Dispensa di Machine Learning

Federico Luzzi
Christian Uccheddu

Indice

1 Dati

Gli ambiti più importanti nei quali vengono applicate tecniche di machine learning nella vita reale sono diversi, in particolare si usano in: finanza, sanità, agricoltura, e-commerce, social, chatbot, sensoristica (come i veicoli a guida autonoma).

Vista la grande mole di dati la necessità è capire come trattare i dati.

L'obiettivo del Machine Learning è sviluppare metodologia per dare valore ai dati in funzione di una particolare domanda che ci stiamo facendo.

Tipicamente le tecninche di machine learning si dividono nelle seguenti tre macro categorie:

- Apprendimento supervisionato o predittivo: qualcuno ha gia catalogato ad esempio delle immagini o dei dati e noi prendendo questi modelli dovremmo essere in grado di predire.
- Apprendimento descrittivo: ci sono delle funzioni obiettivo che vanno ottimizzate. Non usiamo etichette della singola istanza ma in qualche modo sappiamo dove arrivare.
- Apprendimento rinforzato: E' quello più utile in questa epoca: funziona sui premi.

In questo corso ci concentreremo sui primi due tipi di apprendimento.

L' *apprendimento supervisionato* si divide a sua volta nelle seguenti due categorie:

- Classificazione: quando il problema consiste nel dividere in classi delle quantità discrete.
- Regressione: quando il problema consiste nel ricostruire una certa variabile date delle condizioni pregresse.

L' *apprendimento non supervisionato* si divide a sua volta in:

- Clustering: Quando il problema consistere nel ricostruire delle classi delle istanze che ci vengono consegnate, senza sapere nulla sulla storia pregressa.
- Associativa: Quando il problema consiste nel scoprire pattern che descrivono bene caratteristiche associate ad un certo fenomeno.

Per alcuni compiti la correlazione statistica va benissimo, in alcuni casi però essa è addirittura deleteria. Se infatti provassi a vedere la correlazione tra il numero di omicidi in america e il numero di fondi investiti sulla ricerca scientifica vedrei che statisticamente sono strettamente correlate. Questo è un no-sense ed è il classico esempio di *correlazione spuria.*

A confermare questa visione in cui il modello per una certa trattazione dati è fondamentale è il cosidetto paradosso di Simpson.

Paradosso di Simpson: Se uso solo i dati senza modello no c'è alcun modo di scoprire la verità.

1.1 Data types

Il primo passo fondamentale è sicuramente quello di prendere confidenza coi dati. E' fondamentale capire la natura intrinseca dei dati che abbiamo a disposizione, di solito essi sono organizzati in strutture che chiamiamo **dataset**. Per le analisi con le tecniche di Machine Learning esse non sono altro che tabelle fatte da righe e da colonne.

Le colonne del dataset sono chiamate **Attributi.**

Le righedel dataset sono chiamate **Istanze.**

Ogni attributo è caratterizzato dal fatto di avere un tipo, la sua conoscenza è fondamentale perché ci permette di sapere le proprietà che questo possiede. In particolare gli attributi si dividono in due grandi gruppi:

- Categorici:
  - Nominali: ad esempio il colore degli occhi.
  - Ordinali: ad esempio possono essere i giudizi.
- Numerici:
  - Intervallo: ammettono operazioni di somma e sottrazione.
  - Ratio: possiamo applicare tutte le operazioni logico/matematiche.

	ATTRIBUTE TYPE	DESCRIPTION	EXAMPLES	OPERATIONS
CATEGORICAL (QUALITATIVE)	NOMINAL	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another (=, ≠).	Area Code	mode
			Churn	entropy
			State	contingency
			eye color	
			gender	
	ORDINAL	The values of an ordinal attribute provide enough information to order objects (<, >).	{bad, good, excellent}	median
			grades	percentiles
			street numbers	rank correlation
				run tests
				sign tests
NUMERIC (QUANTITATIVE)	INTERVAL	For interval attributes, the difference between values are meaningful, i.e., a unit of measurements exists (+, -).	calendar dates	mean
			temperature in Celsius or Fahrenheit	standard deviation
				Pearson's correlation
				t and F tests
	RATIO	For ratio attributes, both differences and ratios are meaningful, (*, /).	Day Mins	geometric mean
			Eve Mins	harmonic mean
			monetary quantities	percentiles
			length	variation
			electrical current	

Figura 1: Distinzione degli attributi con le relative proprietà

Dall'alto al basso il livello gerarchico sale e le proprietà aumentano.

C'è un'ulteriore divisione che può essere fatta ed è quella in: Si possono anche dividere in attributi *discreti* che possono essere:

- Discreti (in cui la serie dei valori è finita o una infinità numerabile), che a loro volta si suddividono in:
  - Categorici
  - Numerici
  - Binari: sono i più particolari da trattare, e hanno una serie di proprietà strane.
- Continui, i cui valori sono numeri reali.

1.2 Data exploration

Dobbiamo però anche sapere come esplorare i dati in modo intelligente usando tutti i nostri strumenti a nostra disposizione. Diamo ora un elenco degli strumenti statistici più utili che ci permetton di avere un'analisi completa dei dati.

1.2.1 Definizioni

Definizione 1.1. Si definisce **quantile** α il valore q_α che divide la popolazione in due parti proporzionali rispettivamente ad α e a $(1-\alpha)$ e caratterizzate da valori rispettivamente minori e maggiori di q_α

Un quantile molto importante è il quantile di ordine $\frac{1}{2}$ e si chiama **mediana** che è quello che ha esattamente minori di lui metà dei dati.

Definizione 1.2. Si definisce **media** il seguente valore:

$$mean = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La media non è un buon modo di visualizzare i dati perché dice poco riguardo alla loro distribuzione. Siccome la media è basata su singole osservazioni è fortemente soggetta a variazioni quando si hanno valori fortemente discostati dalla distribuzione dei dati. Tali valori si chiamano *outlier* e sono particolarmente interessanti da trattare.

Per ovviare a questo problema si è soliti usare la **media trimmed** in cui si buttano via il valore più piccolo e il valore più grande.

Definizione 1.3. Si definisce **media trimmed** il seguente valore:

$$mean_{trimmed} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad x_i = x - \max x - \min x$$

Se si trova un grosso scostamento tra la media e la media trimmed probabilmente si ha la presenza di almeno un outlier.

