Concept Description

Gianluca Amato

Corso di Laurea in Economia Informatica
Università “G. D'Annunzio”di Chieti-Pescara

Metodi predittivi e descrittivi

- Ricordiamo che i metodi di data mining si dividono in predittivi e descrittivi
  - Descrittivi: descrivono gli insiemi di dati in oggetto in maniera concisa e semplificata, presentandone interessanti proprietà generali;
  - Predittivi: costruiscono modelli dei dati il cui scopo è predirre il comportamento di nuovi insiemi di dati.
- Alcuni algortimi di data mining possono essere considerati sia di natura predittiva che descrittiva:
  - Ad esempio un albero di decisione può essere usato per predirre il comportamento di nuove istanze, ma da anche una descrizione dell'insieme dei dati di addestramento.

Descrizione di Concetti

- Il tipo più semplice di data mining descrittivo prende il nome di Concept Description (Descrizione di concetti).
- In questo contesto, un concetto è un insieme di istanze.
  - Ad esempio, l'elenco degli acquirenti abituali di una azienda.
- Si distingue in Concept Characterization e Concept Discrimination.
  - Concept Characterization: dato un concetto, produce una descrizione dello stesso insieme di dati a un livello di astrazione più alto
  - Concept Discrimination: dati più concetti, fornisce una descrizione che ne mette in evidenza le differenze e somiglianze.

Caratterizzazione di Concetti

Generalizzazione dei Dati

- Uno dei metodi standard per descrivere un concetto è la generalizzazione dei dati:
  - eliminare alcuni attributi irrilevanti (ad esempio il nome, per il concetto degli "acquirenti abituali")
  - cambiare il livello di granularità di altri attributi (ad esempio la residenza viene astratta con la sola provincia)
  - stesse operazioni che abbiamo già visto nei sistemi OLAP.
- Due approcci principali:
  - Approccio basato sui Data Cube (OLAP)
  - Approccio basato sul processo di "Attribute Oriented Induction" (AOI)

OLAP e AOI

- Attribute Oriented Induction
  - Può operare su attributi e misure di tipo complesso.
  - Tipicamente basato su un database relazione
  - Eseguito su richiesta dell'utente (nessuna precomputazione)
  - un processo automatizzato: il sistema determina il miglior livello di granularità per descrivere il concetto.
- OLAP
  - Tipicamente operara su attributi e misure di tipo base.
  - Tipicamente basato su un data-warehouse
  - Pre-computazione dei cuboidi per migliorare le prestazioni.
  - E` un processo manuale: l'utente decide quali attributi considerare e a quale livello di granularità

Attribute Oriented Induction

- Proposto nel 1989, prima dell'introduzione del modello OLAP.
- Come funziona?
  - Parte dalla raccolta dei dati rilevanti (il concetto da descrivere) che è la relazione iniziale.
  - Esegue dei passi di generalizzazione o rimuovendo attributi o generalizzando attributi.
  - Aggrega i dati, mettendo assieme le tuple identiche e accumulando i relativi conteggi. Si ottiene la relazione prima generalizzata.
  - Presenta i dati all'utente, sotto varie forme, possibilmente in maniera interattiva consentendo operarazioni di drill-down, roll-up, etc..

Passi di generalizzazione

- Rimozione di attributi: rimuovi un attributo A se A assume un numero elevato di valori diversi e
  - Non esiste un operatore che generalizzi A, oppure
  - Le versioni generalizzate di A sono espresse in termini di altri attributi presenti nell'insieme di dati.
- Generalizzazione di attributi: se A assume un numero elevato di valori diversi ed esiste una possibile generalizzazione di A, sostituisci A con la sua versione generalizzata.
- Ovviamente il sistema deve conoscere in anticipo le possibili generalizzazioni di un attributo, ovvero la gerarchia dei concetti.

Controllo del processo di generalizzazione

- Soglia sugli attributi: per ogni attributo si setta una soglia. Se il numero di valori distinti per un attributo è superiore alla soglia, l'attributo va generalizzato o eliminato.
- Soglia sulla relazione generalizzata: si setta una soglia per la dimensione della relazione generalizzata. Se essa ha dimensione maggiore della soglia, bisogna continuare col processo di generalizzazione, altrimenti ci si ferma.
- Le due tecniche si possono applicare assieme.

Aggregare i risultati (1)

- Gli utenti sono spesso interessati ad avere delle informazioni statistiche circa i dati a differenti livelli di astrazioni
  - È importante aggregare i valori durante il processo di induzione.
- Nella forma più semplice, per ogni tupla generalizzata si tiene il conteggio di tutte le corrispondenti nella relazione iniziale.

Aggregare i risultati (2)

- È possibile anche istruire il sistema a considerare degli attributi come delle “misure”, e quindi a fondere i loro valori nelle tuple secondo determinate funzioni di aggregazione.
  - Ad esempio, si può dire che l'attributo “unità vendute” va aggregato con una funzione somma. Ogni tupla generalizzata come “unità vendute” la somma delle unità vendute nelle singole tuple della relazione iniziale.

Esempio

Relazione Iniziale

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC,Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng,Bus	Country	Age range	City	Removed	Excl, VG,..

Relazione prima generalizzata

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Presentazione dei risultati (1)

- Relazione Generalizzata: la relazione generalizzata così come è stata calcolata dal sistema, con i valori per i conteggi e le altre aggregazioni.

location	item	sales (in million dollars)	count (in thousands)
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

Table 5.3: A generalized relation for the sales in 1997.

