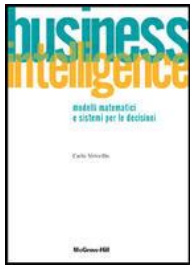
ANALISI DESCRITTIVA E GRAFICA

Analisi Descrittiva e Grafica

Parte dei contenuti della presente lezione sono tratti dai testi elencati di seguito.



Carlo Vercellis (2006). *Business intelligence: modelli matematici e sistemi per le decisioni*, McGraw-Hill.

Introduzione

L'analisi esplorativa ha lo scopo di evidenziare:

- *caratteristiche degli attributi*
- *relazioni esistenti tra attributi*

per un determinato dataset.

Il processo di analisi esplorativa può essere riassunto in tre fasi principali

- **Univariata:** vengono studiate le proprietà di ogni attributo, valutato singolarmente (UNI-variato),
- **Bivariata:** si considerano coppie di attributi (BI-variato) misurando il legame esistente tra loro,
- **Multivariata:** studia i legami esistenti tra più attributi (MULTI-variato).

Analisi Univariata

Consente di studiare il comportamento di un attributo trattandolo come indipendente dal resto degli attributi del dataset. Si studia la propensione dei valori dell'attributo a posizionarsi in prossimità di un *valore centrale* e la propensione ad assumere valori in un *intervallo* più o meno ampio *intorno al valore centrale*.

Utile in quanto modelli di apprendimento presuppongono ipotesi sul comportamento di un attributo per garantire che condurranno a risultati affidabili.

$$D = \{\underline{x}_1, \ldots, \underline{x}_m\}$$

dataset contenente "m" osservazioni, $\underline{\mathbf{X}}_k$ k-ma osservazione.

$$\underline{X}_k = (x_k^1, \ldots, x_k^n)$$

si considerano "n" attributi, $\mathbf{X}^j_k$ k-ma osservazione dell'attributo j-mo.

Analisi Univariata: attributi categorici

Trattandosi di analisi univariata, per semplificare la notazione, scriveremo **X** in luogo di $\underline{X}^j$ e

$$\underline{X} = \{x_1^j, \ldots, x_m^j\}$$

per indicare le "m" realizzazioni appartenenti al dataset "**D**" dell'attributo $\mathbf{X}^j$.

Analisi grafica di attributi categorici

Si utilizzano differenti forme di rappresentazione della distribuzione empirica dell'attributo.

$$V = \{v_1, \ldots, v_H\}$$

viene detto supporto dell'attributo, insieme dei valori che l'attributo categorico può assumere.

Nella fattispecie si assume che l'attributo **X** possa assumere "H" valori distinti.

Analisi Univariata: attributi categorici

Consideriamo un database che contenga informazioni circa la vendita di vetture per i seguenti campi:

