

Database and data mining group, Politecnico di Torino 

Data mining Introduzione

Elena Baralis
Politecnico di Torino

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 1 *Elena Baralis
Politecnico di Torino*


Database and data mining group, Politecnico di Torino 

Knowledge discovery

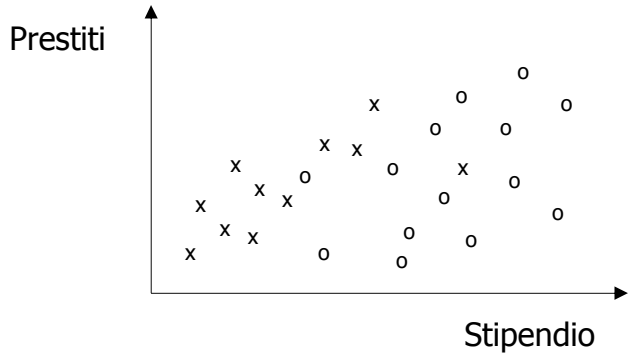
- Processo di estrazione dai dati di pattern
 - validi
 - precedentemente ignoti
 - potenzialmente utili
 - comprensibili

[Fayyad, Piatesky-Shapiro, Smyth 1996]

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 2 *Elena Baralis
Politecnico di Torino*

Database and data mining group, Politecnico di Torino



Esempio



Stipendio

Persone che hanno ricevuto un prestito dalla banca:
 x: persone che hanno mancato la restituzione di rate
 o: persone che hanno rispettato le scadenze

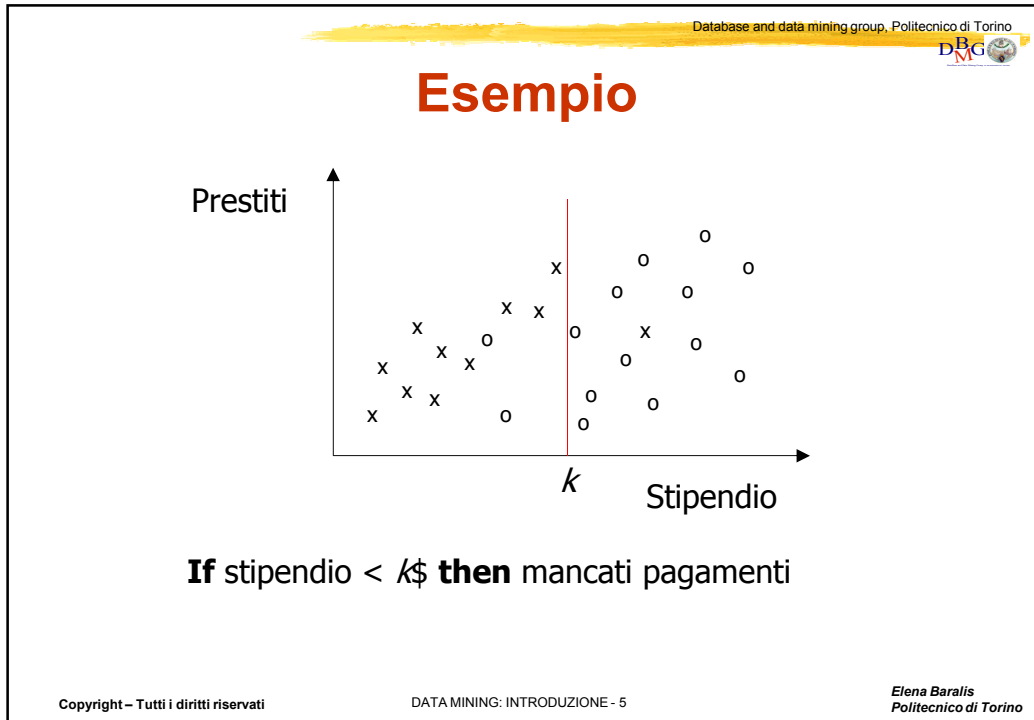
Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 3 Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino


Knowledge discovery

- Dati: insieme di informazioni contenute in una base di dati o data warehouse
- Pattern: espressione in un linguaggio opportuno che descrive in modo succinto le informazioni estratte dai dati
 - regolarità
 - informazione di alto livello