Definizione 1.4. Si definisce **range** il seguente valore:

$$range = \max x - \min x$$

Esso serve a quantificare la dispersione dei dati, ovviamente questo valore può essere fuorviante qualora i valori siano concentrati in una stretta banda di valori. Per ovviare a questo problema si usa la varianza.

Definizione 1.5. Si definisce **varianza** il seguente valore:

$$var = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

E' preferibile però usare la deviazione standard in quanto per come è definita risulta essere della stessa unità di misura dei dati.

Definizione 1.6. Si definisce **deviazione standard** il seguente valore:

$$std = \sigma = \sqrt{\sigma^2}$$

Essendo queste due grandezze definite a partire dalla media esse soffrono dello stesso problema: sono fortemente condizionate dagli outlier, con lo stesso metodo precedente si definiscono quindi altre due grandezze in grado di ovviare a questo problema.

Definizione 1.7. Si definisce **deviazione media assoluta** il seguente valore:

$$AAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - mean|$$

Definizione 1.8. Si definisce **deviazione mediana assoluta** il seguente valore:

$$MAD = mediana(x_1 - mean, x_2 - mean, ..., x_{n0} - mean)$$

E' utile anche definire il range interquartile (IQR) sempre per ovviare alla presenza di outliar.

Definizione 1.9. Si definisce **deviazione mediana assoluta** il seguente valore:

$$IQR = q_{75\%} - q_{25\%}$$

Ci sono anche diverse grandezze utili da definire nei casi in cui si ha a che fare con copppie di attributi, in tal caso:

Definizione 1.10. Si definisce **covarianza** il seguente valore:

$$cov(X,Y) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})$$

Per come è costruita questa matrice è necessariamente quadrata, e il suo valore ij^{th} rappresenta la covarianza tra il valore i^{th} dell'attributo x e il valore j^{th} dell'attributo y.

Un'ulteriore misura dell'associazione tra le coppie di attributi quantitativi che non dipende dalla varianza di ciascun attributo è la seguente:

Definizione 1.11. Si definisce **correlazione di Pearson** il seguente valore:

$$corr(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

Per come è definita si ha ovviamente che: $cov(x,y) \in [-1,1]$.

E' molto utile visualizzare i dati quando bisogna lavorarci, questo fondamentalmente per due motivi:

- Ci permette di trovare pattern tra le variabili che valori puntuali non ci permetterebbero di trovare.
- Ci permette di visualizzare i risultati di una lavorazione fatta sui dati.

Di seguito proponiamo un rapido elenco dei grafici più utili e descriviamone le varie caratteristiche.

1.2.2 Visualization

Possono organizzare i dati in istogrammi caratterizzati dalla presenza di bin: essi indicano la larghezza in cui i dati sono organizzati. A seconda dell'ampiezza che uso posso ottenere due disegni molto diversi come mostrato in figura.

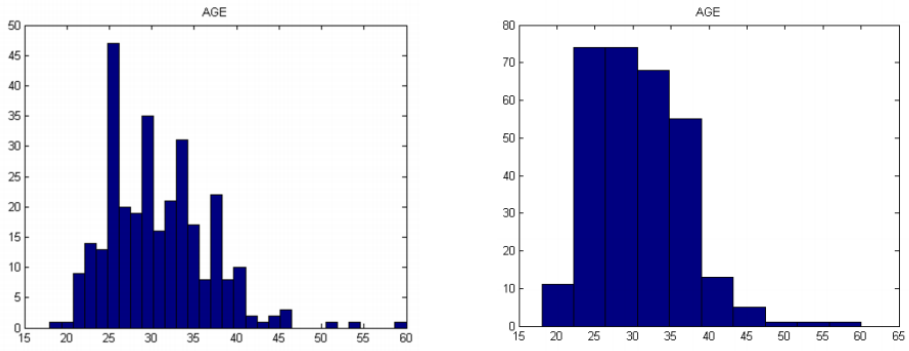

Figura 2: Differenza tra due istogrammi fatti sulla stessa distribuzione di dati ma con numero di bin diversi.

Si possono allo stesso modo creare istogrammi per dati qualitativi. La differenza rispetto agli istogrammi sui dati quantitativi è quella per cui ogni bin corrisponde ad una categoria diversa.

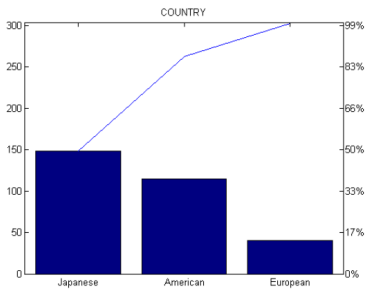

Figura 3: Istogramma di dati qualitativi.

In questo caso la linea blu disegnata sopra è una linea cumulativa, ossia rappresenta dove si trova il livello sommando tutti i dati che man mano incontro, per come è costruita ovviamente dovrà finire sul valore 100%

Un altro modo di rappresentare i dati particolarmente utile è il **grafico Box and Whiskers** applicato solo ad attributi quantitativi. Rispetto agli istogrammi questo è decisamente più utilizzato perché permette di estrarre decisamente più informazioni. In particolare proponiamo un grafico in cui sono esplicitate le informazioni ottenibili:

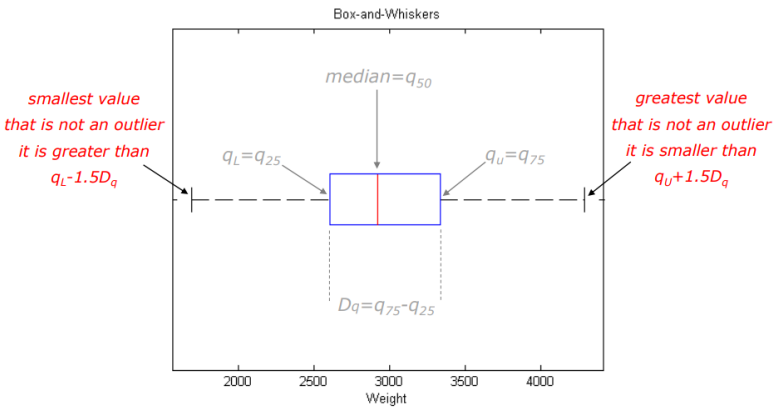

Figura 4: Informazioni ottenibili da un box plot.