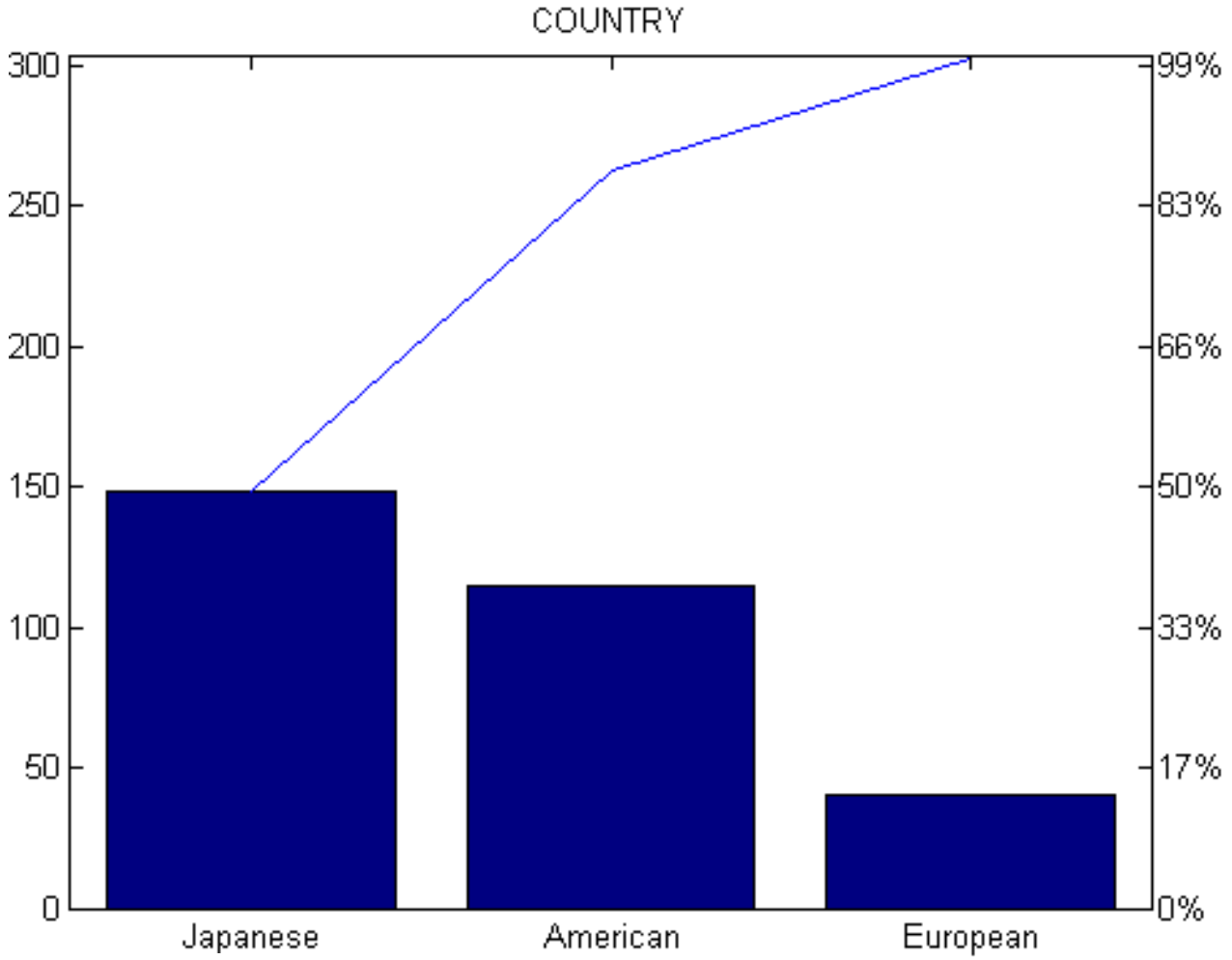
GENDER:	genere dell'acquirente
MARITAL STATUS:	stato civile dell'acquirente
AGE:	età dell'acquirente
COUNTRY:	area geografica di provenienza della vettura
TYPE:	tipo della vettura
SIZE:	dimensione della vettura

e si ponga l'attenzione al campo (attributo) **COUNTRY**, questo attributo può assumere solo i seguenti valori

$$V = \{\text{American}, \text{European}, \text{Japanese}\}$$

Analisi Univariata: attributi categorici

Una rappresentazione grafica per analizzare un attributo categorico è senza dubbio offerta da un *diagramma a barre verticali* il quale riporta in ordinata la frequenza empirica per ogni valore del supporto della variabile considerata.



Indichiamo con "e_h" la frequenza del valore "h-simo" per l'attributo considerato, per quante delle "m" osservazioni del dataset l'attributo considerato assume il valore "h-simo".

