Slide 2: Cosa Sono le Regole di Associazione?

Qui iniziamo a definire cosa sono le regole di associazione. Ci viene detto che sono "modelli di apprendimento non supervisionato". Questo è un concetto chiave: "non supervisionato" significa che a differenza di altri modelli di machine learning (come la classificazione), non abbiamo un "attributo target" o una variabile che vogliamo prevedere. Invece, l'obiettivo è scoprire strutture o pattern nascosti nei dati.

L'obiettivo principale è "identificare pattern e ricorrenze regolari all'interno di un ampio insieme di transazioni". Pensa a un supermercato: ogni acquisto è una transazione. Le regole di associazione cercano di capire quali articoli vengono acquistati insieme frequentemente. Sono descritte come "semplici e intuitive" e vengono spesso utilizzate per analizzare le transazioni di vendita (il classico esempio è l'analisi del "cestino della spesa" o market basket analysis) e i percorsi di navigazione sui siti web (web mining). Per il tuo corso, questo significa capire come i dati transazionali possono rivelare comportamenti dei clienti o degli utenti.

Slide 3: Applicazione 1 - Market Basket Analysis

Questa slide approfondisce l'applicazione più comune delle regole di associazione: la "market basket analysis". Si tratta di un'analisi di data mining che mira a identificare regole ricorrenti che collegano l'acquisto di un prodotto (o gruppo di prodotti) all'acquisto di un altro prodotto (o gruppo di prodotti).

L'esempio fornito è molto chiaro: "un cliente che acquista cereali per la colazione acquisterà anche latte con una probabilità di 0,68". Questo non significa che *sempre* comprerà il latte, ma che c'è un'alta probabilità basata sui dati storici. Per i marketing manager, queste informazioni sono estremamente utili. Permettono di:

- Pianificare iniziative promozionali (es. "compra cereali e avrai uno sconto sul latte").
- Definire l'assortimento dei prodotti (assicurarsi che ci sia sempre latte disponibile se si vendono cereali).
- Decidere la disposizione dei prodotti sugli scaffali (mettere latte e cereali vicini per facilitare l'acquisto congiunto).

In sintesi, la market basket analysis aiuta le aziende a capire meglio il comportamento d'acquisto dei clienti e a ottimizzare le strategie di vendita.

Slide 4: Applicazione 2 - Web Mining

Oltre all'analisi dei carrelli della spesa, le regole di associazione sono molto utili anche nel "web mining". Qui l'attenzione si sposta dai prodotti fisici ai percorsi di navigazione online. L'obiettivo è comprendere i "pattern dei percorsi di navigazione" e la frequenza con cui "combinazioni di pagine web vengono visitate da un dato individuo durante una singola sessione o sessioni consecutive".

L'esempio è un utente che visita siti di notizie: "se un individuo visita il sito timesonline.co.uk allora entro una settimana visiterà anche il sito economist.com con una probabilità di 0,87". Questo tipo di informazione è prezioso per chi gestisce siti web, poiché permette di:

- Migliorare la struttura dei link tra le pagine per facilitare la navigazione.
- Consigliare percorsi di navigazione specifici agli utenti.

- Posizionare banner pubblicitari e altri messaggi promozionali in modo più efficace.

Questo dimostra come le regole di associazione possano essere applicate anche a dati non transazionali nel senso stretto del termine, ma a sequenze di azioni o eventi.

Slide 5: Altre Applicazioni

Le regole di associazione sono versatili e hanno altre applicazioni interessanti:

- **Acquisti con carta di credito**: Possono essere usate per analizzare gli acquisti fatti dai titolari di carte di credito al fine di indirizzare promozioni future. Il testo sottolinea che in questo scenario, i prodotti e servizi potenzialmente accessibili al titolare della carta sono "virtualmente infiniti", rendendo l'analisi più complessa ma anche potenzialmente più ricca di scoperte.
- **Rilevamento frodi (Fraud detection)**: In questo caso, le transazioni possono essere "segnalazioni di incidenti e richieste di risarcimento per il danno subito". L'esistenza di "combinazioni speciali" (pattern insoliti) può rivelare "comportamenti potenzialmente fraudolenti e giustificare un'indagine approfondita da parte della compagnia assicurativa". Questo è un esempio cruciale per un ingegnere informatico, dato l'importanza della sicurezza e dell'integrità dei dati nel business digitale.

Slide 6: Struttura delle Regole e Problematiche

Ora entriamo più nel dettaglio della struttura di una regola di associazione. Una regola è espressa come "Y⇒Z", che si legge "se Y è vero, allora Z è anche vero". Spesso si parla di "regola probabilistica", dove "p" è la probabilità che Z sia vero dato che Y è vero.

