Regole associative con Weka
Soluzioni degli esercizi

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

Apriori parametri e output

- In questa fase utilizzeremo il data set CensusTrainining.arff che riporta dati del censimento USA (http://cps.ipums.org/)
  - ✓ 1000 istanze per il training

Attributo	Descrizione
age	Età in anni
workclass	Classe di lavoro
fnlwgt	"Final sampling weight" peso dell'istanza (campione) rispetto alla popolazione
education	Titolo ottenuto
education-num	Numero di anni di studio
marital-status	Stato civile
occupation	Occupazione
relationship	Tipo di relazione con il capo famiglia
race	Razza
sex	Sesso
capital-gain	Utili da capitali (plus valenza)
capital-loss	Perdite da capitali (minus valenza)
hours-per-week	Ore di lavoro settimanali
native-country	Nazionalità
Total Income	**L'individuo guadagna più o meno di 50K$**

Preprocessing

- Gli algoritmi di ricerca delle RA operano solo con attributi discreti
  - ✓ Discretizzare gli attributi numerici mediante il filtro Discretize
    - 10 bins
    - UseEqualFrequency=true
  - ✓ Eliminare l'attributo Fnlwgt
  - ✓ Salvare il file come CensusTrainingDiscrete.arff
- Eseguire un'analisi manuale dei dati al fine di identificare eventuali correlazioni tra coppie di attributi

Apriori: i parametri

- car: specifica se ricercare gereriche AR (false) oppure AR che abbiano l'attributo classIndex come conseguente
- lowerBoundMinSupport: valore minimo per il supporto di una regola
- metricType: metrica da utilizzare per la valutazione della regola
  - ✓ Confidence è l'unica utilizzabile se car=true
  - ✓ Lift
  - ✓ Leverage, Conviction
- minMetric: valore minimo per la metrica utilizzata
- numRules: numero massimo di regole da restituire
- outputItemSets: se true il sistema restituisce anche gli itemset frequenti
- delta – fattore di riduzione della soglia di supporto minimo da upperBoundMinSupport a lowerBoundMinSupport. Le iterazioni si fermano se si raggiunge il valore del lower bound o si scopre il numero richiesto di regole

Apriori: interpretare i risultati

- Eseguire Apriori
  - ✓ Porre outputItemSet=true

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Large Itemsets L(1):
race=White 847
capital-gain='(-inf-297]' 919
capital-loss='(-inf-326.5]' 950
native-country=United-States 902
class=<=50K 768

Size of set of large itemsets L(2): 6

Large Itemsets L(2):
race=White capital-gain='(-inf-297]' 774
race=White capital-loss='(-inf-326.5]' 805
race=White native-country=United-States 778
capital-gain='(-inf-297]' capital-loss='(-inf-326.5]' 869
capital-gain='(-inf-297]' native-country=United-States 826
capital-loss='(-inf-326.5]' native-country=United-States 861

Size of set of large itemsets L(3): 1

Large Itemsets L(3):
capital-gain='(-inf-297]' capital-loss='(-inf-326.5]' native-country=United-States 785

Best rules found:

 1. native-country=United-States 902 ==> capital-loss='(-inf-326.5]' 861    conf:(0.95)
 2. race=White 847 ==> capital-loss='(-inf-326.5]' 805    conf:(0.95)
 3. capital-gain='(-inf-297]' native-country=United-States 826 ==> capital-loss='(-inf-326.5]' 785    conf:(0.95)
 4. capital-gain='(-inf-297]' 919 ==> capital-loss='(-inf-326.5]' 869    conf:(0.95)
 5. race=White 847 ==> native-country=United-States 778    conf:(0.92)
 6. native-country=United-States 902 ==> capital-gain='(-inf-297]' 826    conf:(0.92)
 7. capital-loss='(-inf-326.5]' 950 ==> capital-gain='(-inf-297]' 869    conf:(0.91)
 8. race=White 847 ==> capital-gain='(-inf-297]' 774    conf:(0.91)
 9. capital-loss='(-inf-326.5]' native-country=United-States 861 ==> capital-gain='(-inf-297]' 785    conf:(0.91)
10. capital-loss='(-inf-326.5]' 950 ==> native-country=United-States 861    conf:(0.91)
```

Supporto conseguente

Supporto antecedente

Supporto regola

- ✓ Le regole sono ordinate in base al valore della metrica utilizzata
- ✓ Il supporto della regola ossia dell'itemset che include antecedente e conseguente è >= al supporto dell'antecedente per la proprietà di anti-monotona del supporto

Ricerca iterativa a supporto variabile

- In questa particolare implementazione Apriori esegue più cicli di generazione delle regole riducendo progressivamente il supporto richiesto da upperBoundMinSupport a lowerBoundMinSupport a passi di delta
  - ✓ A ogni iterazione sono restituite le regole che superano la soglia minMetric
- Il ciclo si interrompe quando:
  - ✓ è stato raggiunto il valore lowerBoundMinSupport
  - ✓ sono state individuate numRules regole
- Le regole sono comunque sempre riordinate in base al valore della metrica

Ricerca iterativa a supporto variabile

- Aumentando il numero di regole richieste:
  - ✓ Sono possibili più iterazioni
  - ✓ Saranno individuate regole con supporto più basso ma potenzialmente con confidenza più elevata
- Alzando il valore di lowerBoundMinSupport
  - ✓ si trovano regole con supporto elevato
  - ✓ tende a ridursi il numero delle regole trovate, la relativa confidenza e lunghezza
- Limitando l'intervallo [lowerBound, upperBound] del min support è possibile analizzare fenomeni con supporto definito.
  - ✓ Ciò consente di evitare pattern cross-dimensionali poiché di fatto si limita min e max del supporto dei singoli attributi

Ricerca iterativa a supporto variabile

- Impostare outputItemsets=TRUE, lanciare Apriori con i rimanenti parametri di default e discutere il risultato
- Impostare numRules=100 e discutere il risultato
- Sulla base del risultato precedente tarare il parametro lowerBoundMinSupport in modo da ottenere regole di lunghezza 3. Spiegare il ragionamento fatto e discutere i risultati

Ricerca iterativa a supporto variabile

- Impostare outputItemsets=TRUE, lanciare Apriori con i rimanenti parametri di default e discutere il risultato
- Impostare numRules=100 e discutere il risultato
- Sulla base del risultato precedente tarare il parametro lowerBoundMinSupport in modo da ottenere regole di lunghezza 3. Spiegare il ragionamento fatto e discutere i risultati

Regole associative per uno specifico attributo

- Impostare car=TRUE e classIndex=14 e verificare il comportamento di Apriori
- Le regole associative fanno scelte similari a quelle di un algoritmo di classificazione
- Vincolare le regole ad avere uno specifico conseguente serve a capire fenomeni correlati a quell'attributo

Market Basket Analysis

- Market Basket Analysis: scopo dell'analisi è l'individuazione dei comportamenti/abitudini di acquisto dei consumatori, per progettare opportune azioni di marketing, ad esempio:
  - ✓ promozione prodotti
  - ✓ Collocazione prodotti negli scaffali dei supermarket
  - ✓ Composizione e invio cataloghi pubblicitari
- Utilizziamo il data set «MarketBasket.arff», relativo a un ipotetico supermarket
  - ✓ Istanze 651
  - ✓ Attributi 56 (binari)
    - Uno per ogni prodotto in vendita

Formato del file

- Un attributo per ogni prodotto
  - ✓ Ogni riga rappresenta una transazione di acquisto
- Sono possibili due formati
  - ✓ Formato denso: lunghezza di ogni transazioni pari al numero di prodotti in vendita. Utilizzo di dati missing per prodotti non acquistati
    - Weka non riconosce attributi asimmetrici quindi se si utilizzasse {t,**f**} il sistema restituirebbe regole associative legate alle implicazioni logiche tra prodotti non acquistati

```
@relation transaction_example_dense
@attribute product1 {t}
@attribute product2 {t}
@attribute product3 {t}
@attribute product4 {t}
@data
?,t,?,?
t,?,t,?
t,t,t,t
```

Formato del file

- Un attributo per ogni prodotto
  - ✓ Ogni riga rappresenta una transazione di acquisto
- Sono possibili due formati
  - ✓ Formato sparso: lunghezza variabile delle transazioni che contengono solo i prodotti acquistati specificati nella forma (indice prodotto, valore).

```
@relation transaction_example_sparse
@attribute product1 {f,t}
@attribute product2 {f,t}
@attribute product3 {f,t}
@attribute product4 {f,t}
@data
{2 t}
{0 t, 2 t}
{0 t, 1 t, 2 t, 3 t}
```

Market Basket Analysis

- Ricercare regole associative interessanti
  - ✓ Con i parametri di default (*minSupport=0.1 numRules=10*; *MinMetric*=0.9 *metricType*= Confidence)
  - ✓ Portando *minSupport=0.01*
  - ✓ Portando *minSupport=0.05*
  - ✓ Portando *minMetric=0.7*
- Le regole non evidenziano pattern interessanti! Quali le possibili cause?

Market Basket Analysis

- Discutere il legame tra Hamburger Buns, 98pct Fat Free Hamburger e White Bread