1.3 Missing replacement

E' il problema che si occupa di studiare come sostituire i valori mancanti di certi attributi in un dataset. Essendo un problema di così larga portata possiamo capire bene che una trattazione esaustiva riempirebbe le ore di un intero corso, tuttavia ne forniamo le basi per essere in grado di poter effettuare una analisi dati efficace.

Le motivazioni principali del perché un database presenta delle mancanze sono le seguenti:

- Tale valore non è stato possibile misurarlo.
- Non si conosce con esattezza tale valore.
- Si è verificato qualche errore durante la presa dati.
- Quel determinato attributo fino ad un certo punto dell'analisi non è mai stato considerati importante e quindi non è mai stato registrato.

Ci sono diversi metodi che ci permettono di riempire quel valore mancante, di seguito diamo una rapida trattazione dei più usati:

- **Record removal:** è il metodo più drastico e consistere nell'eliminare tutta l'istanza che contiene quel valore. Non viene usato frequentemente perché si rischia di perdere valori che possono essere fondamentali per l'analisi dati.
- **Global constant:** consiste nel sostituire tutti i missing values con un valore costante chiamato *place holder*. Tale metodo non è molto efficiente perché un valore costante non può essere rappresentativo di una distribuzione.
- **Manual imputation:** consiste nel sostituire manualmente i missing values tramite delle osservazioni. Lo svantaggio è che è tremendamente svantaggioso dal punto di vista computazionale.
- **Moda replacement:** consiste nel rimpiazzare tutti i missing values con la moda di quel dato attributo, valgono le stesse considerazioni fatte per il metodo della global constant. Qualora gli attributi fossero continui si effettuerebbe la stessa cosa ma sostituendo la media.
- **Conditional mean replacement:** consiste nel rimpiazzare tutti i missing values con la media a condizione del fatto che sia presente un'ulteriore condizione posta su un secondo attributo. E' un po' più efficace degli altri elementi ma presenta anch'esso delle criticità.
- **Most probable:** consiste nell'eseguire una regressione sul nostro dataset utilizzando un altro attributo. In questo caso è possibile usare deei modelli di regressione anche molto complessi ed è possibile lavorare anche con dati qualitativi.

E' ora chiaro che il confine tra l'esplorazione dati e la modellizzazione degli stessi è molto sottile.

1.4 Data Preprocessing

Il preprocessing è un'area dell'analisi dati che consiste nel **rendere i dati più fruibile per una analisi dati.**

1.4.1 Aggregation

L'aggregazione consiste nel combinare due o più record in un unico oggetto. Ci sono diversi vantaggi ottenibili nell'aggregare i dati, in particolare ne elenchiamo alcuni:

- Dataset più piccoli: durante un'analisi dati abbiamo bisogno di usare il minor quantitativo di memoria e di tempo, l'aggregazione in particolare diminuisce il tempo di esecuzione di un algoritmo.
- Cambiamento di scopo: ci permette di avere una visione più ampia dei dati qualora cambiassimo lo scopo della nostra nalisi in corso d'opera.
- Varianza ridotta: gli attributi calcolati su record aggregati sono più stabili rispetto a quelli associati ai records nativi, questo per un effetto statistico.

1.4.2 Sampling

In molti casi avere una quantità enorme di dati può essere deleterio dal punto di vista computazionale, questo perché bisognerebbe usare algoritmi più semplificati in modo che la computazione possa essere svolta su un maggior numero di dati. Ciò che dobbiamo preferire è invece usare un algoritmo migliore su un dataset ridotto. Per questo entra in gioco il concetto di **campionamento.**

Il problema diventa quindi: *quando un campione è rappresentativo?*

Definizione 1.12. Un campione si dice **rappresentativo** quando ha approssimativamente le stesse proprietà del dataset di partenza.

Dobbiamo quindi trovare uno schema che ci permetta di scegliere *con grande probabilità* dei campioni rappresentativi. In questo caso il problema si riduce a trovare le appropriate:

- Dimensioni del campione.
- Tecniche di campionamento.

Esistono moltissime tecniche di campionamento ma di seguito mostreremo solo le più basilari:

- **Simple Random Sampling:** Ogni record del dataset ha la stessa probabilità di essere incluso nel campione. Tale record può essere rimosso o meno dal dataset di partenza. Quando i campioni sono molto

piccoli rispetto al dataset di partenza la rimozione o meno genera due campioni che sono molto simili.

Questo metodo fallisce quando il datset consiste di attributi qualitativi in modo che i possibili valori che possono avere hanno frequenze fortemente diverse.

- **Stratified Sampling:** Vengono presi record in modo che all'interno del campione vengano rispettate le proporzioni tra gli attributi presenti nel dataset di partenza.

Una volta scelta la tecnica di campionamento dobbiamo occuparci di scegliere la grandezza del campione. Ovviamente tenere un campione di grande ampiezza aumenta la probabilità che un campione sia rappresentativo, di contro elimina la maggior parte dei vantaggi computazionali di avere un campionamento. Avere un campione troppo piccolo invece manda in contro al rischio di eliminare pattern potenzialmente importanti o addirittura di mantenere pattern erronei. La giusta dimensione del campionamento va scelta in base al nostro dataset di riferimento effettuando delle prove.

1.4.3 Dimensionality reduction

In diversi casici capiterà di analizzare dataset caratterizzati da un gran numero di attributi. Diminuirne il numero porta a diversi vantaggi:

- Molti algoritmi lavorano se la dimensionalità è minore.
- L'interpretabilità del modello implementato aumenta perché dipende da un numero minore di attributi.
- La rappresentazione grafica dei dati è facilitata.
- L'ammontare di tempo e memoria diminuisce drasticamente.

Contrariamente a ciò che ci si aspetterebbe *avere una dimensionalità maggiore dei dati non implica un aumento delle performance.* Mostriamo questo dato con il sgeuente grafico.

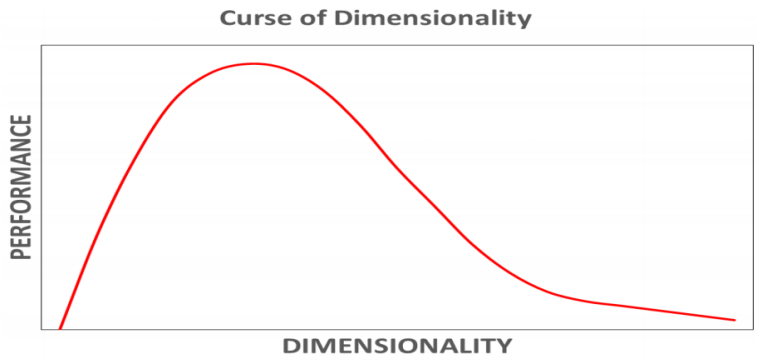

Figura 5: Andamento delle performance di un algoritmo in funzione della dimensionalità del dataset.

