

## Estrazione dati

- Altre attività:
- Data cleaning utili, come ~~preprocessing~~ (rimozione dati non necessarie)
  - Visualizzazione dati

## Pattern di classe:

- Valori: veri per tutti i dati
  - Utile
  - Inutile
  - Congruibili
- 
- Devono avere qualche caratteristica

Un diagramma di clustering è

- Descrittivo: identificare pattern comprensibili e descrivibili dei dati
- Preddittivo: creare un modello per predire i dati futuri.

## I dati con cui si ha a che fare

- Grotti
- Eleganti dimensioni (molti attributi)
- Confini

Si usano strumenti avanzati per le tracce automatiche dei dati

Ottenerli è la principale analisi.

Dati  $\neq$  componenti

Dopo l'analisi dei dati, veniamo a patti con i dati

Il Dato Numpy non è solo estremo Tanti  
dati e analizzati Trammi, bisogna estremare  
i dati BENE, quelli che servono.

Non riferisco estremi dati inutili per  
analizzare,abbiamo estremi a rigore  
che servono essere utili

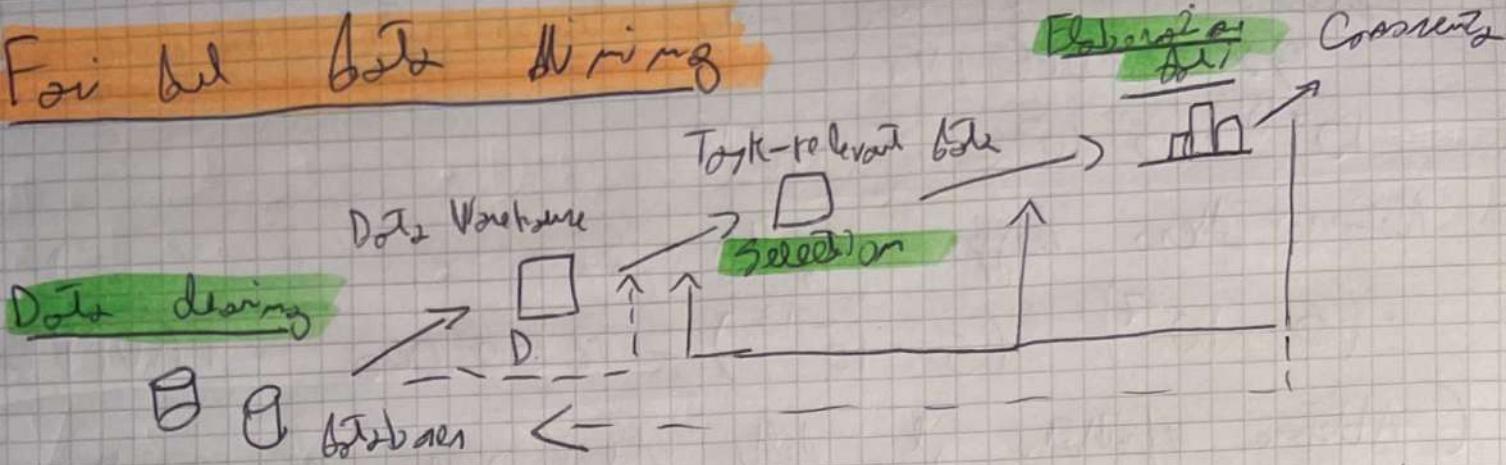
Avrò troppi dati se i confronti

Piaggio & Brambilla: se il numero di occorrenze  
altro, nell'ipotesi di bimodale, è impossibile  
già alto del numero di frequenze così che si  
riesce di trovare altro le informazioni alcuna  
posta essere considerate spazzatura

Dunque condiziona una proprietà cui dati abbiano  
essere vicini che questa proprietà non venga  
oversata troppo volte nel caso in cui i dati  
sono bimodali

Potremo dire che c'è un  $N_2$  un lower  
e Upper bound del numero di dati da  
analizzare

## Fai del best mining



Data troppo grandi  $\leftrightarrow$  Eseguo di applicazione di Boltzman

- $10^9$  persone trascorre
- Ogni  $\approx 1\%$  del tempo in hotel
- $10^7$  persone in hotel al giorno
- ovvero  $10^{10}$  persone in hotel in 1000 giorni
- $10^5$  hotel, ognuno con 100 persone
- Quante coppie al giorno in un intero hotel in due giorni diversi?
- Prob. una persona vada in un hotel  $\frac{1}{100}$
- P e Q vanno nello stesso hotel stessa giorno

$$\frac{1}{100} \cdot \frac{1}{100} \cdot \frac{1}{10^5} = \frac{1}{10^9} = 10^{-9} \rightarrow \text{probabilità 1000 persone in 2 hotel}$$

$$- \text{probabilità 1000 persone in 2 giorni diversi} \quad 10^{-9} \cdot 10^{-9} = 10^{-18}$$

$$\text{Coppie giornaliari} \quad \binom{1000}{2} = \frac{1000!}{2! \cdot (998)!} = 5 \cdot 10^{15}$$

$$\text{come persone distinte} \quad \frac{10^9 \cdot (10^9 - 1)}{2} \cong 5 \cdot 10^{17}$$

## Calcolo delle probabilità

Evento: probabilità che avvenga o non avvenga per un esperimento

Esperimento: Sarà testo  $\rightarrow$  Evento: cada voci

Variabile aleatoria: quantità numerica che descrive il possibile esito di un evento

L'insieme ~~fatto~~ insieme dei valori possibili della variabile  $X$  è detto range  $R(X)$

Range continuo  $\rightarrow$  Variabile aleatoria continua  
// Discreto  $\rightarrow$  // // discreta

Alla ogni variabile  $A$  e un valore  $\alpha$  associano una probabilità  $P(A=\alpha)$   $\rightarrow$  indica quanta è la probabilità che  $A=\alpha$

A volte si dice  $P(\alpha)$

Distribuzione di probabilità  $P(X)$ : insieme delle probabilità che  $X$  assume bei certi valori (uno b. en.)

Distribuzione cumulativa: trovare le probabilità che  $X$  assume valori minori o uguali a  $x$

$$\bullet F(x=x) = \sum_{x' \leq x} P(X=x')$$

Per una variabile continua non vi può parlare di distribuzione (i valori sono infiniti).