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 4 Elena Baralis Politecnico di Torino



- Database and data mining group, Politecnico di Torino

- ## Caratteristiche dei pattern
- Validità: i pattern scoperti devono essere validi su nuovi dati con un certo grado di certezza
 - Esempio: spostamento a destra del valore di k porta riduzione del grado di certezza
 - Novità: misurata rispetto a variazioni dei dati o della conoscenza estratta
- Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 6 Elena Baralis Politecnico di Torino


Caratteristiche dei pattern

- Utilità
 - Esempio: aumento di profitto atteso dalla banca associato alla regola estratta
- Comprensibilità: misure di tipo
 - sintattico (numero di bit del pattern)
 - semantico

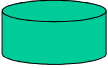
Processo di estrazione

- Richiede l'esecuzione di un insieme di passi
 - preparazione dei dati, ricerca dei pattern, valutazione dell'informazione estratta, raffinamenti
 - operazioni non "ovvie"

Database and data mining group, Politecnico di Torino



Processo di estrazione



Dati


Sviluppo della conoscenza del dominio applicativo

Copyright – Tutti i diritti riservati

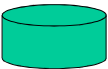
DATA MINING: INTRODUZIONE - 9

Elena Baralis
Politecnico di Torino


Database and data mining group, Politecnico di Torino



Processo di estrazione



Dati




Passo di **selezione dei dati**:
focalizzazione su un sottoinsieme
significativo dei dati

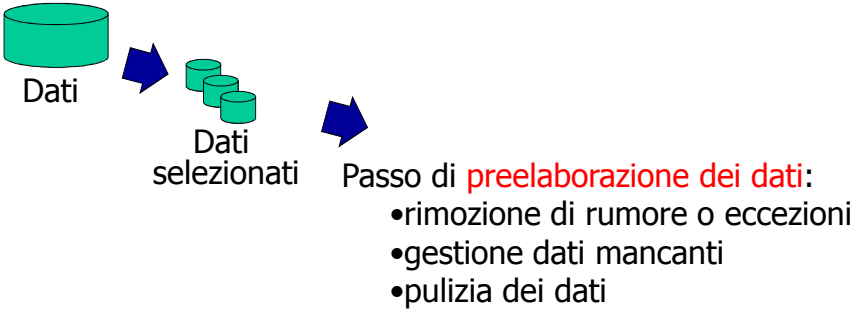
Copyright – Tutti i diritti riservati

DATA MINING: INTRODUZIONE - 10

Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino



Processo di estrazione



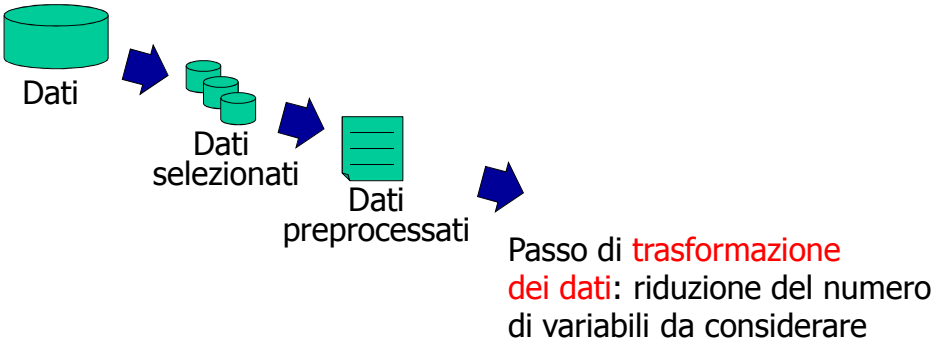
Passo di preelaborazione dei dati:

- rimozione di rumore o eccezioni
- gestione dati mancanti
- pulizia dei dati

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 11 *Elena Baralis Politecnico di Torino*