Data la frequenza "e_h", la *frequenza empirica relativa* o *densità empirica* "f_h" è definita come segue:

$$f_h = \frac{e_h}{m}$$

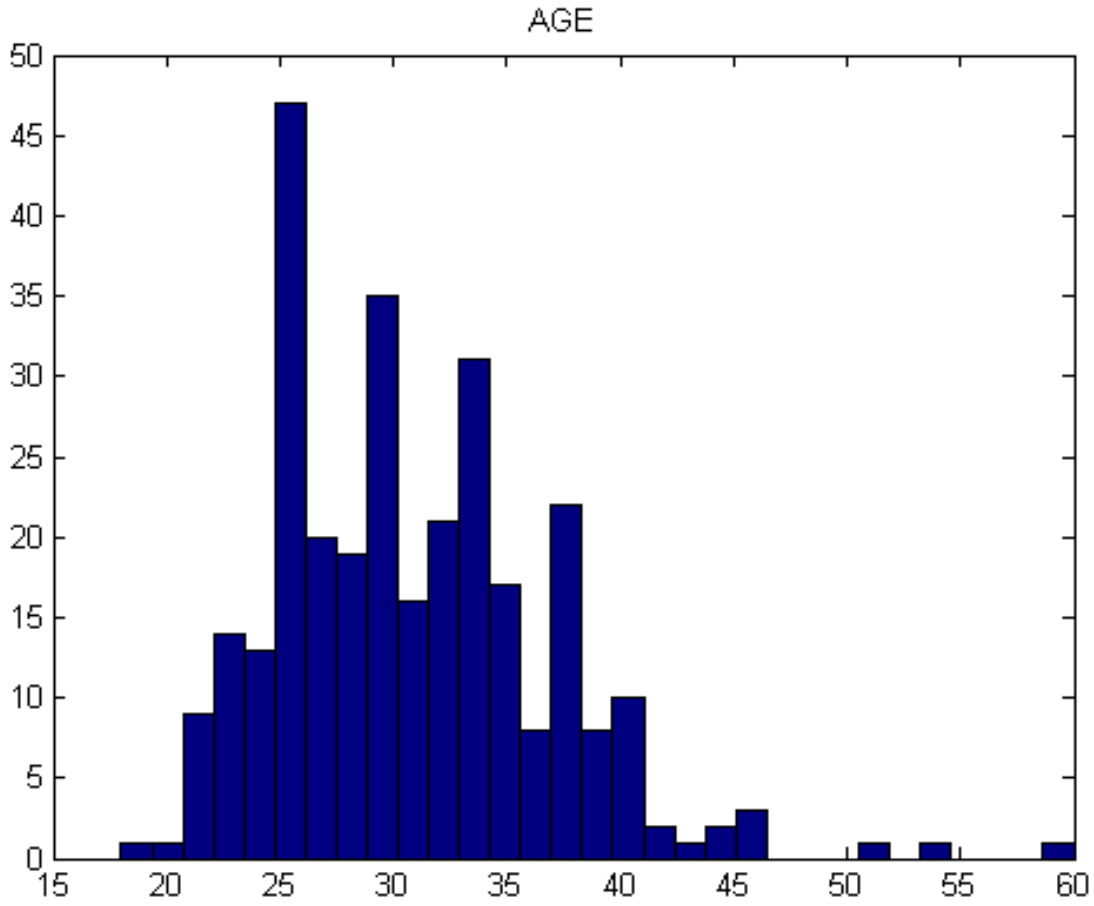
Se il campione a disposizione ha dimensione sufficientemente grande, in base al teorema limite centrale, la funzione di frequenza empirica relativa costituisce una buona approssimazione della densità di probabilità dell'attributo **X**. Più precisamente risulta possibile affermare che per un campione di numerosità sufficientemente elevata vale la seguente relazione

$$f_h \approx p_h = \Pr\{X = v_h\}$$

dove "p_h" rappresenta la probabilità che l'attributo **X** assuma il valore "v_h".

Analisi grafica di attributi numerici

Per *attributi numerici discreti* che assumono un *numero finito e limitato di valori* è possibile ricorrere ad una *rappresentazione mediante diagrammi a barre* come mostrato per gli attributi categorici.



Analisi Univariata: attributi numerici

Per *attributi continui o discreti che assumono infiniti valori* risulta impossibile utilizzare la medesima rappresentazione; si procede alla suddivisione, dell'asse delle ascisse, corrispondente ai valori assunti dall'attributo, in intervalli, solitamente di egual ampiezza, che vengono di fatto assimilati a classi distinte.