Si parla di densità di probabilità (L e R potrebbero essere  $-\infty$  e  $\infty$ )

Dato  $X$  e  $L, R$  estremi inferiori e superiori della range

$n$  intere la funzione di densità  $f(x)$ .

La densità di probabilità in  $(a, b)$  con

$$L \leq a \leq b \leq R$$

$$\bullet P(a < X < b) = \int_a^b f(x) dx$$

Funzione cumulativa: la probabilità: probabilità che  $X$  sia  $\leq x$

$$\bullet F(x \leq x) = \int_L^x f(x) dx$$

## Proprietà Probabilità

Nel discreto

$$\bullet P(X=x) \geq 0, \forall x \in R(X)$$

$$\bullet \sum_{x \in R(X)} P(X=x) = 1$$

Nel continuo

$$\bullet f(x) \geq 0 \quad \forall x \in R(X)$$

$$\bullet \int_L^R f(x) dx = 1$$

Normalizzazione

Le probabilità assumono valori tra 0 e 1 escluso.

0 → valore non possibile

1 → certo

Di solito si cerca lavoro più variati.  
Si cercano modelli più variati alle stesse condizioni contemporaneamente

### Probabilità di interazione

$P(X=x, Y=y)$  oppure  $P(X=x \text{ and } Y=y)$   
Ovvero probabilità che accadano entrambi così

### Unione

$P(X=x \text{ or } Y=y)$  o uno dei due eventi  
Come per la legge degli insiem

$$P(X=x \text{ or } Y=y) = P(X=x) + P(Y=y) - P(X=x, Y=y)$$

### Distribuzione di prob. Condiz.

$P(X, Y)$ : insieme probabilità che  $X$  e  $Y$  assumano certi valori contemporaneamente (in  $R(X)$  e  $R(Y)$  riguardante)

$P(x, y)$  bba  $\forall x$

$$\cdot P(x, y) \geq 0$$

$$\cdot \sum_{x \in R(X)} \sum_{y \in R(Y)} P(x, y) = 1$$

Nel caso continuo  $\rightarrow$  densità di probabilità continua

$f(x, y)$

$$\cdot f(x, y) \geq 0$$

$$\cdot \int_x \int_y f(x, y) dx dy = 1$$

## Distribuzione marginale

La distribuzione marginale fornisce informazioni anche sulle componenti di una singola variabile.

Distribuzione marginale  $\rightarrow$  distribuzione di una delle sue variabili scelte e pertanto dalla distribuzione congiunta

Questo process è detto marginalizzazione

Nel discutere

$$P(X=x) = \sum_{Y \in R(Y)} P(X=x, Y=y)$$

$$P(Y=y) = \sum_{X \in R(X)} P(X=x, Y=y)$$

Nel continuo  $\rightarrow$  integrazione al posto di somma

Dato:

$$P(X=z) = \sum_{Y \in \mathbb{R} \cup \{\infty\}} P(X=z, Y=y) \quad \text{ENIGMA}$$

La marginalizzazione si estende a 3 o più variabili

$$P(X_1=x_1, \dots, X_n=x_n) = \sum_{x_i \in R(X_i)} P(X_1=x_1, \dots, X_n=x_n)$$

Le variabili da marginalizzare possono anche 2 o più

post.

## Probabilità Condizionale

Dati  $X = x$  e  $Y = y$  con  $y$  da  $t_2$  in alto valore

La prob. dc  $X=x$  b22  $Y=y$  è probabilità condizionale  
e denotata com  $P(X=x | Y=y)$

$$P(X=x | Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

$$\cdot P(X=x | Y=y) \neq P(Y=y | X=x)$$

Si può generalizzare a  $n$  variabili

$$P(X_{k+1}=x_{k+1}, \dots, X_n=x_n | X_1=x_1, \dots, X_k=x_k) = \frac{P(X_1=x_1, \dots, X_n=x_n)}{P(X_1=x_1, \dots, X_k=x_k)}$$

## Terremoto di Bayes

- $P(X=x, Y=y) = P(Y=y) \times P(X=x | Y=y)$
- $P(Y=y, X=x) = P(X=x) \times P(Y=y | X=x)$
- $P(X=x, Y=y) = P(Y=y, X=x) \quad \downarrow$
- $P(Y=y | X=x) = \frac{P(Y=y) \times P(X=x | Y=y)}{P(X=x)}$

$X$  e  $Y$  sono indipendenti se

$$P(X=x, Y=y) = P(X=x) \times P(Y=y)$$

Da cui  $P(X=x | Y=y) = P(X=x)$

In genere  $X_1, \dots, X_n$  indipendenti se

$$P(X_1=x_1, \dots, X_n=x_n) = P(X_1=x_1) \times \dots \times P(X_n=x_n)$$

## Combinazione variabili

Un'operazione sui variabili aleatori si chiama  $X$  nuova variabile aleatoria e il suo  $R(X)$  che è dato dai valori ottenuti combinando in ogni modo possibile i valori delle  $n$  variabili secondo l'operazione

Variabile somma  $X = X_1 + X_2 + \dots + X_n$

## Valore atteso o media

$$\mathbb{E}[X] \text{ è } \sum_{x \in R(X)} x \cdot p(X=x)$$

$\stackrel{n \text{ usi}}{\text{anche}}$   
 $N_X$

## Varianza

$$G_X^2 = \mathbb{E}(X - \mathbb{E}[X])^2 = \sum_{x \in R(X)} (x - \mathbb{E}[X])^2 p(X=x)$$

Varianza: media dei quadrati delle differenze tra i valori che può avere  $X$  e il valore atteso

- risulta la rappresentazione della variabile aleatoria  $X$  in base al valore medio
- valore probato di  $G_X^2 \rightarrow G_X$  detta deviazione standard

Se  $X$  continua ha definizioni di media

e varianza analoghe

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot p(x) dx$$

$$G_X^2 = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p(x) dx$$

Mefo e Varianzza di variabili

$$X = X_1 + \dots + X_n$$

$$\cdot E[X] = E[X_1] + \dots + E[X_n]$$

$$\cdot \sigma^2_X = \sigma^2_{X_1} + \dots + \sigma^2_{X_n}$$

Ma cosa è combinazione lineare  $X = \alpha_1 X_1 + \dots + \alpha_n X_n$

$$\cdot E[X] = \alpha_1 E[X_1] + \dots + \alpha_n E[X_n]$$

$$\cdot \sigma^2_X = \cancel{\alpha_1^2 \sigma^2_{X_1} + \dots + \alpha_n^2 \sigma^2_{X_n}}$$