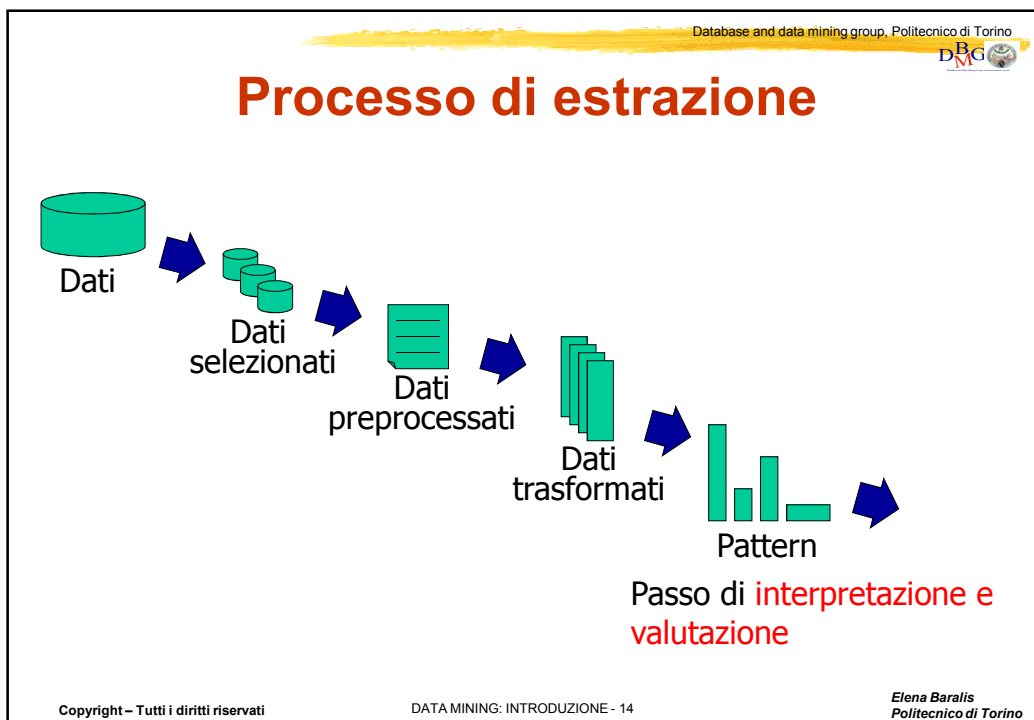
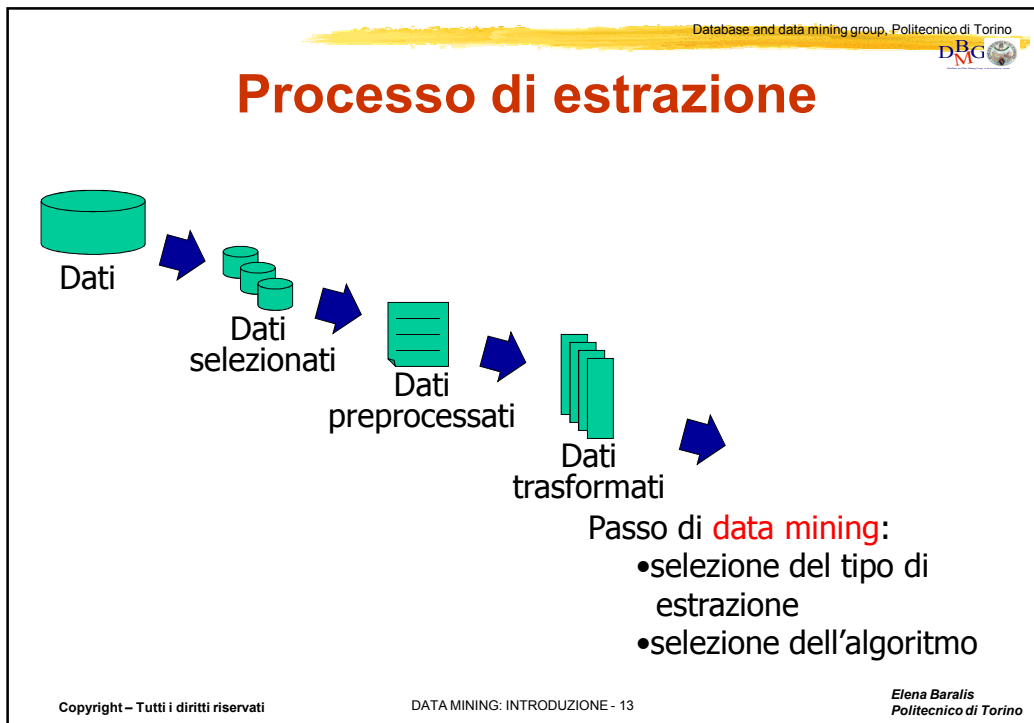
Database and data mining group, Politecnico di Torino