In altre parole si procede alla *discretizzazione* (***istogramma***) di un attributo continuo mediante un numero finito e limitato di classi, introducendo pertanto un certo grado di approssimazione.

È opportuno procedere ad una suddivisione in "R" intervalli il più possibile ristretti, compatibilmente con l'esigenza di mantenere limitato il numero di classi distinte così generate.

Una volta determinata la suddivisione, si procede al conteggio del numero di osservazioni "e_r", $r=1, \ldots, R$, che cadono in ciascun intervallo.

Procedura – Istogrammi per la densità empirica

Si sceglie il numero di classi "R", dipendente dal numero "m" di osservazioni del campione, si cerca di ottenere un numero di classi tra 5 e 20 con l'accorgimento che per ogni classe vi siano almeno 5 osservazioni.

Si definisce il range totale e l'ampiezza "l_r" di ciascuna classe, usualmente si divide la differenza tra valore massimo e valore minimo dell'attributo per il numero di intervalli che si desidera ottenere, in modo da ottenere intervalli di egual ampiezza.

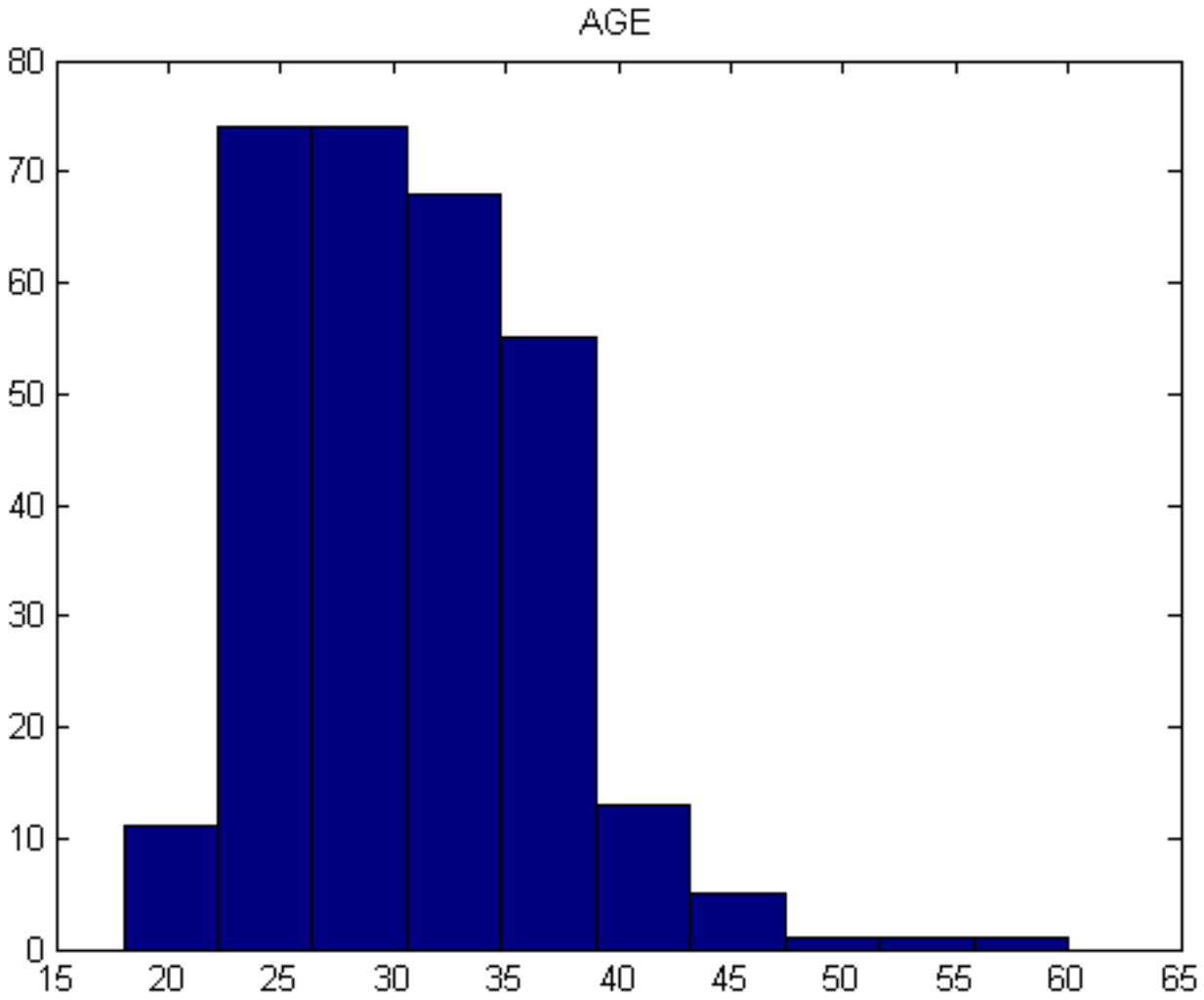
Si assegnano i confini di ogni classe per renderle disgiunte, evitando che vi siano valori che appartengono contemporaneamente a classi contigue. Ogni intervallo è chiuso a sinistra ed aperto a destra.