Entropia (tirofisi)

Così  $X$  è distribuzione di prob.  $P(X)$ , l'entropia è  $H(P(X))$

$$H(P(X)) = - \sum_{x \in \text{dom}(X)} P(X=x) \log(P(X=x))$$

E' un valore positivo e il log è in base 2

L'entropia indica quanto è difficile prevedere il valore della variabile avendo la distribuzione  $P(X)$

Entropia: minima (0) se tutta  $X$  è dist. di prob.  $P(X)$  passa tutte le possibili con certezza il valore delle variabili

: Massima se tutti i valori della distribuzione sono equiprobabili

Se  $p$  rappresenta questa prob (uguale per tutti)

$$H(P(X)) = \log \frac{1}{p}$$

Entropia e Varianza misurano entrambi l'incertezza  
del valore di una variabile

- Perciò:
- Entropia definita solo in base alle probabilità dei possibili valori e non in base ai valori stessi
  - Varianza invece dipende anche dai valori stessi

### Entropia relativa

Date  $P(X_0)$  e  $Q(X_1)$  due distribuzioni di prob. associate a  $X_0$  e  $X_1$ , che hanno stesso range.

L'entropia di  $P$  rispetto a  $Q$  è

$$H(P||Q) = \sum_{x \in R(X)} P(X_0=x) \log \frac{P(X_0=x)}{Q(X_1=x)}$$

L'entropia relativa minima la somma delle entropie

$$\cdot H(P||Q) \neq H(Q||P)$$

$$\cdot J(P||Q) = H(P||Q) + H(Q||P) \text{ è detta entropia relativa tra le due distribuzioni}$$

## Covarianza e matrice di covarianza

Nel caso di due var. aleatorie  $X$  e  $Y$  si definisce  
la covarianza

$$\begin{aligned} \text{Cov}_{X,Y} &= E[(X - E[X])(Y - E[Y])] = \\ &= \sum_{x \in R(X), y \in R(Y)} (x - E[X])(y - E[Y]) P(x,y) \end{aligned}$$

nel caso continuo  $\rightarrow$  integrale

Date  $n$  variabili  $X_1, X_n$  costituiscono una  
matrice simmetrica detta di covarianza.

Gli elementi sono le covarianze di tutte le possibili  
coppie

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{pmatrix}$$

E' simmetrica perché

$$\sigma_{21} = \sigma_{12} \text{ ecc ecc.}$$

Usando covarianza e deviazioni standard delle

$X$  e  $Y$  posso calcolare la correlazione di  $X$  e  $Y$  (di Pearson)

$$r_{X,Y} = \frac{\text{Cov}_{X,Y}}{\sigma_X \sigma_Y}$$

i valori vanno da -1 a 1

1) Ottengo 1 se  $X = Y$

-1) Ottengo -1 se

Se  $r_{X,Y} > 0$  sono correlate positivamente

Se  $r_{X,Y} < 0$  sono correlate negativamente

Se  $r_{X,Y} = 0$  sono indipendenti

Con cosa si correlazione?

Minimizza il grado di dipendenza lineare tra le 2 variabili

Quindi se ci è dipendenza lineare ovvero

$Y$  approssima  $aX + b$  le variabili sono correlate

$a > 0 \rightarrow P_{X,Y} > 0$  (pos. correlate)

$a < 0 \rightarrow P_{X,Y} < 0$  (negat. correlate)

Se le due var.  $X$  e  $Y$  sono indipendenti  $\rightarrow P_{X,Y} = 0$

Ma non è tutto il viceversa, (potrebbe avere una correlazione a tipo quadratico o altro tipo)

## Condizione di Spearman

• confrontiamo i rank

Sono facendo la correlazione di Pearson tra i rank delle due variabili

Minimizza la correlazione tra rank dei valori delle due variabili

Rank: posizione nell'insieme dei valori ordinato in maniera crescente

## Calcolare il rank

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8 \quad x_9 \quad x_{10}$$
$$3 \quad 4 \quad 5 \quad 6 \quad 5 \quad 1 \quad 4 \quad 3 \quad 2 \quad 5$$

• Ordiniamo

$$\frac{x_6}{1} \quad \frac{x_9}{2} \quad \frac{x_1}{3} \quad \frac{x_8}{3} \quad \frac{x_2}{5} \quad \frac{x_3}{5} \quad \frac{x_5}{2} \quad \frac{x_7}{5} \quad \frac{x_3}{5} \quad \frac{x_{10}}{5}$$

Se gli  $x$  condividono la "posizione" di  $i$  allora le loro posizioni sono uguali

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
3,5	6,5	8,5	6,5	6,5	1	6,5	3,5	2	9,5

•  $x_1$  e  $x_8$  hanno prima 3 e 5  $\rightarrow (3+5)/2 = 3,5$

Dunque  $\bar{x} \rightarrow$  rank  $R_x$  ottiene uno ciak di ogni valore il suo rank

La distribuzione di prob. di  $R_x$  è uguale a quella di  $X$  (escluso per la densità nel continuo)

Combinazione di Spearman:

$$P_{X,Y} = \frac{6 R_{X,Y}}{6 R_X R_Y}$$

Spearman misura la concordanza fra rank e non le variabili

Inoltre se esiste una relazione monotonica tra le variabili ovvero se esiste una f monotonica (non per forza lineare) che descrive  $Y$  in funzione di  $X$

Spearman è in generale più applicabile a Pearson

Se la funzione che lega  $Y$  e  $X$  non è monotonica Spearman non riesce a catturare questa relazione

2 colpire lo effett.

Nelle prob.: fino a ora Pearson è più quello di Spearman (se quello di Pearson non ci ha dato info)

## Matrice di confusione

come con le variabili date  $X_1, \dots, X_m$

Continuiamo:

$$P = \begin{pmatrix} 1 & p_{12} & \dots & p_{1m} \\ \vdots & & & \vdots \\ p_{m1} & \dots & \dots & 1 \end{pmatrix}$$

elementi: combinazioni  
tra tutte le  
possibili copie

## Distribuzione di Bernoulli

(2 esiti possibili)

$Y$  può avere solo 2 valori: 0 e 1

Sia  $P(Y=1)$  con prob  $p$ , allora  $P(Y=0) = 1-p$

La distribuzione di Bernoulli di  $Y$  è

$$P(Y=y) = p^y (1-p)^{1-y} \quad \text{per } y = 0, 1$$

$$\mathbb{E}[Y] = p \quad \text{probabilità media}$$

$$\text{Var}[Y] = p(1-p)$$

## Distribuzione binomiale

E' possibile vedere come una sequenza di esperimenti

Bernoulli indipendenti tra loro.