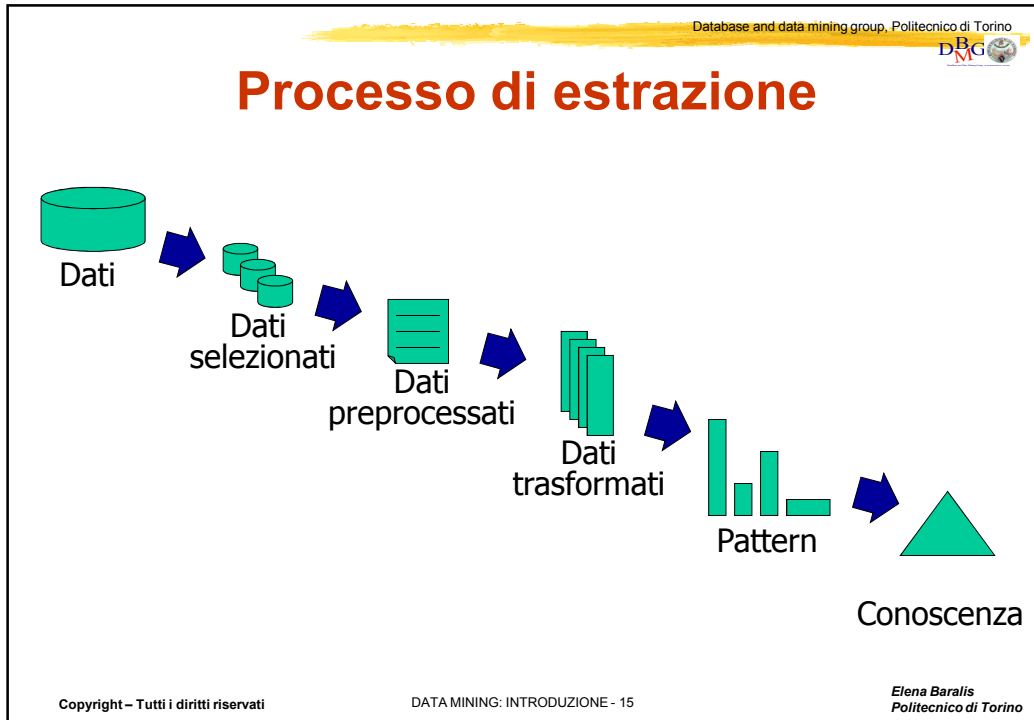
Processo di estrazione



Passo di trasformazione dei dati: riduzione del numero di variabili da considerare

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 12 *Elena Baralis Politecnico di Torino*





Database and data mining group, Politecnico di Torino
DBG

Data mining

Preparazione dei dati

Elena Baralis
Politecnico di Torino


Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 16 Elena Baralis Politecnico di Torino

Passi principali

- Pulizia dei dati
 - dati incompleti
 - dati rumorosi
 - riconoscimento di outliers ed eccezioni
 - gestione delle inconsistenze
- Integrazione dei dati
- Trasformazione dei dati
 - normalizzazione
 - aggregazione
- Riduzione dei dati
 - rappresentazione ridotta in volume, che genera risultati analitici simili
 - discretizzazione
 - campionamento

Dati incompleti


- Mancanza del dato dovuta a cause diverse, tipicamente legate al processo di data entry
 - malfunzionamento di strumenti
 - non inserito perché non importante
- Soluzioni
 - ignorare la tupla con informazione mancante
 - usare un valore speciale (NULL, N/A)
 - usare come valore il valor medio dell'attributo
 - usare come valore il valor medio dell'attributo all'interno della stessa classe