Si conta il numero di osservazioni in ogni intervallo e si assegna al corrispondente rettangolo un'altezza pari alla densità empirica "f_r" definita come

$$f_r = \frac{e_r}{m}$$

Analisi Univariata: attributi numerici

Considerando l'attributo **AGE**, e fissando il numero di intervalli pari a dieci (R=10) si ottiene l'istogramma riportato sotto.



Analisi Univariata: attributi numerici

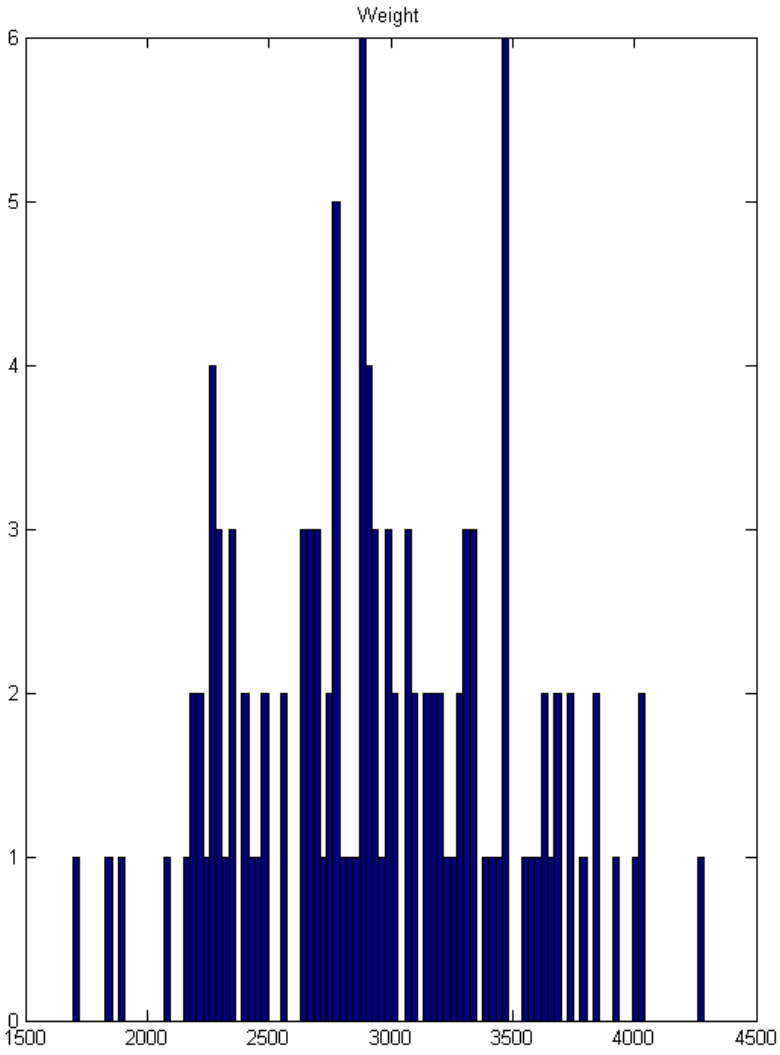
Consideriamo un database che contenga i seguenti campi:

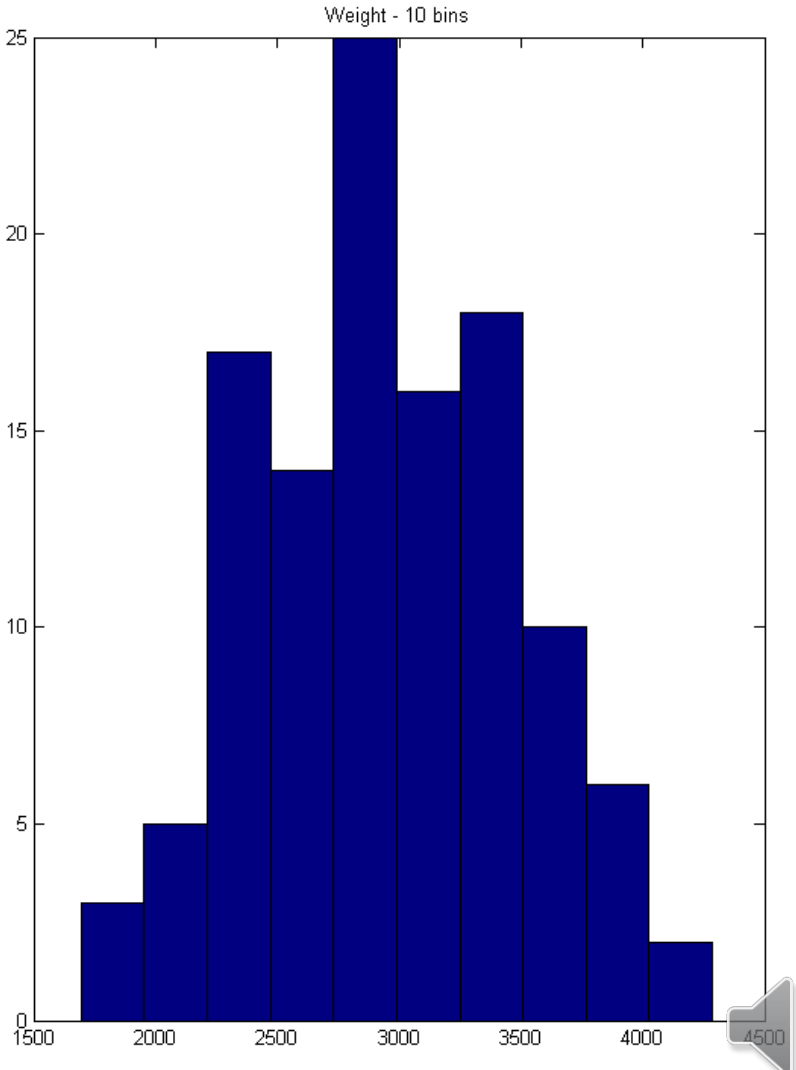
Model:	modello dell'autovettura
Country:	area geografica di provenienza della vettura
Type:	tipo della vettura
Weight:	peso della vettura
Turning Circle:	caratteristica tecnica della vettura
Displacement:	cilindrata
Horsepower:	potenza del motore
Gas Tank Size:	capacità del serbatoio

relativi a diverse vetture.

Analisi Univariata: attributi numerici

Consideriamo l'attributo **Weight**, per il quale riportiamo di seguito il diagramma a barre in due versioni.