(il loro totale numero è indipendente dal probabile)

•  $M$ : esperimenti di Bernoulli (indipendenti)

• ogni esperimento ha prob  $p$

Allora la distribuzione binomiale per  $Y$  sarà il numero

$$P(Y=y) = \binom{M}{y} p^y (1-p)^{M-y} \quad \text{per } y = 0, 1, \dots, M$$

$$\text{NB} \quad \binom{M}{y} = \frac{M!}{y!(M-y)!}$$

combinazioni  
di  $y$

$$\begin{aligned} P(C \cap T) &= P(C) \cdot P(T) \cdot P(C) = \\ &= (1-p) \cdot p \cdot (1-p) = p^1 (1-p)^{M-1} \end{aligned}$$

## Molti e somme di variabili binarie

Si parla di somma come moltiplicazione di variabili binarie. Somma di molti e variazioni delle singole variabili è Bernoulli.

$Y_1, Y_2, Y_3, \dots, Y_n$  variabili di Bernoulli.

$$Y = Y_1 + \dots + Y_n$$

$$\mathbb{E}[Y] = \mathbb{E}[Y_1] + \dots + \mathbb{E}[Y_n] = np$$

$$\text{Var}^2 = \text{Var}^2_{Y_1} + \dots + \text{Var}^2_{Y_n} = np(1-p)$$

## Distribuzione igeometrica

Uma e  $N$  palline  $\rightarrow$   $m$  buone,  $N-m$  bende

Percorso in palline buone (caso binomiale fermo)

$Y$ : palline buone entro il ball' una

$$P(Y=y) = \frac{\binom{m}{y} \binom{N-m}{m-y}}{\binom{N}{m}}$$

con  $y = A, A+1, \dots, B$

$$A = \max(0, m + N - n)$$

$$B = \min(m, n)$$

$A$  e  $B$  determinano

i "confini" dei valori possibili di  $Y$

## Distribuzione Uniforme

$Y$  ha la distribuzione uniforme se i valori di  $Y$  sono

$$a, a+1, \dots, a+b-1 \quad \text{con } b > 1 \quad \text{e la}$$

probabilità che  $Y$  assume un dei  $b$  valori è

$$\frac{1}{b}$$

$$P(Y=y) = \frac{1}{b} \text{ per } y=a, \dots, a+b-1$$

media  $E[Y] = a + \frac{b-1}{2}$

$$\text{Var}_Y = \frac{b^2 - 1}{12}$$

### Distribuzione geometrica

E' simile alla binomiale

Considerare nrg. di Bernoulli trials indipendenti con prob di successo  $p$

$n$  NON è finito

$Y$  è il numero di successi prima di un fallimento

Se  $Y=y$ : ho ottenuto  $y$  successi già ora e  
un fallimento

La variab.  $Y$  ha una distribuzione geometrica:

$$P(Y=y) = (1-p)p^y \quad \text{con } y=0, 1, \dots$$

### Distribuzione di Poisson

Ho un esperimento di Bernoulli ripetuto infinite volte in un certo tempo  $T$  e il numero medio di successi è  $\lambda (> 0)$

$Y$  è il numero di successi in  $T$

E' simile alla binomiale ma invece di avere  $n$  ho il numero medio di successi nell' $T$

$$P(Y=y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{per } y=0, 1, \dots$$

media e varianza  $\Rightarrow \lambda$

C'è una relazione tra  $\lambda$  e la frequenza

La Poisson ci può definire un'approssimazione della binomiale se  $n \rightarrow \infty$ ,  $p \rightarrow 0$  e  $np = \lambda$  costante.

$$n \rightarrow \infty, p \rightarrow 0, np = \lambda$$

- Queste condizioni su  $n$  e  $p$  sono molto frequenti.
- Ponendo queste condizioni, la Poisson è detta della binomiale.
- La Poisson è già facile da usare dato che ha solo un parametro ( $\lambda$ ) mentre la binomiale  $2(n, p)$ .

### Distribuzione multinomiale

È una generalizzazione della binomiale.

Abbiamo  $m$  Bernoulli Trials, con  $M \geq 3$  OUT

biensi

La prob di avere  $i$  ( $i=1..m$ ) è la stessa per tutti i trials ed è  $p_i$ .

Ho  $Y_1, Y_2, \dots, Y_m$  variabili dove  $Y_i$  indica il risultato quando i quanti volte è stato osservato i negli  $m$  trials.