Database and data mining group, Politecnico di Torino


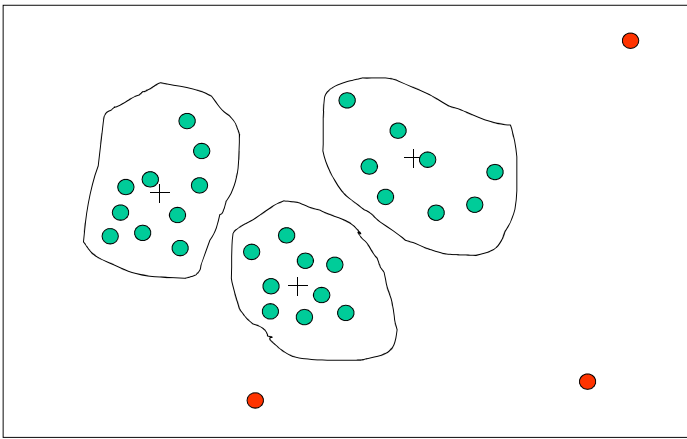
Dati rumorosi

- Errori casuali o varianza significativa di una misura
 - malfunzionamento di strumenti/trasmissione in rete
 - limitazioni tecnologiche
 - problemi di data entry
- Soluzioni
 - clustering o regressione per riconoscere ed eliminare gli outliers
 - discretizzazione

Copyright – Tutti i diritti riservati
DATA MINING: INTRODUZIONE - 19
Elena Baralis
Politecnico di Torino

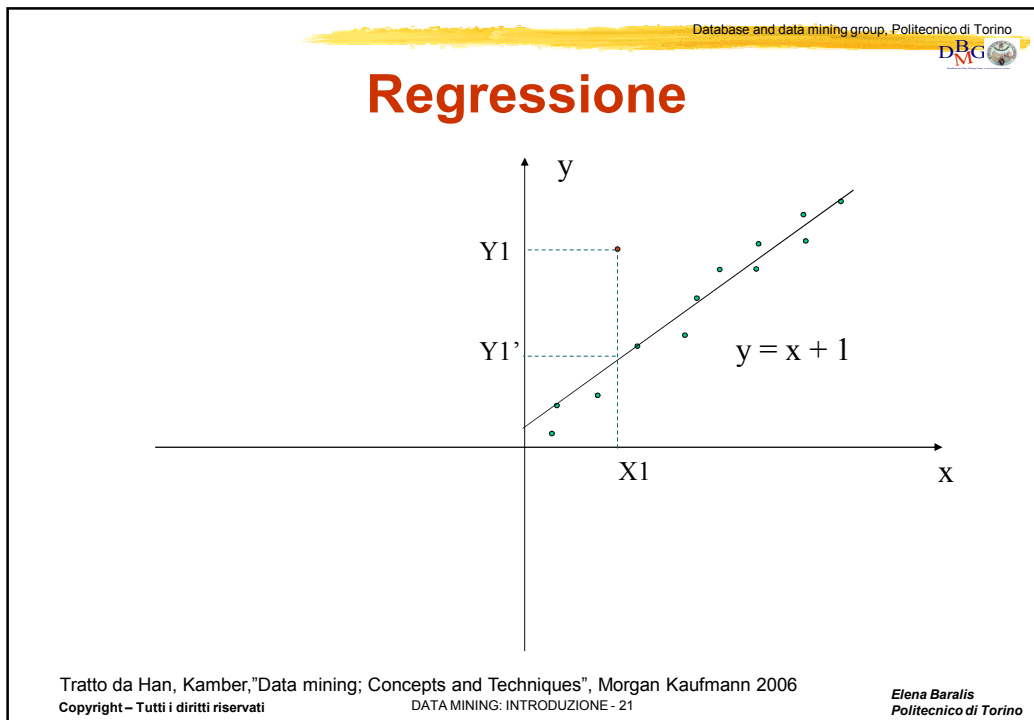
Database and data mining group, Politecnico di Torino


Clustering



The diagram shows a 2D space with three clusters of green points. Each cluster is enclosed by a hand-drawn black outline and contains a small black cross representing its centroid. There are three red points scattered outside these clusters, representing outliers.