Molte delle tecniche usate per ridurre le dimensionalità sfruttano tecniche prese dall'algebra lineare per proiettare i dati da uno spazio dimensionale maggiore ad uno spazio dimensionale minore.

La **Principal Component Analysis (PCA)** trova nuovi attributi in modo che:

- Siano *combinazione lineari* degli attributi di partenza.
- Siano *mutualmente ortogonali.*
- Catturino il *massimo ammontare di variabilità* nei dati.

In particolare la **Singular Value Decomposition (SVD)** è una tecnica di algebra lineare correlata alla PCA ed è largamente usata per ridurre la dimensionalità dei dataset.

Definizione 1.13. Si chiama **binarizzazione** la tecnica di trasformazione di attributi continui e categorici in uno o più attributi binari

Per farlo associamo k possibili valori a k valori interi nell'intervallo $[0, k-1]$. Successivamente trasformiamo questi k interi in un numero binario, come sappiamo il numero di cifre necessarie per rappresentare un numero intero son le sguenti:

$$s = [\log_2 k]$$

In questo caso è necessario introdurre un attributo binario per ogni attributo categorico con cui il dato può manifestarsi. Ovviamente si è più interessati ai casi in cui si manifesta il valore 1 perché indica la presenza di tale attributo.

Quando invece abbiamo a che fare con attributi durante un'analisi di classificazione o di associazione ricorrdiamo alla tecninca di **discretizzazione.** Come facilmente intuibile la miglior discretizzazione dipende dal tipo di algoritmo che sto utilizzando.

La discritezzazione può essere:

- **Supervisionata.**
- **Non-supervisionata.**

La *discritezzizazione supervisionata* utilizza ulteriori informazioni (attributi di classe) per discretizzare gli attributi. Questa tecninca divide i punti in modo che qualche misura di *purità* sia massimizzata. La misura di purità più spesso applicata è l'**entropia.**

$$e_i = -\sum_{k=1}^{K} p_{ki} \log_2 p_{ki}$$

Per come è costruita si ha che:

- Se contiene solo record di una data classe $\implies e_i = 0$, massima purità.
- Se contiene egualmente spesso tutte le class $\implies e_i = \max$, minima purità.

La discretizzazione supervisionata basata sull'entropia permette di trovare i punti di divisione degli attributi *continui* tali che l'entropia totale sia minimizzata.

Definizione 1.14. Si definisce entropia totale la seguente espressione:

$$E = \sum_{i=1}^{n} w_i e_i$$

In cui:

$$w_i = \frac{m_i}{m}$$

E le variabili sono così definite:

- n: numero di intervalli.
- m: numero di record.
- m_i: numero di record nell'intervallo i-esimo.

La *discretizzazione non-supervisionata* non utilizza alcuna informazione eccetto il valore dell'attributo continuo da discretizzare. Possono essere a loro volta divisi in due categorie:

- **equal width unsupervised discretization:** gli intervalli in cui viene discretizzati hanno tutti la stessa ampiezza
- **equal frequency unsupervised discretization:** gli intervalli in cui viene discretizzato hanno approssimativamente la stessa frequenza.

Può succedere però che gli attributi categorici abbiano troppi valori. Se gli attributi in questione sono ordinali allora si applicano le tecniche viste in precedenza, se invece sono nominali allora bisogna trovare un altro approccio. In questo caso possiamo infatti raggruppare i valori solo se il raggruppamento si traduce in un miglioramento nelle performance di classificazione o nel raggiungimento di qualche obiettivo del trattamento dati.

1.4.4 Variable transformation

Definizione 1.15. Una trasformazione di variabile sè una trasformazione che è applicata a tutti i valori di quella variabile.

Ci sono due tipi di trasformazione di variabile:

- **Funzioni semplici**: una semplice funzione matematica è applicata a tutti i valori di una variabile individualmente. Bisogna stare attenti all'ordine in cui la applichiamo e soprattutto a come si comporta la funzione per valori negativi e vicini allo 0.
- **Standardizzazione:** trasforma tutto il dataset in modo che acquisiscano una particolare proprietà.

  $$Z = \frac{x - \mu}{\sigma}$$

  E' molto importante perché il nuovo dataset avrà $\mu = 0$ e $\sigma = 1$ per come l'ho costruito. In questo modo la somma di diversi attributi continui permette ad uno o a pochi attributi di prendere grandi valori e di dominare il nuovo attributo somma. E' molto utile quindi perché fa saltare immediatamente all'occhio la presenza di outlier.

  In questo caso la media è sostituita dalla mediana, e la deviazione standard è sostituita dalla AAD che ricordiamo essere definita come:

  $$\sigma_x = \frac{1}{m} \sum_{i=1}^{m} |x_i - \mu|$$

2 Association Analysis

2.1 Introduction (*)

L'analisi di associazione ci riporta al concetto di causalità di variabili: dati certi valori di attributi cosa posso dire del valore di un altro attributo?

Le regole associative ci permettono di prendere delle scelte molto operative in diversi ambiti, in particolare in quello che viene chiamato *market busket analysis*. Il problema del carrello tratta il posizionamento di prodotti in un negozio. Si sa che all'acquisto di un certo prodotto si tende a acquistare altri prodotti connessi, quindi si cercano queste associazioni basandosi sul carrello della spesa dei clienti (appunto market basket) per poi capire come impostare la disposizione sugli scaffali.

Obiettivo: identificare quali siano gli **item associati** per poter prendere delle decisioni. In sostanza si generano **Regole associative** formate da coppie di insiemi *antecedente* e *conseguente*.

es. {Beer} → {swiss cheese}

{antecedente} → {conseguente}

NB: non è una causalità ma un'associazione

L'analisi si basa sullo studio di due diversi dataset:

- Product set: contiene informazioni legate ai prodotti come il nome e il prezzo
- Transaction set: contiene informazioni legate agli acquisti dei clienti, ogni record corrisponde ai prodotti presenti nel carrello del cliente

Si organizza il dataset delle transazioni in formato binario, ovvero ogni colonna indica se in una data transazione un certo prodotto sia o meno presente.

