
	459		
	342		
	475		
	225		
	622		

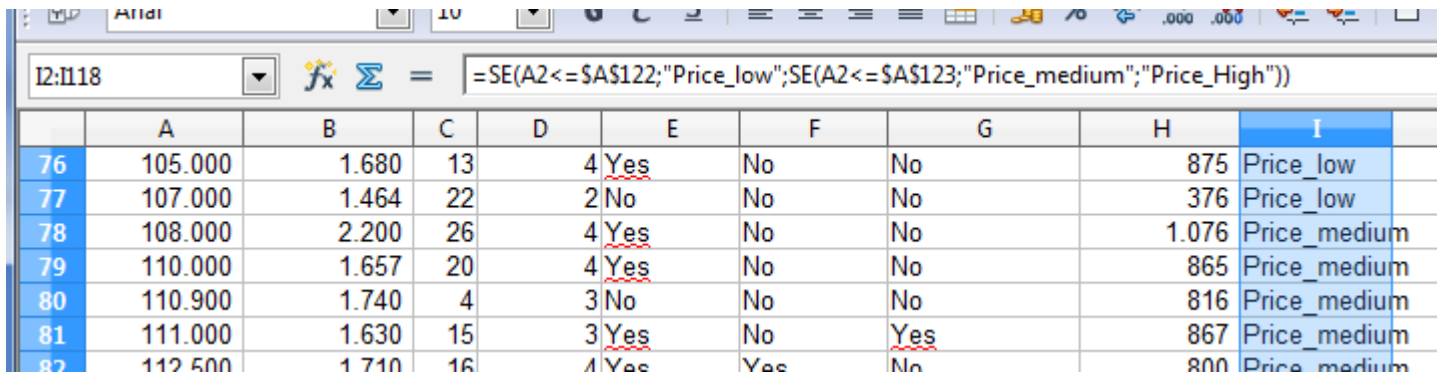
Esercitazione (4)

- I valori da confrontare sono tre, seguiremo il seguente algoritmo
 - SE valore < I soglia
 - ALLORA "Price_low"
 - ALTRIMENTI
 - SE valore < II soglia
 - ALLORA "Price_medium"
 - ALTRIMENTI "Price_high"
- Utilizzeremo due funzioni SE annidate
- Per comodità inseriamo le tre soglie sul foglio nelle celle A122:A124
- $\text{SE}(\text{A2} \leq \$\text{A}\$122; \text{"Price_low"}; \text{SE}(\text{A2} \leq \$\text{A}\$123; \text{"Price_medium"}; \text{"Price_High"}))$



Esercitazione (5)

- Estendiamo la formula a tutte le celle della colonna “Price Label”
- Nota1: il riferimento alle celle delle soglie è effettuato utilizzando il simbolo “\$” nelle coordinate (es. \$A\$122); in caso contrario, durante l’estensione della formula il riferimento sarebbe modificato in maniera progressiva: A123, A124, A125, ecc.
- Nota2: dato che avevamo ordinato i valori in senso crescente, le label risultanti sono anche esse ordinate. Sfruttiamo questa osservazione per commentare una strategia diversa di labeling



	A	B	C	D	E	F	G	H	I
76	105.000	1.680	13	4	<u>Yes</u>	No	No	875	Price_low
77	107.000	1.464	22	2	No	No	No	376	Price_low
78	108.000	2.200	26	4	<u>Yes</u>	No	No	1.076	Price_medium
79	110.000	1.657	20	4	<u>Yes</u>	No	No	865	Price_medium
80	110.900	1.740	4	3	No	No	No	816	Price_medium
81	111.000	1.630	15	3	<u>Yes</u>	No	<u>Yes</u>	867	Price_medium
82	112.500	1.710	16	4	<u>Yes</u>	<u>Yes</u>	No	800	Price_medium

Esercitazione (I)

- Usando il file HOMEDATA
 - Modificare i valori dei prezzi delle case in base a tre valori di soglia scelti in modo tale che ogni intervallo abbia la stessa ampiezza
 - Le etichette da utilizzare per i tre intervalli sono:
 - price_low
 - price_medium
 - price_high

Esercitazione (2)

- Come prima:
 - Ordinare i valori di Price, determinare il minimo (54000) e il massimo (215000) valore e suddividere il range totale (161000) in tre parti (53666)
 - I valori soglia per i tre intervalli sono dunque:
 - 107666
 - 161333
 - 215000

Esercitazione (3)

- Aggiungere la colonna “Price Label”
- Nella prima cella inserire la stringa “Price_low”

	A	B	C	D	E	F	G	H	I
1	<u>Price</u>	<u>Square Feet</u>	<u>Age</u>	<u>Features</u>	<u>NE Sector</u>	<u>Corner Lot</u>	<u>Offer Pending</u>	<u>Annual Tax</u>	Price Label
2	54.000	1.142	21	0 No	No	No	No	223	Price low
3	58.000	1.051	15	2 No	No	No	No	426	
4	60.000	1.198	14	4 No	No	No	No	459	
5	61.900	837	10	2 Yes	No	No	No	342	

- Scorriamo i valori della colonna “Price” fino a trovare un valore che supera la prima soglia (107666)
 - In corrispondenza di questo valore (il primo valore dell’intervallo successivo), inseriamo la stringa “Price_medium” nella colonna di destra

	A	B	C	D	E	F	G	H	I
76	105.000	1.680	13	4 Yes	No	No	No	875	
77	107.000	1.464	22	2 No	No	No	No	376	
78	108.000	2.200	26	4 Yes	No	No	No	1.076	Price medium
79	110.000	1.657	20	4 Yes	No	No	No	865	
80	110.900	1.740	4	3 No	No	No	No	816	

Esercitazione (4)

- Cerchiamo la riga corrispondente alla seconda soglia (161333)
- Inseriamo la label “Price_high” in corrispondenza

	A	B	C	D	E	F	G	H	I
106	158.000	2.563	14	2	No	Yes	No	1.189	
107	159.900	2.440	19	5	Yes	Yes	No	1.265	
108	169.500	2.931	28	3	Yes	No	Yes	1.142	Price high
109	180.000	2.774	2	4	Yes	No	No	1.765	
110	184.400	2.250	40	6	No	Yes	No	915	

- Partendo dalla prima cella in alto, estendiamo il contenuto della cella (l'angolo in basso a destra della cella) verso il basso, fino a trovare la prima label in corrispondenza del cambio di classe

	A	B	C	D	E	F	G	H	I
70	102.000	1.478	53	3	Yes	No	Yes	626	Price low
71	103.000	1.540	6	2	No	No	Yes	826	Price low
72	104.500	1.630	6	4	No	No	No	750	Price low
73	104.900	1.900	34	3	Yes	No	No	690	Price low
74	105.000	1.620	6	4	No	No	No	800	
75	105.000	1.920	8	4	No	No	No	944	
76	105.000	1.680	13	4	Yes	No	No	875	
77	107.000	1.464	22	2	No	No	No	376	
78	108.000	2.200	26	4	Yes	No	No	1.076	Price_medium