Tratto da Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006
Copyright – Tutti i diritti riservati
DATA MINING: INTRODUZIONE - 20
Elena Baralis
Politecnico di Torino



Database and data mining group, Politecnico di Torino



Normalizzazione dei dati

- È un tipo di trasformazione dei dati
 - i valori di un attributo sono riportati in un intervallo prefissato
 - tipicamente (-1,+1) o (0,+1)
- Tecniche
 - min-max

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$
 - z-score

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$
 - scalamento decimale

$$v' = \frac{v}{10^j} \quad j \text{ intero minimo tale che } \max(|v'|) < 1$$

Copyright – Tutti i diritti riservati

DATA MINING: INTRODUZIONE - 22


Elena Baralis
Politecnico di Torino

Riduzione dei dati

- Generazione di una rappresentazione ridotta in volume, che genera risultati analitici simili
 - campionamento
 - riduzione della cardinalità dell'insieme
 - feature selection
 - riduzione del numero di attributi
 - discretizzazione
 - riduzione della cardinalità del dominio di un attributo

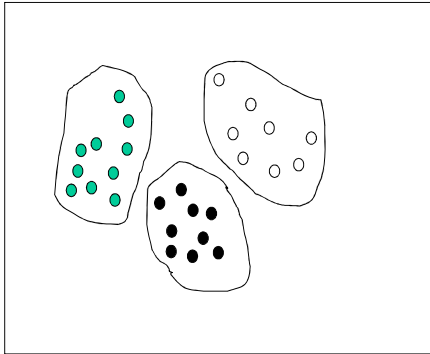
Campionamento

- Selezione di un sottoinsieme (campione) dei dati di partenza
 - rappresentativo
 - riduce la complessità degli algoritmi di data mining
 - non sempre riduce effettivamente il numero di I/O di pagine
- Modalità
 - campionamento casuale
 - senza sostituzione
 - con sostituzione
 - campionamento stratificato
 - per ogni sottoclasse di interesse, il campione contiene una frazione proporzionale alla popolazione della sottoclasse nella base dati di partenza
 - appropriato per dati con distribuzione non uniforme

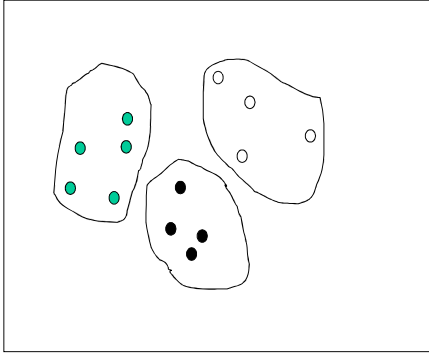
Database and data mining group, Politecnico di Torino


Campionamento


Dati di partenza



Campionamento stratificato




Tratto da Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006
 Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 25 Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino


Feature selection

- Selezione di un sottoinsieme degli attributi che fanno parte dello schema di partenza
 - selezione di un insieme minimo di attributi tale da conservare la distribuzione dei dati di partenza
 - utile specialmente per gli algoritmi di clustering
- Causa anche una riduzione del numero di pattern estratti
 - maggiore interpretabilita` del risultato
- Tecniche euristiche
 - esplorazione esaustiva impossibile a causa del vasto numero di scelte possibili
 - si aggiunge/toglie un attributo alla volta e si osserva l'effetto


Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 26 Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino


Discretizzazione

- Divisione del dominio di un attributo continuo in un insieme di intervalli
 - riduce la cardinalità del dominio di un attributo
- Tecniche
 - N intervalli con la stessa ampiezza $W=(v_{\max} - v_{\min})/N$
 - semplice
 - sensibile agli outliers e ai dati sparsi
 - incrementale
 - N intervalli con (approssimativamente) la stessa cardinalità
 - si adatta meglio a dati sparsi e outliers
 - non incrementale
 - clustering
 - si adatta bene a dati sparsi e outlier

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 27 *Elena Baralis Politecnico di Torino*

Database and data mining group, Politecnico di Torino


Discretizzazione

Prezzo	Ampiezza (W = 10)	Cardinalità (N = 2)	Clustering
7	[0, 10]	[7, 20]	[7, 7]
20	[11, 20]	[22, 50]	[20, 22]
22	[21, 30]	[51, 53]	[50, 53]
50	[31, 40]		
51	[41, 50]		
53	[51, 60]		

Copyright – Tutti i diritti riservati DATA MINING: INTRODUZIONE - 28 *Elena Baralis Politecnico di Torino*