Presentazione dei risultati (2)

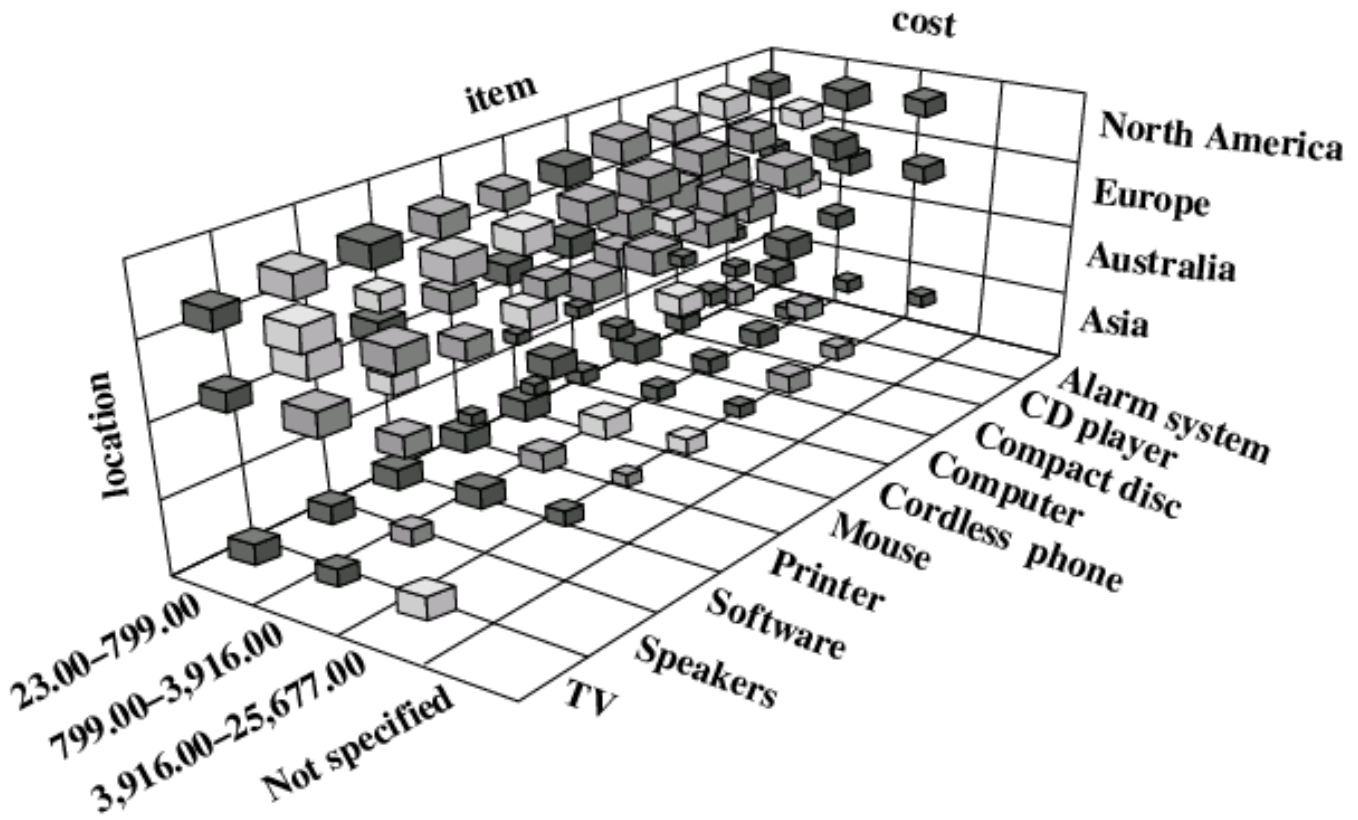
- Cross-tabs: rappresentazione bidimensionale della relazione generalizzata. Righe e colonne contengono i possibili valori per gli attributi e le misure stanno dentro gli incroci.

location \ item	TV		computer		*both_items*	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

Presentazione dei risultati (3)

- Visualizzazione grafica: diagrammi a barre, torte, cubi 3D.



t-weight

- Il valore di conteggio di una tupla generalizzata viene spesso espresso come percentuale rispetto al numero di tuple totali
  - si ottiene il t-weight (t sta per typicality)
- Formalmente
  - q_a è una tupla generalizzata e q_1, ... q_n l'insieme di tutte le tuple generalizzate
  - count(q_a) = numero di tuple nella relazione iniziale che corrispondono alla tupla generalizzata.

$$t-weight(q_a)=\frac{count(q_a)}{\sum_{i=1}^{n} count(q_i)}$$

Regole caratteristiche quantitative

- Regole caratteristiche quantitative: regole logiche con associate delle informazioni quantitative. Ad esempio
  - $\forall X.\ \text{vendita}(X) \Rightarrow$
    $(\text{location}(X)=\text{'Asia'} \wedge \text{item}(X)=\text{'Computer'})\ [t: 20\%] \vee$
    $(\text{location}(X)=\text{'Asia'} \wedge \text{item}(X)=\text{'TV'})\ [t: 6\%] \vee$
    $(\text{location}(X)=\text{'Europe'} \wedge \text{item}(X)=\text{'Computer'})\ )\ [t: 24\%] \vee$
    .....
- La testa della regola è la classe considerata, la coda è una disgiunzione di condizioni, ognuna delle quali corrisponde a una tupla generalizzata
  - Il valore in percentuale è il t-weight
  - È possibile eliminare dalla regola quei disgiunti con t-weight troppo basso.