È fondamentale che le regole estratte siano "non-banali e interpretabili in piani d'azione concreti". Ciò significa che una regola non dovrebbe essere ovvia (es. "chi compra scarpe compra due scarpe") o semplicemente riflettere campagne promozionali passate. Inoltre, è importante fare attenzione alla "confusione tra causa ed effetto". L'esempio "gli acquirenti di una polizza assicurativa compreranno anche un'auto con una probabilità di 0,98" è "inutile per i marketing manager" perché la causa è l'acquisto dell'auto, non della polizza. Questo sottolinea l'importanza di una comprensione del dominio (business knowledge) oltre alla mera analisi statistica.

Slide 7: Rappresentazione del Dataset (Lista di Transazioni)

Questa slide ci mostra come i dati di input per le regole di associazione sono tipicamente strutturati. Il dataset contiene una "lista di transazioni", e "ciascuna contiene una lista di item", con un "identificatore unico".

Viene presentata una tabella d'esempio con 10 transazioni (ti) e gli item che contengono:

- 001: {a, c}
- 002: {a, b, d}
- ...
- 010: {a, e}

Poi, per chiarezza, viene data una "legenda" per gli item: {a, b, c, d, e} = {pane, latte, cereali, caffè, tè}. Questo è l'esempio del "cestino della spesa" in pratica.

Slide 8: Rappresentazione del Dataset (Matrice Binaria) e Frequenza

Il dataset può essere rappresentato anche come una "matrice bidimensionale", dove ogni riga è una transazione e ogni colonna è un item. Il valore nella cella sarà "1 o 0 se l'item è o non è presente in una transazione". A volte, per valori continui, potrebbe indicare la frequenza con cui l'oggetto appare, ma per le regole di associazione classiche è spesso binario.

Viene mostrata la matrice corrispondente all'esempio precedente.

- "identificatore ti" | "a" | "b" | "c" | "d" | "e"
- 001 | 1 | 0 | 1 | 0 | 0
- ...

Qui vengono introdotti due concetti fondamentali:

- **Frequenza empirica di un itemset**: è il "numero di transazioni Ti esistenti nel dataset D che contengono l'itemset". Ad esempio, la frequenza di {a, c} è 4 (si trova nelle transazioni 001, 005, 007, 009).
- **Probabilità di occorrenza di un itemset**: è il "rapporto tra la frequenza empirica e il numero totale di transazioni". Quindi, per {a, c}, la probabilità di occorrenza è 4/10=0.4. Questi concetti sono cruciali per la fase successiva.

Slide 9: Regole di Associazione a Dimensione Singola - Confidenza e Supporto

Questa slide introduce i due metriche principali per valutare le regole di associazione:

- **Supporto (Support)**: È la "proporzione di transazioni che contengono sia L (antecedente) che H (conseguente)". In altre parole, misura "la frequenza con cui una coppia antecedente-conseguente appare insieme nelle transazioni di un dataset". Un "basso supporto" indica che la regola potrebbe essere occasionale e di scarso interesse, quindi il supporto è usato per "scartare regole di scarsa significatività".
- **Confidenza (Confidence)**: È la "proporzione di transazioni contenenti l'itemset H (conseguente) tra quelle che includono l'itemset L (antecedente)". Misura l'affidabilità della regola. Una "alta confidenza" significa un'alta probabilità che H esista in una transazione che contiene anche L.

In formula:

- Support(L⇒H)=P(L∪H) (probabilità che L e H appaiano insieme)
- Confidence(L⇒H)=P(H|L)=Support(L∪H)/Support(L) (probabilità che H appaia dato che L è presente)

Slide 10: Esempio di Confidenza e Supporto

Questa slide applica i concetti di confidenza e supporto all'esempio "se a⇒c".

- Per il **Supporto**: Dobbiamo trovare la proporzione di transazioni che contengono sia 'a' che 'c'. Dalla tabella (che viene ripetuta per comodità), le transazioni che contengono sia 'a' che 'c' sono 001, 005, 007, 009. Ci sono 4 transazioni su un totale di 10. Quindi, il supporto è 4/10=0.4.
- Per la **Confidenza**: Dobbiamo trovare la proporzione di transazioni che contengono 'c' *tra quelle che contengono 'a'*. Le transazioni che contengono 'a' sono: 001, 002, 005, 007, 008, 009, 010. Ci sono 7 transazioni che contengono 'a'. Di queste 7, quelle che contengono anche 'c' sono 001, 005, 007, 009 (4 transazioni). Quindi, la confidenza è 4/7≈0.57.

Questo esempio pratico aiuta a consolidare la comprensione di queste due metriche fondamentali.