Transaction ID	swiss cheese	cherry coke	bio coke	Peppers	scrambled egg	Pomegranate	strawberries
0	1	0	0	0	1	0	0
1	0	1	0	0	0	0	1
2	0	0	0	1	1	0	0
3	0	0	0	0	0	1	0
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

Figura 6: esempio: transaction set binarizzato

Consideriamo:
$I = \{i_1, i_2, ..., i_d\}$ il set di tutti gli item nel market basket
$T = \{t_1, t_2, ..., t_N\}$ l'insieme di tutte le transazioni

Definizione 2.1. Una collezione di zero o più item è chiamata **Itemset**.

Definizione 2.2. Se un itemset contiene k item è detto **K-Itemset**.

Definizione 2.3. L'itemset che non contiene alcun elemento è detto **empty set**.

Definizione 2.4. La **transaction width** è il numero di item presenti in una transazione

Una transazione t_j contiene l'itemset X, se X è un sottoinsieme di t_j.

Definizione 2.5. Il **Support count** è il numero di transazioni che contentono uno specifico itemset.

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}$$

Definizione 2.6. Una **regola di associazione** viene rappresentata come:
$X \to Y$
dove X e Y sono itemset disgiunti ($X \cap Y = \emptyset$).

Una regola viene valutata in termini di *supporto* e di *confidenza*.

Definizione 2.7. Support determina quanto spesso una regola sia applicabile dato un data set:

$$s\{x \to Y\} = \frac{\sigma(X \cup Y)}{N}$$

Definizione 2.8. Confidence determina quanto frequente Y è presente in una transazione che contiene X (si assume che l'universo sia rappresentato partendo da X):

$$c\{x \to Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Es.
$X = \{swisscheese, cheddar\}$, $Y = \{dietcoke\}$
Assumiamo che:

- $\sigma(x) = 8$ (support count)

- $N = 20$ (numero di transazioni)
- $\sigma(x, y) = 6$

Allora il support e la confidence della regola $X \to Y$ sono:

$$s\{X \to Y\} = \frac{\sigma(X \cup Y)}{N} = \frac{6}{20} = 0.3$$

$$c\{x \to Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{6}{8} = 0.75$$

Perchè utiliziamo il supporto:

- se troppo basso potrebbe esserci un'associazione casuale
- potrebbe non valere la pena seguire associazioni che si applicano in modo poco significativo dal punto di vista dei profitti

Supporto utilizzato per eliminare regole non desiderate e condivide interessanti proprietà che possono essere sfruttate per la ricerca di regole associative efficaci.

La confidenza è molto importante perchè misura l'affidabilità dell'inferenza e:

- un alta confidenza significa che Y sarà molto presente in transazioni con X
- si stima la probabilità condizionata di Y dato X

Definizione 2.9. Association Rule Mining Problem può essere formalmente definito come: dato un set di transazioni T, trovare tutte le regole con $support \geq minsup$ e $confidence \geq minconf$, dove $minsup$ e $minconf$ sono i threshold corrispondenti alle due misure.

Approccio a forza bruta di association rule non è molto praticabile in quanto i tempi di computazione aumentano in modo esponenziale:
$R = 3^d - 2^{d+1} + 1$.

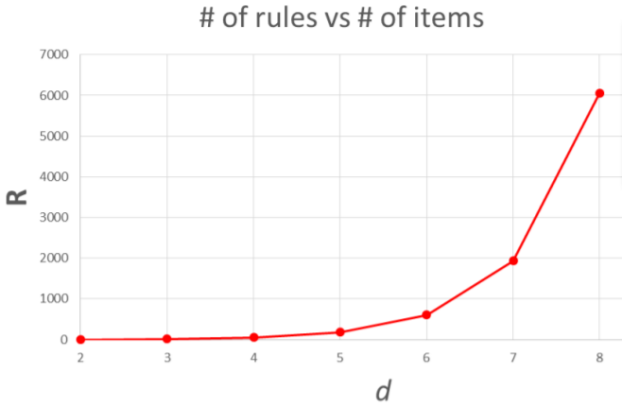

Figura 7: numero di regole calcolate in base al numero di itemset

Una strategia comunemente adottata in molti algoritmi è decomporre il problema in 2 grandi supertask:

- *Generazione dei frequenti itemset*: specifichiamo tutte e sole quelle regole per cui il supporto è maggiore del *minsup*, gli itemset generati sono chiamati **Frequent Itemset**
- *Generazione delle regole*: estriamo tutte le regole con alta confidenza (maggiore di $minconf$) dai Frequent Itemset trovati precedentemente, queste regole vengono chiamate **Strong Rules**

La complessità maggiore è richiesta dalla generazione dei Frequent Itemset.

2.2 Rule Extraction

Per comprendere l'inefficienza della generazione con l'approccio forza bruta pensiamo a questo esempio:

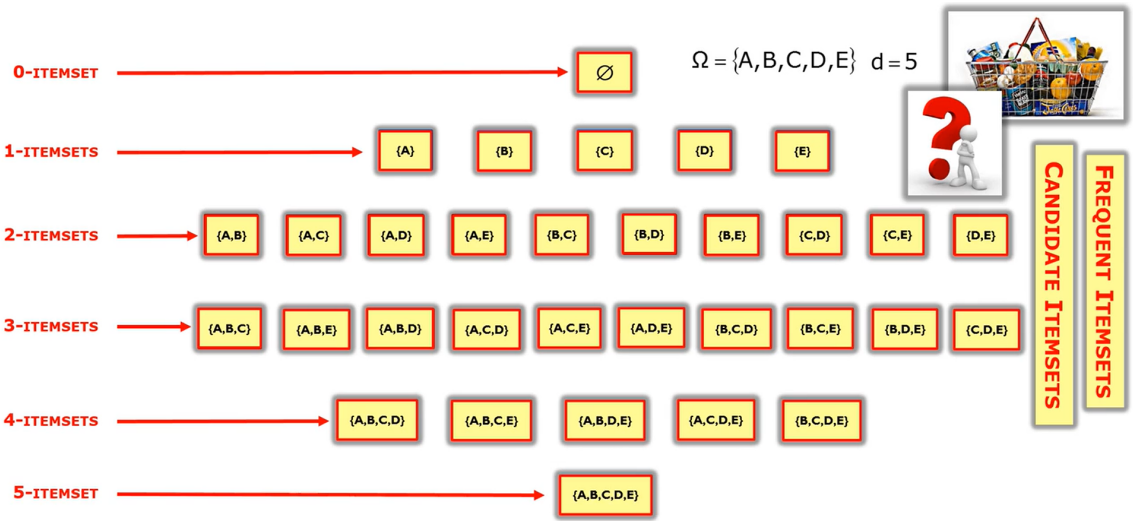

Figura 8: k-itemset brute-force

Definizione 2.10. Candidate Itemset: l'insieme di tutti gli itemset che possiamo formare. Avranno un numero di item diverso.