Caratterizzazione analitica

Analisi di rilevanza degli attributi (1)

- Un problema per la Concept Description:
  - Che attributi inserire nella relazione iniziale?
- Se la scelta la fa l'utente, si va incontro a due rischi:
  - Vengono lasciati fuori degli attributi che invece sono importanti per descrivere il concetto;
  - Vengono inclusi troppo attributi, la maggior parte irrilevanti. Il metodo di Attribute Oriented Induction viene rallentato e potrebbe non accorgersi che alcuni attributi sono inutili.
- Occorre eseguire delle analisi di rilevanza degli attributi per eliminare gli attributi inutili.

Analisi di rilevanza degli attributi (2)

- Intuitivamente, un attributo è considerato rilevante per un concetto se può essere usato per distinguere quel concetto (target class) da altri (contrasting classes).
  - Ad esempio, il colore di una automobile non può essere usato per distinguere le macchine economiche, ma la cilindratA sì.
- All'interno della stessa dimensione, livelli di granularità differenti possono avere rilevanza diversa.
  - Il mese di nascista difficilmente è rilevante per descrivere il concetto “impiegati con elevato salario”... ma il decennio di nascista sì.
- L'analisi di rilevanza va eseguita a livelli multipli di astrazione.

Caratterizzazione analitica

- Si parla di analytical characterization (caratterizzazione analitica) per un processo di concept characterization che integra un processo di analisi della rilevanza degli attributi.
- Analogamente si parla di analytical comparison.
- Cosa sono la target class e le constrasting classes?
  - Per il confronto analitico, corrispondono ai vari concetti che si vogliono confrontare;
  - Per la caratterizzazione analitica, la classe target è il concetto da analizzare, ma la contrasting class non è ovvia
    - può essere ottenuta da un insiemi di dati simili presi dal database ma che esclude quelli che fanno parte della prima classe;
    - ad esempio, per caratterizzare gli studenti della laurea specialistica, la classe contrastante può essere quella degli studenti della triennale.

Misure di rilevanza

- L'idea fondamentale della analisi di rilevanza è quella di determinare delle misure di rilevanza:
  - Information gain
  - Gain ratio
  - indici di Gini
  - coefficienti di correlazione
- Riassumiamo la definizione di Information Gain.

Information Gain

- Siano date le classi C_i con i=1...m. La classe C_i contiene s_i istanze. Sia A un attributo con valori $\{a_1,...,a_v\}$ e s_{ij} sia il numero di istanze in C_i con $A=a_j$.

$$I(s_1,\ldots,s_m)=\sum_{i=1}^{m}\frac{s_i}{s}\log\left(\frac{s_i}{s}\right)$$

Entropia Iniziale

$$I(s_1,\ldots,s_m|A)=\frac{\sum_{j=1}^{v} s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

Entropia condizionata media per A

$$IG(A)=I(s_1,\ldots,s_m)-I(s_1,\ldots,s_m|A)$$

Guadagno di Informazione per A

Caratterizzazione Analitica

- Raccolta dati
- Analisi preliminare dei dati usando AOI: si eliminano gli attributi che hanno un numero di valori distinti elevato.
  - Si usano soglie degli attributi molto elevate.
  - Si ottiene la relazione candidata.
- Rimozione di attributi irrilevanti o debolmente rilevanti usando la misura di rilevanza
  - Si ottiene la relazione iniziale di lavoro.
- Si applica di nuovo il metodo di AOI, con delle soglie più stringenti, per ottenere la caratterizzazione del concetto.

Esempio (1)

- Come per l'esempio precedente, vogliamo ottenere una descrizione dei “graduate student” (studenti di dottorato) ma usando la caratterizzazione analitica.
- Siano dati
  - Attributi: *name, gender, major, birth_place, birth_date, phone#, and gpa*
  - Valori di soglia per gli attributi da usare nella fase AOI iniziale.
  - Valori di soglia per attributi e/o relazione da usare nella fase AOI finale.
  - R=0.1: valore di soglia per la rilevanza degli attributi.

Esempio (2)

- Raccolta dati
  - Target class: studenti di dottorato
  - Contrasting class: altri studenti
- Analisi preliminare con AOI
  - Attributi rimossi: *name, phone#*
  - Attributi generalizzati: *major*, *birth_place*, *birth_date, gpa*
  - Misure aggregate: *count*
- Relazione candidata
  - Attributi: *gender, major, birth_country, age_range, gpa*

Esempio (3)

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

Relazione candidata per Target class: Studenti di dottorto (Σ=120)

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

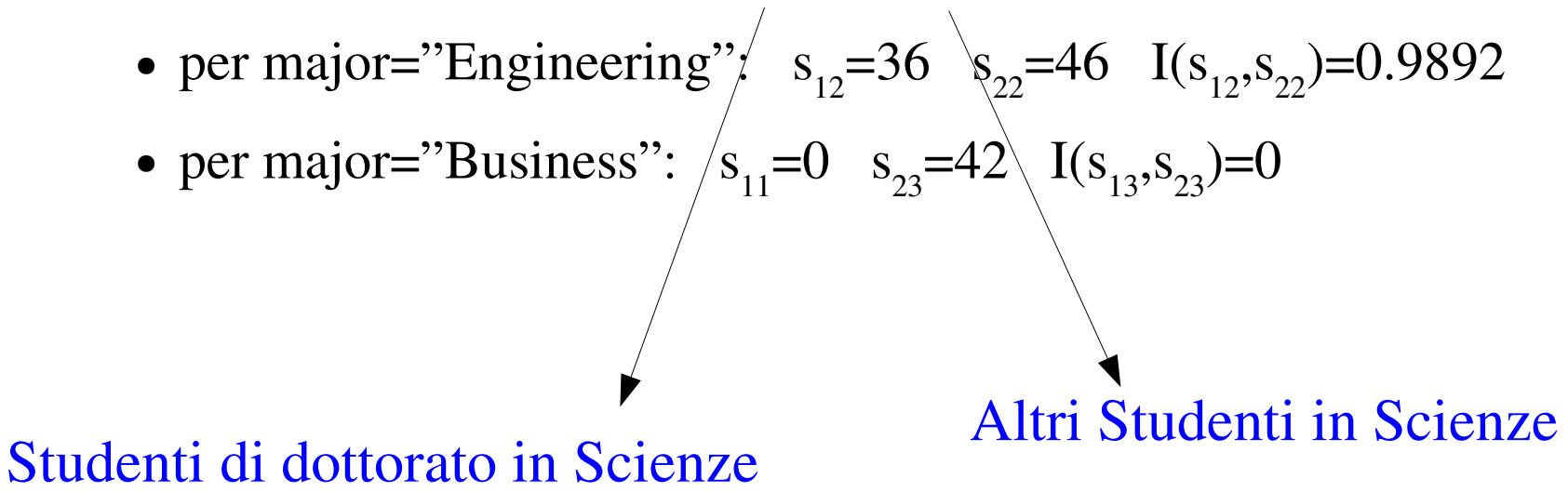
Relazione candidata per Contrasting class: Altri studenti (Σ=130)