Slide 11: Fasi della Scoperta delle Regole di Associazione

La scoperta delle regole di associazione si articola in due fasi principali:

1. **Generazione di frequent itemsets (basata sul supporto)**: Questa fase si concentra sull'identificazione di insiemi di item che appaiono frequentemente insieme nel dataset, superando una "soglia minima di supporto". Vengono utilizzati "algoritmi per ottenere i frequent itemsets in modo efficiente (Apriori algorithm)". È la fase più computazionalmente intensa.
2. **Generazione di strong rules (basata sulla confidenza)**: Una volta identificati tutti gli itemset frequenti, si generano le regole di associazione da questi itemset. Per ogni regola candidata, si "verifica se la confidenza della regola a sua volta supera una soglia minima" (confidence threshold). Solo le regole che superano entrambe le soglie (supporto e confidenza) sono considerate "forti".

Slide 12: Algoritmo Apriori

L'algoritmo Apriori è il metodo più conosciuto per estrarre regole di associazione forti. È suddiviso in due fasi, che riflettono quelle della slide precedente:

- **Prima fase**: Identificare i frequent itemsets in modo sistematico, senza esplorare lo spazio di tutti i candidati.
- **Seconda fase**: Estrarre le regole forti.

Il cuore dell'algoritmo Apriori si basa su un'assunzione chiave, nota come "principio di Apriori": "se un itemset è frequente, allora tutti i suoi sottoinsiemi sono anch'essi frequenti".

- Questo significa che "se l'itemset (a, b, c) è frequente, allora anche i 2-itemsets (a, b), (a, c), (b, c) e gli 1-itemsets (a), (b), (c) sono frequenti".
- La conseguenza inversa è altrettanto importante: "se un itemset (a, b, c) non è frequente, allora i suoi super-insiemi (itemsets che lo contengono) non sono frequenti". Questo è il punto chiave per l'efficienza: "se un itemset non frequente è stato identificato, eliminare tutti gli itemset che lo contengono". Questo permette di ridurre drasticamente lo spazio di ricerca e aumentare l'efficienza complessiva dell'algoritmo.

Slide 13: Esempio dell'Algoritmo Apriori

Questa slide illustra un esempio del funzionamento dell'algoritmo Apriori, utilizzando una "soglia di supporto minima (pmin) di 0.2".

1. **Inizializzazione (Iterazione 1)**: Si calcola la frequenza (relativa) per ogni singolo item (1-itemsets). Vengono considerati solo quelli la cui frequenza è maggiore o uguale a 0.2. La tabella mostra che 'a', 'b', 'c', 'd', 'e' sono tutti frequenti.
2. **Generazione di candidati 2-itemsets (Iterazione 2)**: Si combinano gli itemset frequenti della fase precedente per creare candidati 2-itemsets (es. {a,b}, {a,c}, ecc.). Si calcola la loro frequenza relativa e si scartano quelli che non superano la soglia di supporto (0.2). Per esempio, {a,d} è "not frequent" perché ha una frequenza di 1/10 = 0.1. Analogamente, {b,e}, {c,d}, {c,e} non sono frequenti.
3. **Generazione di candidati 3-itemsets (Iterazione 3)**: Si combinano i 2-itemsets frequenti della fase precedente per creare candidati 3-itemsets. Si calcola la loro frequenza. Le tabelle mostrano {a,b,c} e {a,b,d} come frequenti, entrambi con frequenza di 2/10 = 0.2.
4. **Stop del processo**: Se non si possono generare candidati itemset di dimensione superiore (es. 4-itemsets), la procedura si ferma.

Alla fine, il processo ha identificato tutti gli itemset frequenti: 5 1-itemsets, 6 2-itemsets e 2 3-itemsets, per un totale di 13 itemset frequenti. Questa è la prima fase dell'Apriori.

Slide 14: Generazione delle Regole Forti

Questa è la seconda e ultima fase del processo di scoperta delle regole di associazione.

- **Input**: La "lista di tutti i frequent itemsets", ciascuno associato alla sua frequenza relativa (che è maggiore della soglia minima di supporto). Questi sono i risultati della fase precedente (l'algoritmo Apriori).
- **Processo**: Per ogni frequent itemset, si generano tutte le possibili regole di associazione dividendo l'itemset in antecedente (L) e conseguente (H). Per ciascuna di queste regole candidate, si "calcola la confidenza".
- **Selezione**: "Se la confidenza della regola è maggiore o uguale a una soglia minima (pmin), la regola viene inclusa nella lista delle regole forti, altrimenti viene scartata".

In sintesi, questa fase filtra le regole generate dagli itemset frequenti, mantenendo solo quelle che mostrano una sufficiente "affidabilità" misurata dalla confidenza.