Nel nostro caso il numero di itemset candidato è: $M = 2^d - 1 = 2^5 - 1 = 31$

Come si può notare abbiamo un sistema a doppio cono che è tipica delle distribuzioni binomiali. Una volta che gli abbiamo considerati tutti ci interessano solo i più frequenti.

Se usassimo la forza bruta dovremmo calcolare per ogni itemset candidato il suo support count, e vedere se il suo supporto lo configura come un itemset frequente (molto dispendioso).

I confronti da effettuare sono nell'ordine di $O(NMw)$, dove:

- N = numero di transazioni
- M = numero di itemset candidato
- w = massima lunghezza delle transazioni

È decisamente troppo come numeri confronti contando che molti dei quali sono inutili o poco significativi.

Vi sono due approcci per ridurre il costo computazionale della generazione di itemset frequenti:

- ridurre il numero di candidati itemset (M). Il principio Apriori è un metodo per eliminare alcuni candidati itemset senza contare il support count.
- riduce il numero di confronti anzichè controllare tutte le possibili combinazioni, lo si fa con strutture dati avanzate

Principio Apriori se un itemset è frequente, allora tutti i suoi sottoinsiemi sono frequenti.
Quindi se una regola ha una frequenza bassa allora tutte le regole che prevedono come sottoinsieme la stessa non supereranno quella frequenza pertanto è inutile considerarle. Si procede attraverso il **pruning** dell'albero delle sequenze per queste soluzioni, viene chiamato **support-based pruning** (vedi immagine).

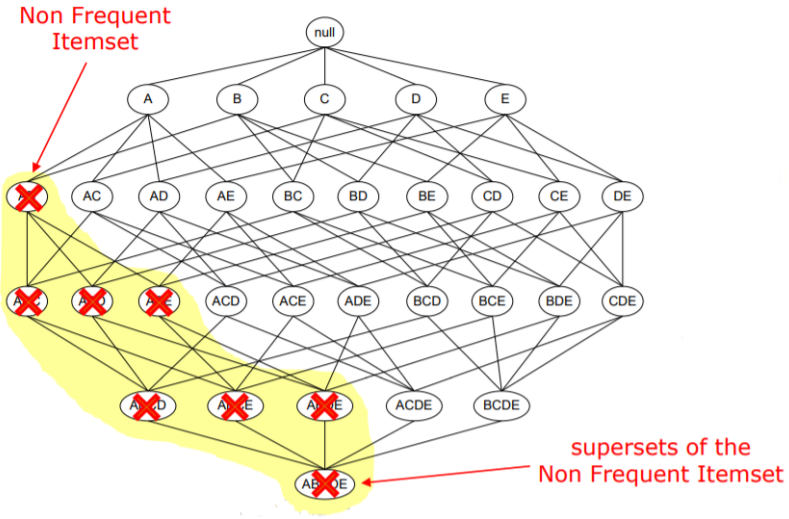

Figura 9: esempio di support-based pruning

2.2.1 Algoritmo apriori

Genera due operazioni:

- **Candidate Generation**: genera nuovi candidati k-itemset basati su (k-1)-itemset frequenti calcolati nella precedente iterazione
- **Candidate Pruning**: questa operazione elimina alcuni candidati k-itemset usando la strategia del support-based pruning

La complessità computazionale soffre di 4 limiti:

- *Support threshold*: la soglia di supporto se troppo bassa non taglio molto l'albero, però non deve essere neanche troppo elevata altrimenti non considero associazioni rilevanti.
- *Numero di item (dimensionalità)*: se il numero di item cresce, ci sarà bisogno di più spazio in memoria per registrare il support count degli

item, inoltre bisogna considerare anche il costo dell'I/O per passare i dati.

- *Numero di transazioni*: l'algoritmo scorre più volte tutta la lista di transazioni, pertanto un numero alto di transazioni inficia sui tempi.
- *Avarage transaction width*: per dataset densi la lunghezza media delle transazioni tende ad essere grande. La massima lunghezza degli itemset frequenti tende ad aumentare quindi più sequenze candidato devono essere esaminate durante la generazione e support counting. In aggiunta aumenta il numero di archi traversi nell'albero durante il support counting.

2.3 Maximal/Closed Frequent Itemsets

2.3.1 Rule Generation

Ogni k-itemset frequente, Y può generare al limite $2^k - 2$ regole di associazione.

Una regola di associazione può essere estratta partizionando l'itemset Y in 2 sottoinsiemi non vuoti $\{X\}$ e $\{Y - X\}$, tale che $X \rightarrow Y - X$ soddisfa il threshold di confidenza.

NB: Tutte le regole generate da itemset frequenti sono esse stesse frequenti.

In pratica, il numero di itemset frequenti prodotti da transazioni possono essere molto grandi. È utile identificare itemset rappresentativi e piccoli con i quali derivare gli itemset grandi. Per questo si ragiona in due rappresentazioni:

1. Maximal Frequent Itemset
2. Closed Frequent Itemset

Definizione 2.11. Maximal Frequent Itemset è definito come un Itemset Frequente per il quale nessuno dei suoi soprainsiemi immediati sono frequenti.

Maximal Frequent Itemset: per ogni nodo si verifica il vincolo della frequenza e si definisce la frontiera dove un nodo non gode più di questa proprietà, corrisponde alla massima frontiera in cui mi posso spingere.