Esempio (4)

- Analisi di rilevanza degli attributi
  - Calcolo della entropia delle due class

$$I(s_1, s_2) = \frac{-120}{250}\log_2\frac{120}{250} - \frac{130}{250}\log_2\frac{130}{250} = 0.9988$$

  - Calcolo delle entropie condizionate all'attributo major
    - per major="Science": $s_{11}=84$ $s_{21}=42$ $I(s_{11},s_{21})=0.9183$
    - per major="Engineering": $s_{12}=36$ $s_{22}=46$ $I(s_{12},s_{22})=0.9892$
    - per major="Business": $s_{11}=0$ $s_{23}=42$ $I(s_{13},s_{23})=0$



Esempio (5)

- Calcolo entropia condizionata media

$$I(s_1, s_2 | major) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

- Calcolo guadagno di informazione per attributo major

$$IG(major) = I(s_1, s_2) - I(s_1, s_2 | major) = 0.2115$$

- Calcolo guadagno di informazione per altri attributi
  - IG(gender)=0.0003
  - IG(birth_country)=0.0407
  - IG(gpa)=0.4490
  - IG(age_range)=0.5971
- Rimozione attributi poco rilevanti
  - Con soglia = 0.1 si rimuove *gender* e *birth_country*.

Esempio (6)

- Relazione di lavoro iniziale

major	age_range	gpa	count
Science	20-25	Very_good	16
Science	25-30	Excellent	47
Science	20-25	Excellent	21
Engineering	20-25	Excellent	18
Engineering	25-30	Excellent	18

- Caratterizzazione partendo dalla relazione di lavoro.
  - Si usano i valori di soglia più stringenti.

Discriminazione di Concetti

Concept Discrimination

- Il processo di Concept Discrimination o Concept Comparison produce una descrizione che confronta una classe target (bersaglio) con una o più classi di contrasto.
- Il processo di prima si può modificare per adattarlo alla discriminazione di concetti:
  - Si individuano le classi target e di contrasto.
  - Si applica il processo di AOI (eventualmente nella versione analitica) generalizzando tutte le classi allo stesso livello
  - Si presentano i dati sotto forma di tuple generalizzate o di regole logiche:
    - Misure di interesse: t-weight e d-weight

Esempio (1)

- Vogliamo confrontare gli studenti di dottorato e gli studenti dei corsi di laurea.
- Siano dati
  - Attributi: *name, gender, major, birth_place, birth_date, phone#, and gpa*
  - Valori di soglia per gli attributi da usare nella fase AOI iniziale.
  - Valori di soglia per attributi e/o relazione da usare nella fase AOI finale.
  - R=0.1: valore di soglia per la rilevanza degli attributi.

Esempio (2)

- Applicando il processo di AOI con analisi di rilevanza, otteniamo:

Birth_country	Age_range	Gpa	Count%
Canada	20-25	Good	5.53%
Canada	25-30	Good	2.32%
Canada	Over_30	Very_good	5.86%
…	…	…	…
Other	Over_30	Excellent	4.68%

t-weight

Relazione prima generalizzata per la classe bersagli**o:** Graduate students

Birth_country	Age_range	Gpa	Count%
Canada	15-20	Fair	5.53%
Canada	15-20	Good	4.53%
…	…	…	…
Canada	25-30	Good	5.02%
…	…	…	…
Other	Over_30	Excellent	0.68%

Relazione prima generalizzata per la classe di contrasto: Undergraduate students

Presentazione dei risultati

- I dati si possono presentare nelle stesse forme viste per la caratterizzazione di concetti
  - Servono misure per confrontare la classe bersaglio e quelle di contrasto.
- Abbiamo già visto come si misura il t-weight
- Adesso analizziamo una misura che discrimina tra varie classi: il d-weight.

d-weight

- d-weight (d sta per discriminante): è una misura che si calcola per ogni tupla generalizzata
  - Siano date le classi C_1,...C_n. Supponiamo C_1 sia la classe target.
  - q_a è una tupla generalizzata
  - count($q_a \in C_i$) = numero di tuple nella relazione iniziale di C_i che corrispondono alla tupla generalizzata.

$$d-weight(q_a) = \frac{count(q_a \in C_1)}{\sum_{i=1}^{n} count(q_a \in C_i)}$$

Regole quantitative discriminanti

- Il d-weight può essere utilizzato per definire delle "regole quantitative discriminanti":

$$\forall X, target\_class(X) \Leftarrow condition(X)[d{:}d\_weight]$$

dove *condition(X)* è una tupla generalizzata q_a espressa in termini logici e d è il corrispondente d-weight.