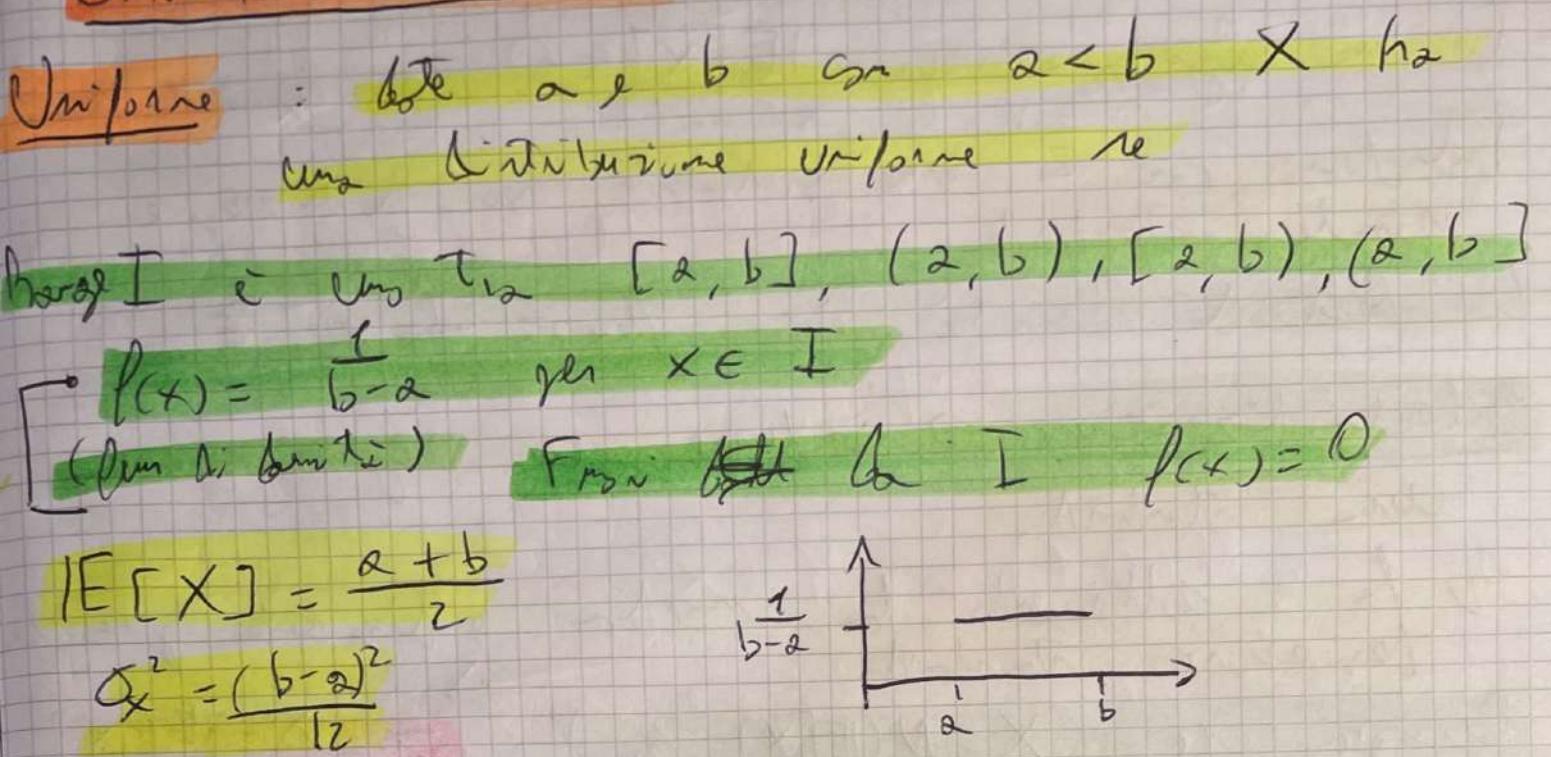
Dove  $Y_i$  è una variabile binomiale con

$$\text{media} = M p_i \quad \text{varianza} = M p_i (1 - p_i)$$

La prob che  $Y_i = y_i$  ( $i=1..M$ ) è

$$P(Y_1=y_1, \dots, Y_m=y_m) = \frac{m!}{\prod_{i=1}^m y_i!} \prod_{i=1}^m p_i^{y_i}$$

## Distribuzione continua

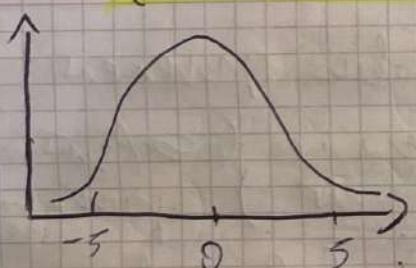


## Normale (Gaussiana)

$X$  ha una dist. Gaussiana se ha campo  $(-\infty, +\infty)$  e  $f(x)$  funzione densità è

$f(x) = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(x-\alpha)^2}{2\beta^2}}$   $\text{Var}[X] = \beta^2$  se  $\alpha = 0$

con  $\alpha$  e  $\beta$  parametri della dist.  
 $(-\infty < \alpha < +\infty, \beta > 0)$



Forma più comune  $\rightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$

Una variabile aleatoria che ha la dist. normale è indicata con  $N(E[X], \sigma_x^2)$

## Scarsità Standard

E' una Gaussian con  $\text{IE}[X] = 0$  e  $\sigma_x^2 = 1$

Indichiamo con  $Z$  una variabile con distribuzione Normale Standard.

La sua densità è  $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

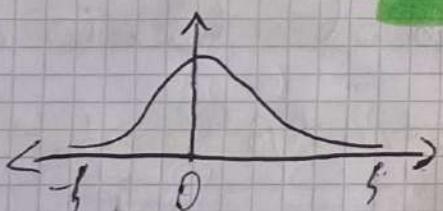
Standardizzazione  $\rightarrow$  Combinare una Gaussiana a una standard

E' fa: sottrarre la media e dividere per la deviazione standard.  $\rightarrow$

Dunque se  $X = N(\text{IE}[X], \sigma_x^2)$   $Z = \frac{X - \text{IE}[X]}{\sigma_x}$

Se  $x$  è valore orario di  $X$

allora  $z = \frac{x - \text{IE}[x]}{\sigma_x}$  è detto z-score



## Applicazioni della normale / Gaussian Standard

1)

Se vogliamo confrontare due valori provenienti da due diverse distribuzioni Gaussiane (corrispondenti a risponditori su valori nella dist.) possiamo usare gli z-score invece di calcolare le probabilità di ottenere valori tali nell'unisono.

2)

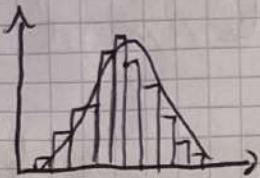
E' possibile usare per calcolare le probabilità di ottenere valori in una qualsiasi distribuzione

Trovare entro 1-2-3 deviazioni standard dalla med. (regola delle 1-2-3 dev standard)

Con questa regola si può approssimare probabilità di variabili con distribuzione simile alla normale

Esiste un rapporto tra normale e binomiale

Inoltre la binomiale si comporta come approssimazione della gauss.



### Distribuzione esponenziale

Applicazione in biologia sopravvivenza

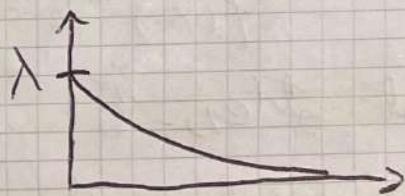
$X$  con dist. esponenziale ha spettrum di valori

$$f(x) = \lambda e^{-\lambda x} \text{ con } x > 0$$

funzione cumulativa  $\rightarrow F(x) = 1 - e^{-\lambda x}$

media (stesso)  $E[X] = \frac{1}{\lambda}$

Varianza  $\sigma^2 = \frac{1}{\lambda^2}$



### Distribuzione Chi-quadro

Usata per testare ipotesi statistiche

H<sub>0</sub>:  $X_1, X_2, \dots, X_n$  n variabili stazio. con una dist. normale standard  $N(0,1)$

Creiamo  $X = X_1^2 + X_2^2 + \dots + X_n^2$  detta  $X^2$  con n gradi di libertà

La funzione di densità è

$$f(x) = \begin{cases} 0 & \text{se } x < 0 \\ \frac{1}{\Gamma(\frac{m}{2})} \left(\frac{1}{2}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} e^{-\frac{1}{2}x} & \text{se } x \geq 0 \end{cases}$$

Dove  $\Gamma(v) = \int_0^{+\infty} e^{-t} t^{v-1} dt$  (Funzione Gamma)

valore  $\mu_{X_m} = M$

$$\text{Varianza} = 2M$$

grafico  $\rightarrow$  simmetrico attorno alla linea  
di y anche da m.