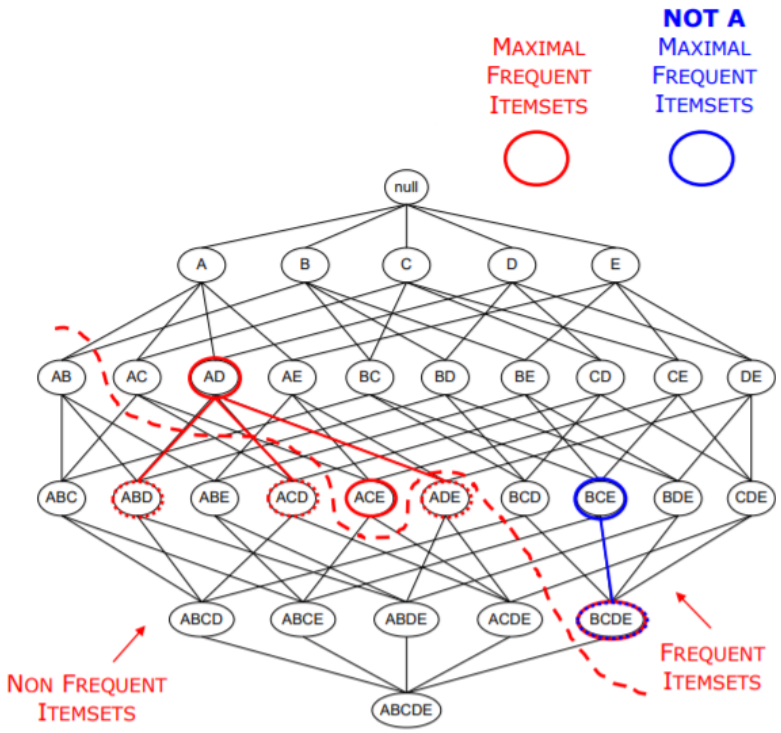

Figura 10: esempio di frontiera di maximal frequent itemset

Nella pratica: Un nodo è un Maximal Frequent Itemset se è frequente e se tutte le sue estensioni non sono frequenti.

- Fornisce una rappresentazione compatta dell'insieme di itemset che cerchiamo, la più piccola espressione in cui gli itemset sono derivabili
- Calcolato dal più piccolo insieme di itemset dal quale tutti gli itemset frequenti possono essere derivati
- È praticabile solo l'algoritmo efficiente usato esplicita la ricerca dei maximal frequent itemset senza numerare tutti i suoi sottoinsiemi

Per costruzione *però* non ci dice quanto è il supporto rispetto ai suoi sottoinsiemi. In alcuni casi potrebbe servire avere una minima rappresentazione degli itemset frequenti che preservano l'informazione sul supporto.

Definizione 2.12. Un itemset X è **Closed Frequent Itemset** se nessun immediato superinsieme ha esattamente lo stesso support count di X. In ogni caso il suo spporto deve essere $\geq minsup$.

Importante qunado vi sono gruppi di prodotti venduti a blocco ignorando gli altri.

Es

TID	a1	a2	a3	a4	a5	b1	b2	b3	b4	b5	c1	c2	c3	c4	c5
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figura 11: esempio di gruppi closed frequent itemset

Come si può notare in figura ogni gruppo di variabili (Gruppo A, B e C) è perfettamente associato e non hanno bisogno di mostrare items collegati ad altri gruppi. Assumendo che il $minsup = 20\%$, il numero totale di itemset frequenti è: $3 \cdot (2^5 - 1) = 93$. In ogni caso vi sono solo 3 closed frequent itemset:

$\{a1, a2, a3, a4, a5\}$

$\{b1, b2, b3, b4, b5\}$

$\{c1, c2, c3, c4, c5\}$

Questo tipo di itemset sono utili per rimuovere **regole di associazione ridondanti**.

Definizione 2.13. Una regola di associazione $X \rightarrow Y$ è **ridondante** se esiste un'altra regola $X' \rightarrow Y'$ che rispetta certe proprietà.

- $X \subseteq X'$
- $Y \subseteq Y'$
- $s(X \rightarrow Y) = s(X' \rightarrow Y')$
- $c(X \rightarrow Y) = c(X' \rightarrow Y')$

Mostriamo ora la gerarchia dei frequent itemset:

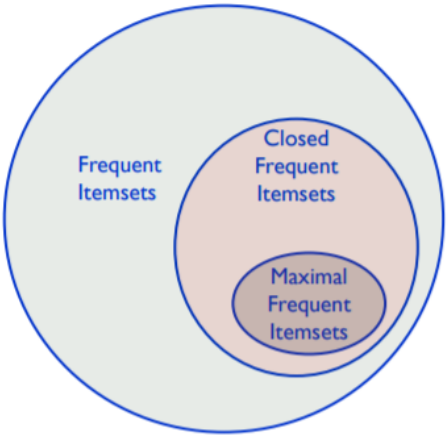

Figura 12: gerarchia dei frequent itemset

Come si può notare i Maximal Frequent Itemset sono inclusi nei Closed Frequent Itemset perchè nessun maxmal può avere lo stesso support count del suo immediato superset.

2.4 Rules Evaluation (*)

Generati gli insiemi di pattern potenzialmente utili, bisogna ordinarli in base al loro livello di attrattività per il dominio di applicazione. La valutazione della qualità viene effettuata seconodo due criteri:

- **Statistical arguments**: per patterns di items indipendenti e coperti da poche transazioni. Il problema è che *possono catturare relazioni spurie* nei dati, per evitare:
  - Objective Interestingness Measure: usare statistiche derivate dai dati per determinare quali pattern sono interessanti
  - Supporto, confidenza e correlazione
- **Subjective arguments**: un pattern è considerato non interessante a meno che riveli informazioni inaspettate riguardo ai dati e alla conoscenza: es. forte associazione tra acquisto di pannolini e birra. È difficile fare questo tipo di valutazioni, richiede una considerevole quantità di informazioni pregresse dagli esperti di domino.

Meglio cercare di applicare un approccio più *oggettivo* alla valutazione: data-driven, indipendente dal dominio in cui si richiede un minimo input dagli utenti (solo dei threshold) e calcolato basandosi sulle frequenze calcolate in una **Contingency Table**.