- Esempio:

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

$$\forall X, graduate\_student(X) \Leftarrow birth\_country(X) = Canada \wedge$$
$$\wedge age\_range(X) = 25-30 \wedge gpa(X) = good[d:30\%]$$

dove d=90/(90+120)=30%.

Vari tipi di regole quantitative

- Regole quantitative caratteristiche:

$$\forall X, target\_class(X) \Rightarrow condition(X)[t{:}t\_weight]$$

- Regole quantitative discriminanti:

$$\forall X, target\_class(X) \Leftarrow condition(X)[d{:}d\_weight]$$

- Regole descrittive quantitative:
  - Mettono assieme regole caratteristiche e discriminanti in una unica regola:

$$\forall X, target\_class(X) \Leftrightarrow \begin{array}{l} condition_1(X)[t:w_1, d:w'_1] \vee \ldots \vee \\ condition_n(X)[t:w_n, d:w'_n] \end{array}$$

Esempio: regole di descrizione quantitative

Location/item		TV			Computer			Both_items	
	Count	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*
Europe	80	25%	40%	240	75%	30%	320	100%	32%
N_Am	120	17.65%	60%	560	82.35%	70%	680	100%	68%
Both_ regions	200	20%	100%	800	80%	100%	1000	100%	100%

- Target class: vendite in Europa
- Contrasting class: vendita in Nord America
- Regola di descrizione quantitativa:

$$\forall X, Europe(X) \Leftrightarrow$$
$$(item(X) = TV)[t{:}25\%, d{:}40\%] \vee (item(X) = computer)[t{:}75\%, d{:}30\%]$$

Incrementalità

- Dato l'enorme ammontare di dati in un database sono preferibili algoritmi incrementali
  - dati nuovi dati, è possibile aggiornare il risultato dell'analisi senza ricalcolare tutto da capo
- L'algoritmo di AOI si può facilmente adattare per essere incrementale
  - dato un nuovo insieme di dati ΔDB, lo si generalizza allo stesso livello di astrazione della relazione prima R per ottenere ΔR
  - si uniscono R e ΔR modificando tutte le misure necessarie
- Stesso approccio si può adottare per utilizzare tecniche di campionamento, o per parallelizzare l'algoritmo.

Misure statistiche descrittive

Misure statistiche

- Un modo del tutto diverso per dare una descrizione dei dati è fornire delle misure statistiche
  - misure di tendenza centrale
    - media, mediana, moda, ...
  - misure di dispersione
    - varianza, percentili, ...
  - grafici della distribuzione dei dati
    - istogrammi, boxplot, ...

Misure di centralità

- Media (misura algebrica su attributi numerici)
  - dati i valori x_1, ... x_n la media è $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Mediana (misura olistica su attributi ordinati)
  - una volta ordinati i dati, la mediana è:
    - il valore medio se il numero di dati è dispari
    - la media dei due valori centrali se il numero di dati è pari
  - meno sensibile agli outliers
- Moda (misura olistica su qualunque attributo)
  - il valore (o i valori) che occorrono più frequentemente
- Midrange (misura algebrica su attributi numerici)
  - la media tra il valore massimo e minimo