### Dati T-student

H<sub>0</sub>  $Z = N(0,1)$  e  $Y = X^2$  con m gradi di libertà

$\rightarrow$  Z valori aleatori indipendenti

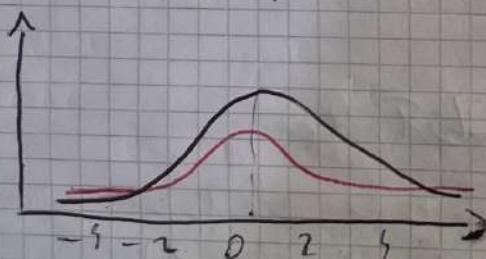
H<sub>0</sub>  $X = \frac{Z}{\sqrt{m}}$  di distribuzione secondo la T-student con m gradi di libertà

La funzione di densità è

$$f(x) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})} \frac{1}{\sqrt{m\pi}} \left(1 + \frac{x^2}{m}\right)^{-\frac{1}{2}(m+1)}$$

valore  $\mu_{X_m} = 0$  se  $m > 1$

$$\text{Varianza} = \frac{m}{m-1} \text{ se } m > 2$$



StD - Normal

grado  $m = 1$

già m è alto già i  
sopravvive alla messa

## Dati multi variabili multivariante

E' un esempio di distribuzione di probabilità congiunta continua.

Si considera come una generalizzazione della distribuzione multivariata del caso di più variabili.

Ho:

- $\vec{X} = (X_1, \dots, X_m)$  vettore di  $m$  variabili
- $\vec{N} = (N_1, N_2, \dots, N_m)$  vettore delle  $m$  variabili
- $\Sigma$  = matrice di covarianza
- $\vec{x} = (x_1, \dots, x_m)$  vettore di  $m$  valori con  $x_1 \in R(X_1), \dots, x_m \in R(X_m)$

$X_1, \dots, X_m$  hanno una

multivariata se la loro distribuzione di probabilità congiunta è

$$f(\vec{X} = \vec{x}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$$

$\nabla |\Sigma|$  determinante della matrice  $\Sigma$  di covarianza

# MARKET BASKET Model

Sezione relativa alle trig di elementi >

- item (oggetti)
- basket (o transazione)

Ogni basket è un itemset (insieme di item)

N-basket >> M oggetti -> in un basket

Comunemente il basket potrebbe essere un DB

Il market è chiamato così perché un'elaborazione

regole trig si sostituisce dai singoli item (i singoli item nel basket) al mercato (market)

Trovare itemset frequenti ovvero insiemi di item che sono osservati spesso in molti basket

Frequenze: definite in 0 fatto solo a rapporto.

Il rapporto di un itemset è il numero di itemset basket in cui è presente

Se il rapporto  $\geq 0$  è detto frequente

## Applicazioni

Cercare concetti condivisi  $\rightarrow$  item = parola basket = documenti  
Trovare coppie di gruppi di parole che appaiono insieme in molti documenti e che quindi potrebbero essere legate tra loro

Programma: documenti che hanno molte parole in comune  $\rightarrow$  parola parola

(Item = documenti  
basket = frase)

## Scelta del problema

D'obbligo la ricerca di itemset frequenti è fatta su DB esistenti.

Avere algoritmi efficienti è fondamentale, bisogna ottimizzare tempo e spazio.

Bisogna sapere bene anche la logica di supporto.

- alto  $\rightarrow$  pochi soluzioni interessanti

- basso  $\rightarrow$  Trope soluzioni

D'obbligo la soglia è l'10% del numero di basket

## Regole di associazione

La ricerca di itemset frequenti è finalizzata alla costruzione di regole if-then delle regole di associazione.

La regola è un'implicazione  $I \rightarrow j$  dove  $I$  è un itemset e  $j$  un item.

$I \rightarrow j$  indica che nei canali dove si osserva l'itemset  $I$  è probabile osservare  $j$ .

## Confidenza di una regola

La probabilità dell'implicazione è quantificata dalla confidenza della regola.

$Supp(X) = \text{support}(X)$

$$\text{Conf}(I \rightarrow j) = \frac{Supp(I \cup \{j\})}{Supp(I)}$$

La confidenza rappresenta la percentuale dei canali in cui si osserva  $I$  e  $j$  insieme.

$Supp(I \cup \{j\})$  supporto della regola.

$Supp(I)$  è detta correlazione della regola, cioè quanto meno si può applicare la regola.

La confidenza è utile se il rapporto dell'influenza a  $I \times (I \rightarrow j)$  è abbastanza alto.

Infatti se abbiamo  $\text{Conf} = 100\%$  in un  $I \times j$ , la confidenza è poco interessante. Ma in realtà anche rapporto e Conf. altri non danno informazioni molto interessanti.

Ese:  $\text{Pasta} \rightarrow \text{Salz}$  E' quasi ovvio, non ha molte info. (anche rapporto e Conf sono alti)

Invece  $\text{Pomodoro} \rightarrow \text{Bitter}$  potrebbe essere interessante con rapporto e Conf alti.

### Interesse di una regola

$\text{Int}(I \rightarrow j)$  definisce l'influenza di un insieme di oggetti su un oggetto.

$$\text{Int}(I \rightarrow j) = \text{Conf}(I \rightarrow j) - \frac{\text{Supp}(j)}{N}$$

$N$ : numero basket

I non ha influenza su  $j \rightarrow$  per centuale di basket va con  $I$  che  $j$  è circa uguale a per centuale di basket con  $j$ .

L'interesse può essere positivo o negativo. Ci indica quanto i loro i basket ~~sono~~ un'associazione.