	B	not B	
A	f_{11}	f_{10}	f_{1+}
not A	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Figura 13: contingency table

dove:

- f_{1+} = support count di A
- f_{+1} = support count di B
- N = nro di transazioni totali

Supponiamo che un direttore di un minimarket voglia analizzare la relazione tra le persone che bevono tè e persone che bevono caffè. La seguente contingency table è ottenuta considerando le transazioni disponibili:

	coffee	**not coffee**	
tea	*150*	*50*	**200**
not tea	*650*	*150*	**800**
	800	**200**	1,000

Figura 14: contingency table

Ora valutiamo l'associazione: tea $\rightarrow$ coffee

- $support = \frac{f_{11}}{N} = \frac{150}{1000} = 15\%$
- $confidence = \frac{f_{11}}{f_{1+}} = \frac{150}{200} = 75\%$

Ad una prima occhiata sembrerebbe che le persone che bevono tè tendono a bere anche caffè (vedi confidenza). Ma se notiamo le persone che bevono caffè, a prescindere dal tè sono l'80% ($\frac{800}{1000}$), mentre la frazione dei bevitori di tè che bevono caffè è solo il 75%.

Da questo ragionamento si può concludere che il fatto di bere tè non influisca sulle persone che bevono caffè. Infatti nonostante l'associazione abbia un alto livello di confidenza (75%) non si può ignorare il supporto dell'itemset conseguente (80%).

Pertanto, vengono definiti altri indici:

Definizione 2.14. Lift: tasso di confidenza rispetto al supporto del conseguente

$$Lift = \frac{c(A \to B)}{s(B)}$$

Definizione 2.15. Interest Factor: equivalente al *Lift* ma per attributi binari, è definito in questo modo:

$$I(A, B) = \frac{s(A, B)}{s(A)s(B)} = \frac{Nf_{11}}{f_{1+}f_{+1}}$$

I valori sono così classificati:

- $= 1$ se A e B sono indipendenti
- > 1 se A e B sono positivamente associati
- < 1 se A e B sono negativamente associati

Definizione 2.16. Analisi di correlazione (per attributi binari simmetrici): analizza la relazione tra coppie di attributi. Attributi continui possono essere analizzati con la correlazione di Pearson, la correlazione per attributi binari è misurata usando il ϕ-coefficient:

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

il valore varia da $[-1, +1]$. Se gli attributi sono statisticamente indipendenti il valore è 0.

Definizione 2.17. IS Measure (per attributi binari asimmetrici)

$$IS(A, B) = \sqrt{I(A, B)s(A, B)} = \frac{s(A, B)}{\sqrt{s(A)s(B)}}$$

Il suo valore è grande quanto l'*Interest Factor* e il supporto sono grandi. Se A e B sono indipendenti:

$$IS(A, B) = \sqrt{s(A)s(B)}$$

ha lo stesso problema della correlazione, il valore può essere grande anche per associazioni incorrelate o negativamente correlate.

Vi sono altri tipi di misure per l'analisi delle relazioni tra coppie di variabili binarie.

Definizione 2.18. Misure Simmetriche
Una misura **M** è **Simmetrica** se: $M(A \to B) = M(B \to A)$
es. Interest Factor è simmetrico

Definizione 2.19. Misure Asimmetriche
Una misura **M** è **Asimmetrica** se: $M(A \to B) \neq M(B \to A)$
es. la Confidenza è asimmetrica

Di seguito gli indici più utilizzati:

Simmetrico	Asimmetrico
Correlazione (ϕ)	Gini index
Odds ratio	Mutual information
Kappa	Certainty factor
Interest (I)	Added value
Cosine (IS)	J-measure
Jaccard	Goodman-Kruskal
Collective strength	

È importante capire che ciascuna misura è adatta per analizzare un certo tipo di associazioni come basket market analysis o document analysis. In base al caso bisogna utilizzare gli indici migliori.

NB: sono stati presentati indici relativi a valutazioni per coppie di attributi binari, ma è possibile estendere l'analisi a più di due attributi usando tabelle delle frequenti in una Contingency Table multi-dimensionale. Gli indici come il Supporto Interest Factor e IS si prestano a ciò.

2.5 Simpson's Paradox

Consideriamo la seguente situazione:

un direttore di un mini-market racconta di una curiosa scoperta, legata agli item birra e hot dogs. Una società di consulenza da lui pagata per analizzare i suoi dati di vendita ha scoperto che i clienti che comprano birra sono meno tentati di acquistare hot dogs rispetto a quelli che non acquistano la birra. Il direttore però è convinto del contrario, lo sa per esperienza professionale che chi acquista birra tende ad acquistare hot dogs.

Come si può risolvere il paradosso?
Consideriamo la seguente tabella:

	hot dogs	NOT hot dogs	Total
beer	70	98	168
NOT beer	45	55	100
	115	153	

Figura 15: es. birra - hot dog

Consideriamo le seguenti regole con le relative confidenze:

- {beer} → {hot dogs} - confidence = 42%
- {NOT beer} → {hot dogs} - confidence = 45%

Possiamo inferire che i clienti che acquistano birra sono meno inclini (42%) ad acquistare hot dog rispetto a quelli che non acquistano birra (45%).

Analizziamo ora gli stessi dati ma categorizzando se il cliente è single o no:

		hot dogs	NOT hot dogs	Total
single	beer	30	10	40
	NOT beer	5	40	45
NOT single	beer	40	88	128
	NOT beer	40	15	55

Figura 16: es. cliente single: birra - hot dog

Pertanto posso derivare queste due coppie di regole:

- Single
  - {beer} → {hot dogs} - confidence = 75%
  - {NOT beer} → {hot dogs} - confidence = 11%
- NOT Single
  - {beer} → {hot dogs} - confidence = 31%
  - {NOT beer} → {hot dogs} - confidence = 73%

Da questi dati posso affermare che: i single che acqustano birra sono più inclini (75%) ad acquistare anche hot dog rispetto a quelli che non acquistano birra (11%).

Quindi sia l'azienda di consulenza che il direttore del mini-market avevano ragione soltanto che per comprenderlo bisognava categorizzare i clienti.

Paradosso di Simpson o Yule-Simpson effect, è un paradosso della probabilità e statistica, in cui una tendenza appare in diversi gruppi di dati ma sparisce o si inverte quando questi gruppi sono combinati.

Bisogna applicare una appropriata stratificazione dei dati per evitare la generazione di associazioni spurie risultati dal paradosso di Simpson.

Es. per i dati del market-basket, la catena di supermercati dovrebbe stratificarli secondo la location del negozio, mentre i dati medici da vari pazienti dovrebbero essere stratificati secondo fattori quali l'età o il genere.