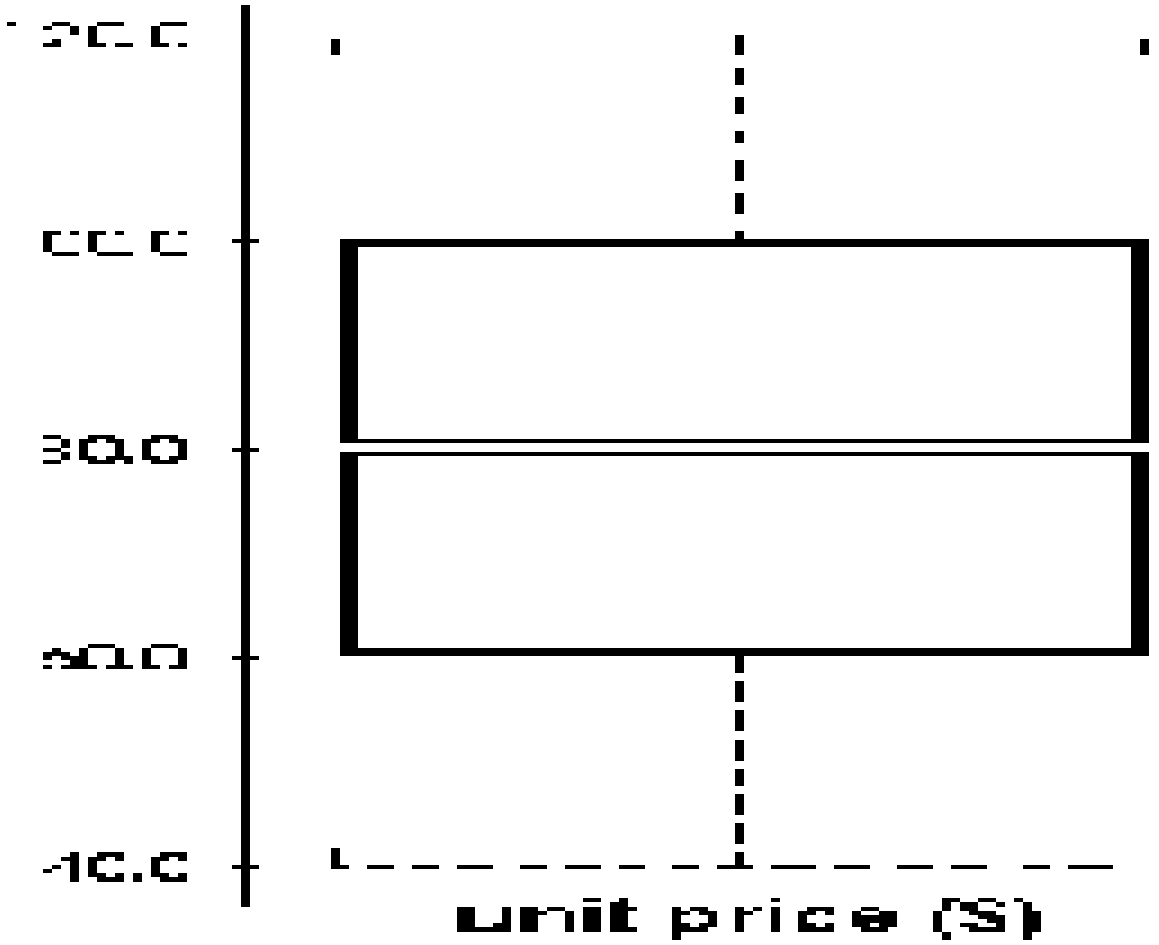
Percentili e quartili

- k-esimo percentile (misura olistica per attributi ordinati)
  - il valore x più grande tale che il k percento dei dati assume valori minori o uguali ad x
  - il 25° e 75° percentile sono detti primo e terzo quartile e indicati con Q_1 e Q_3
  - il 50° percentile corrisponde (più o meno) alla mediana
- intervallo interquartile (misura olistica su attributi ordinati)
  - (IQR) è la differenza tra terzo e primo quartile
  - una regola comune per identificare gli outliers è individuare quei valori x tali che
    - $x - Q3 > 1.5 * IQR$, oppure $Q1 - x > 1.5 * IQR$

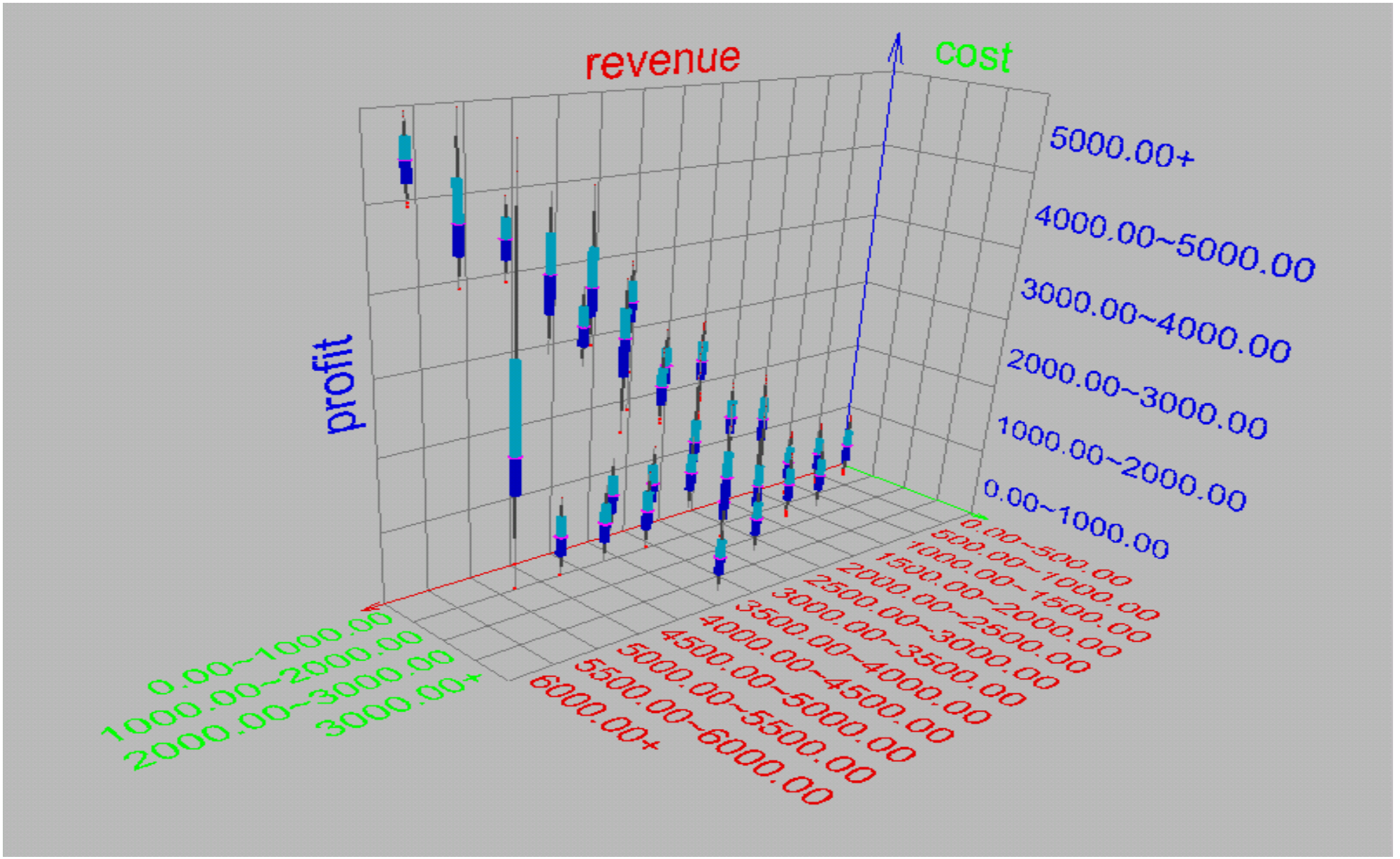
Boxplot (1)