Fatto posso  $\rightarrow$  pasta è banale,  $\text{Int} = 0$

Pomodoro  $\rightarrow$  pasta non è banale

$\text{Lift} < 1$ : più giù basso è lift e maggiore è influenza media di  $I$  su  $j$ ,

$\text{Lift} > 1$ : più alta è lift e maggiore è influenza media di  $I$  su  $j$

Meno attenzivo all'influenza di  $I$  su  $j$

$$\text{lift}(I \rightarrow j) = N \times \frac{\text{Sup}(I \cup j)}{\text{Sup}(I) \times \text{Sup}(j)}$$

$N$ : non bastet

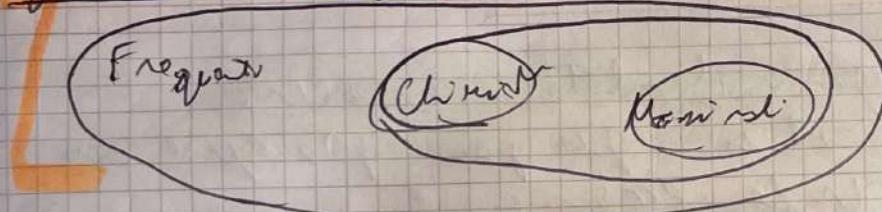
lift corrisponde al rapporto tra il supporto composto dalla regola e il supporto stesso. supponendo non ci sia alcuna dipendenza tra  $I$  e  $j$

$\text{lift} < 1$ :  $I$  influenza negativamente  $j$   
 $\text{lift} > 1$ : positiva

Inicien Frequenti Monostri e chiavi

- Insieme Frequenti e chiavi se non c'è un ruolo nega - insieme con lo stesso supporto
- Insieme Freq. è una singola se non esiste un ruolo nega insieme Frequenti

Gli inizi Frequenti trovati sono anche Chiavi



Un algoritmo che riesce a trovare insieme Frequenti e Chiavi a trovare anche quelli chiavi e monostri sono

Algoritmi

Mitolo Nine: genera tutti gli itemset e calcola per ciascuno supporto e vedi se supera la soglia

Se  $m$  numero item, numero copie è  $\binom{m}{2}$ , numero copie è  $\binom{m}{3}$ .

Mitolo Nine non è particolarmente efficiente

Dato  $I$ , per ogni  $S \subseteq I$ ,  $\text{Sup}(I) \leq \text{Sup}(S)$

Dunque il supporto di un  $I$  non supera mai quello dei suoi sottoinsiemi (il supporto di un  $I$  è sempre  $\geq$  al supporto di un suo sottinsieme).

E' detta anti-monotonicità del supporto.

- Se un itemset  $I$  è frequentante allora ogni sottoinsieme di  $I$  è frequentante oppure
  - Se un itemset  $I$  non è frequentante, allora ~~sussistono~~ nessun itemset che contiene  $I$  è frequentante
- Dato insieme A-Priori (non vale il viceversa)

### L'algoritmo A-Priori → bottom-up

- Setta set tascabili per cardinalità crescente  
portando da singoli item per poi frequentare con le coppie poi con le triple ecc...
- Se itemset con cardinalità  $K$  non è frequentante allora non alcuna estensione è contenuta in itemset con  $K+1$  item (perché quest'ultimo sicuramente non sarà frequentante)

Ciò ci permette di evitare l'esplorazione di molti candidati.

Teniamo conto anche del fatto che itemset frequenti con cardinalità elevata ( $\geq 3$ ) sono rari.

## Algoritmo

- 1) Creare  $L_1$  degli ~~item~~ frequenti
  - 2) Per valori  $k$  te creare (partendo da  $K=1$ )
    - a) Creare  $C_{k+1}$  degli item contatti b' cardinali  $k+1$  partendo da  $L_k$
    - b) Rimuovi da  $C_{k+1}$  gli item che hanno almeno un item con  $k$  item che non sono frequenti
    - c) Collega nippo a ogni item in  $C_{k+1}$
    - d) Continua  $L_{k+1}$  formato dagli item di  $C_{k+1}$  frequenti
- Algoritmo termina quando i contatti finiscono

## Generazione card dati

Creare i tabella con  $k+1$  colonne ( $+1$  è il rapporto)

L'insieme  $C_{k+1}$  si ottiene mediante l'applicazione del join b'  $L_k$

Per esprimere due righe A e B di  $L_k$  ignoriamo

- 1) I primi  $k-1$  item di A e B sono uguali
- 2)  $k$ -esimo item di A  $\leftarrow$  Nessun item di B

Le 2 righe non devono contattarsi con i contatti in  $L_k$  per  $k=2$

itemset	rapporto
b, c	1
b, j	2
b, m	1
c, j	3
c, m	2
j, m	2
m, p	2

itemset	rapporto
//	//
//	//
/	/
/	/

$C_{k+1}$

b, c, j  
b, c, m  
b, j, m  
c, j, m

Eseguo confronto su slide (risposte frequenti: 3E-55)

### Utilizzo delle memoie

Dobbiamo fare i conti con la gestione delle memoie  
Programma attivo: calcoli del rapporto  
Dati in basket e item in basket  
Obiettivo: avere solo in RAM le tabelline con i  
supporti degli item (singoli)

Con una Hash map associamo ad ogni item un intero  
identificativo

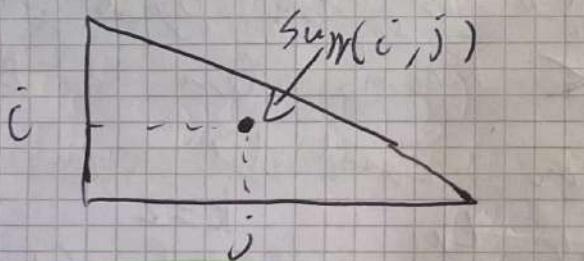
L'hash è più rappresentazione come l'indirizzo dell'item nel supporto

Per calcolare il rapporto delle copie ci sono due tecniche

- metodi matrice triangolare
- metodi delle triple

### Matrice triangolare

Per ogni coppia di item con  $i < j$  con  $i < j$   
memorizziamo con il rapporto della copia  $(i, j)$



Nella pratica viene usata un array (invece di una  
matrice  $M \times M$  dove metti e imutilizzate)

$\text{Sup}(i, j)$  sta nella posizione  $k$ .

$$k = (i-1)M + (k - (i-1)(M - \frac{1}{2}) + j - 1$$

Possiamo avere le coppie  $(1,2)$   $(1,3) \dots (1,n)$  e poi  
 $(2,3)$   $(2,4) \dots (2,n)$  e così via

### Metodo delle triple

Il reporto  $\rightarrow$  di  $(i,j)$  è memorizzato come triple  
 $(i,j,m)$

Le triple sono memorizzate in un basket così  
che la coppia  $(i,j)$  è la chiave e  $m$  il  
corrispondente valore.