- Spesso si riassume una distribuzione di dati indicandone il cosiddetto five-number summary: min, Q1, mediana, Q3, max.
- Un modo per rappresentare queste informazioni è il boxplot
  - i dati sono rappresentati con un rettangolo
  - gli estremi del rettangolo sono il primo e terzo quartile
  - la mediana è marcata con una riga dentro il rettangolo
  - dal rettangolo si protendono due linee (i baffi) che arrivano al minimo e massimo

Boxplot (2)



Boxplot (3)


revenue
cost
profit
5000.00+
4000.00~5000.00
3000.00~4000.00
2000.00~3000.00
1000.00~2000.00
0.00~1000.00
0.00~500.00
500.00~1000.00
1000.00~1500.00
1500.00~2000.00
2000.00~2500.00
2500.00~3000.00
3000.00~3500.00
3500.00~4000.00
4000.00~4500.00
4500.00~5000.00
5000.00~5500.00
5500.00~6000.00
6000.00+
0.00~1000.00
1000.00~2000.00
2000.00~3000.00
3000.00+

Deviazione standard e varianza

- Varianza
  - dati i valori x_1, ... x_n la varianza è

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum x_i^2 - n\left(\sum x_i\right)^2\right]$$

  - o, in alternativa, con n al denominatore

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

- Scarto quadratico medio
  - radice quadrata della varianza
  - più utile perché è misurata con la stessa unità di misura dei dati

Diagramma dei quantili

- Dato un insieme di dati x_1, ... x_n, il quantile corrispondente al dato x_i è $q_i=(i-0.5)/n$
  - corrisponde più o meno al percentile
- Per un attributo A
  - L'asse X assume valori da 0 a 1
  - L'asse Y assume il possibile range di valori per A
  - per ogni valore x_i traccio un punto di coordinate (q_i,x_i)

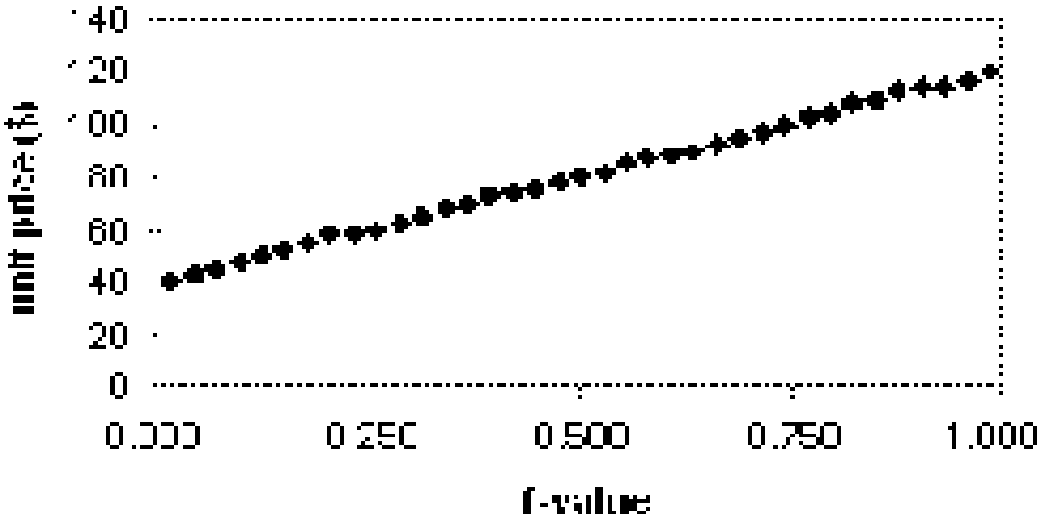
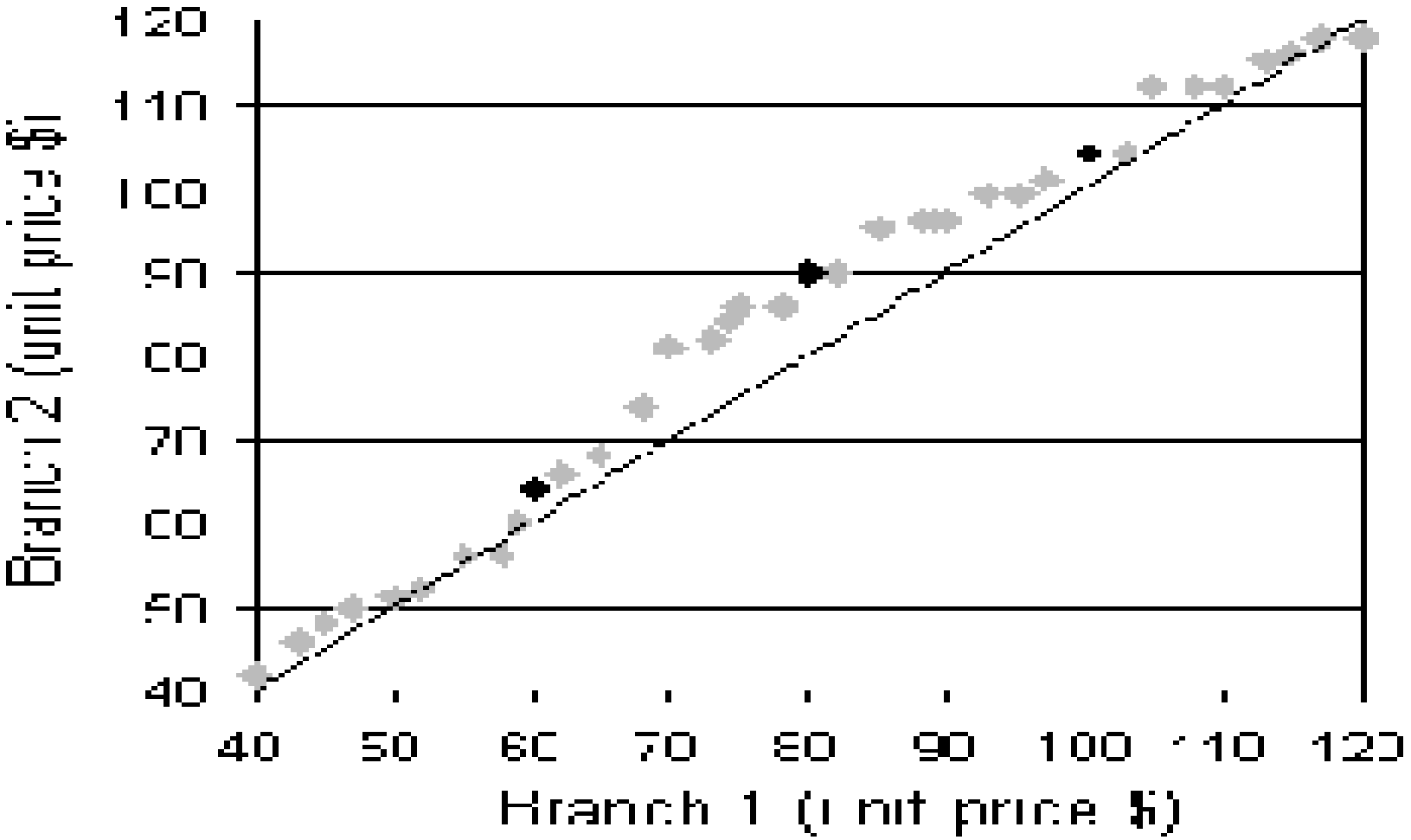