La scelta del metodo dipende dal numero di copie  
che appaiono in qualche basket

- Se almeno  $\frac{1}{3}$  delle  $\binom{M}{2}$  copie di interi sono presenti  
nel basket scriverà la matrice triangolare nel  
caso il metodo delle triple

Per iniziare con 3 o più elementi  $n$ . Usare le triple  
e basket (è più conveniente)

### Triple vs matrice Triangolare

Triple: memorizza solo copie che appaiono in qualche  
basket

Matrice triangolare: richiede un solo basket da memorizzare  
ognes il reporto.

Il metodo delle triple memorizza:  $i, j$  e  $m$

## Algorithmi di Park CHEN e YN PCY

• Prima fase: contano supporto item e leggono copie parziali  
item in ogni basket

Le copie delle aree vengonoate in bucket  $U_{ijkl}$   
funzione hash

Se supporto bucket  $< n_{min} \rightarrow$  rimuovi copia di item  
più frequente (bitwise)

Per determinare se un bucket è frequente usi un bit  $(1 \ 0 \ 0) \rightarrow$  bitmap

Il primo delle copie  $c_{ijkl}$  sarà formato da copie  $i,j$

•  $i, j$  item frequenti

• coppia  $(i, j)$  cade in un bucket frequente

$H(i, j) = x \xrightarrow{\quad} B_1 (i^1, j^1, 1)$

$B_2$

$\xrightarrow{\quad} B_3 (i^2, j^2, 1) (i^{11}, j^{11}, 1) \ S=3$

Quindi se  $B$

questo algoritmo: generare meno copie codificate

## Algorithmi Multistage

Sulle copie  $(i, j)$  che cadono in un bucket frequente effettua un ulteriore raggruppamento in bucket con una 2a fnc hash

E si può applicare il process più volte

## Algorithmi multihash

Si usano 2 o più fnc hash in parallelo nello stesso step.

Le copie codificate sono quelle che cadono in bucket frequenti sulla base di ogni funzione

(le copie non cadono in Bitmap 1 e 2 in bucket frequenti)

## Algoritmi Randomizati

Applicare l'Algoritmo sugli itemset in modo tale che la ricerca di itemset frequenti sia applicata su un sottogruppo random di basket invece di tutto il dataset.

## Algoritmo SON

Divide il dataset in chunk

$S = \text{numero minimo}$   $p = \text{percentuale di basket in ogni chunk}$

- 1) Su ogni chunk applica Algor. (o una sua ottimizzazione). Il numero minimo è ridotto a  $p \times S$
- 2) Considera l'unione di tutti gli ~~elementi~~ gli itemset che sono risultati frequenti in uno o più chunk
- 3) Per ogni itemset combattuto calcola il rapporto nel dataset iniziale

SON è molto buono in architettura distribuita

## TOIVONEN

frontiera negativa: inviare gli itemset che non sono frequenti nel sottogruppo  $S$  del dataset  $D$ )

ma i cui sottoinsiemi immediati (ottenuti togliendo un solo item) sono frequenti in  $S$

Ese.  $\{A, B, C\}$  non è in  $S$  ma  $\{A, B\}$ ,  $\{A, C\}$  e  $\{B, C\}$  lo sono perché  $\{A, B, C\}$  è parte della frontiera negativa di  $S$ .

## Poss Toivonen

- 1) Considera  $S$  e  $D$  forniti da  $p$  bank.  
Se rapporto minore di  $D$   
la lista  $S$  sarà  $\delta \times p \times S$  dove  $\delta$  è  
tra 0 e 1
- 2) Calcola i tempi frequenti in  $S$  e quelli in frontiera  
negativa
- 3) Calcola rapporto degli itemset trovati in 2)
  - a) Se nessun itemset della frontiera negativa è  
frequente in  $D$  bisceguo tutti gli item che  
sono risultati frequenti in  $S$
  - b) Se c'è o gli itemset della frontiera negativa  
sono frequenti in  $D$  allora si ripete l'algoritmo  
in un modo congiunto  $S_2$

Perché ripetere Toivonen? (nel punto 3b)

I punti presentati  
sono corretti; perche potrebbe esserci super-inver-  
sione ( $I$  ( $I$  è nella frontiera negativa di  $S$ ) che  
non sono frequenti in  $S$  ma in  $D$ ).

Dunque la rigetta perché dall'algoritmo sarebbe  
stata riconosciuta con un altro congiunto

5: determina il rapporto minimo richiesto nel  $S$ .

Determina anche la percentuale di successo

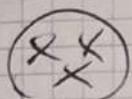
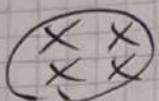
6 buono  $\rightarrow$  memoria usata maggiormente ma l'algoritmo  
finisce in meno passi.

6 cattivo  $\rightarrow$  memoria più grande ma usata meno

## Clustering

Processo che dato un insieme di oggetti li riunisce in gruppi detti cluster segnando una posizione di "distanza" tra gli oggetti.

Distanza corta = oggetti poco simili



$\times \times \leftarrow$  Oggetti con caratteristiche simili

Il clustering è applicato in diversi campi

Il clustering è un processo di unsupervised learning infatti non serve per dividere un insieme di dati in classi senza alcuna conoscenza di quale è quella classe le classi (e etichette)

Invece la classificazione è un processo di supervised learning

Si parte da dei dati etichettati, come la classe di appartenenza della training set

Si ottiene un classificatore partendo dal training set e imparando le regole di associazione tra gli attributi dei dati e la loro corrispondente classe (esempio: poiché nella base di addestramento l'etichetta è già data) \*

Ottimizzato il classificatore può poter stabilire la classe di un nuovo dato senza classe vera e propria (ma esplicitamente) \*\*\*

\* Si mette

• In pratica diamo un oggetto con la sua etichetta e il classificatore rialza le caratteristiche dell'oggetto in modo da capire già o meno le sue caratteristiche che lo sono di un oggetto con cui era confrontato