Diagramma quantile-quantile (1)

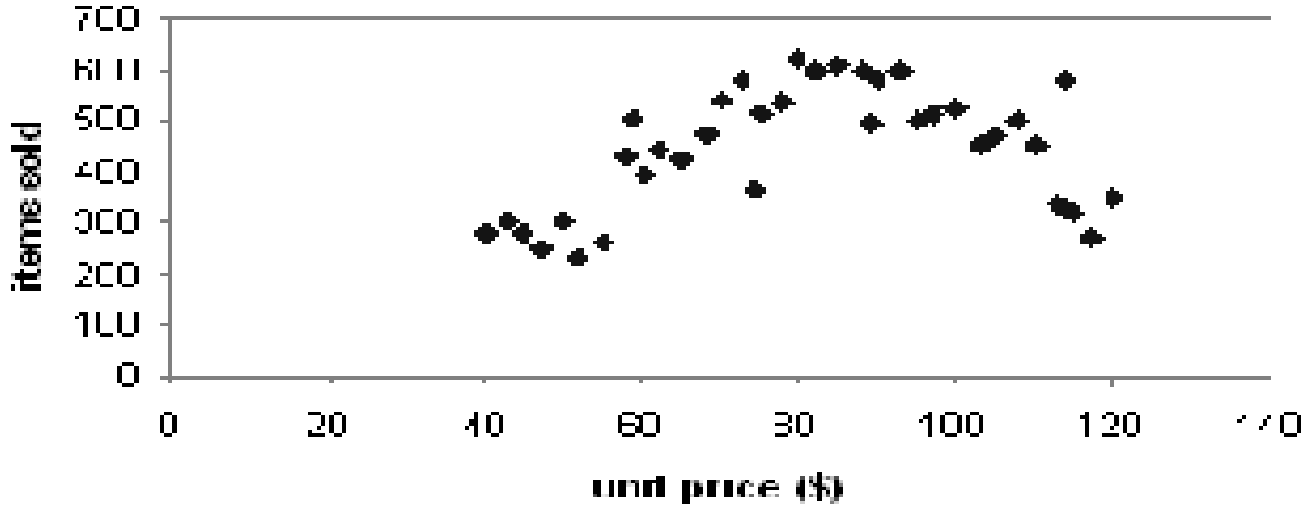
- Visualizza i quantili di una distribuzione univariata rispetto ai quantili di un'altra distribuzione
  - Gli assi X e Y assumono i possibili range di valori per le due distribuzioni
  - Siano x_1, ... x_n i dati della primi distribuzione y_1, ... y_m i dati della seconda
  - se m=n
    - per ogni i, traccio un punto di coordinate (x_i, y_i) (che sono entrambi il quantile (i-0.5)/n della rispettiva distribuzione)
  - se m<n
    - per ogni i=1..m, traccio un punto di coordinate (z_i,y_i) dove z_i è il quantile (i-0.5)/m della distribuzione di A (richiede interpolazione)

Diagramma quantile-quantile (2)



Scatter plot (1)

- Mostra rapidamente se esiste una qualche relazione tra due attributi in una distribuzione bivariata
  - siano (x_1, y_1)... (x_n, y_n) i dati in input per i due attributi scelti
  - gli assi X e Y sono etichettati con i range di possibili valori per gli attributi
  - per ogni i, disegno un punto di coordinate (x_i, y_i)



Scatter plot (2)

- Si può aggiungere una curva ad uno scatter plot che offra una maggiore percezione della relazione tra i due attributi
  - si parla di curva loess (loess=local regression)