Indici di posizionamento centrale per attributi numerici

Ricordiamo di seguito i principali indici di tendenza centrale:

Media $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

Mediana $x^{med} = \begin{cases} x_{(m+1)/2} & \text{se m è dispari} \\ \frac{x_{(m/2)} + x_{(m/2+1)}}{2} & \text{se m è pari} \end{cases}$

Moda x^{mod} *valore più frequente*

Mid-range $x^{midr} = \frac{x^{max} + x^{min}}{2}$

Media geometrica $x^{geom} = \sqrt[m]{\prod_{i=1}^{m} x_i}$

Analisi Univariata: attributi numerici

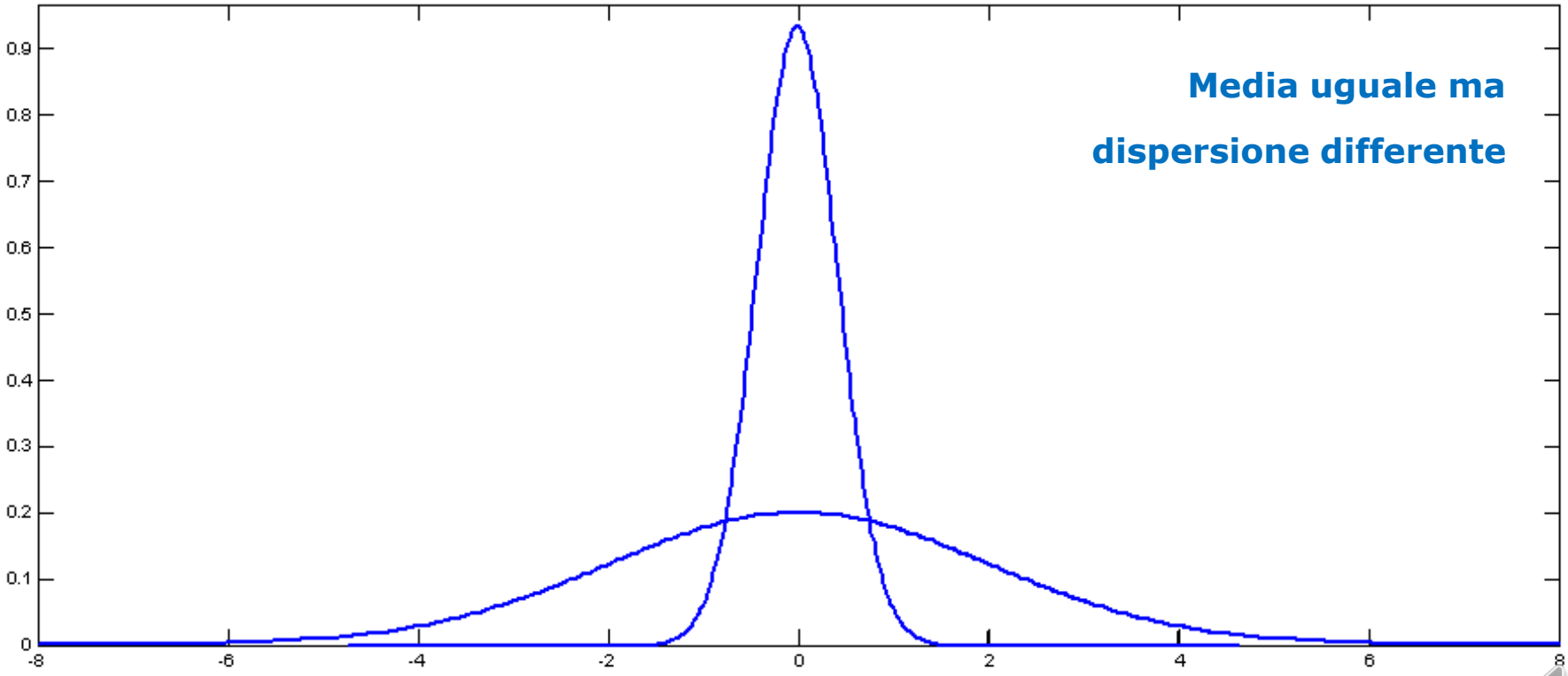
Per l'attributo **Weight**, vengono riportati di seguito i valori degli indici di tendenza centrale.

Weight =

mean:	2.9576e+003
median:	2920
mode:	2885
midrange:	2990
geomean:	2.9085e+003

Indici di dispersione per attributi numerici

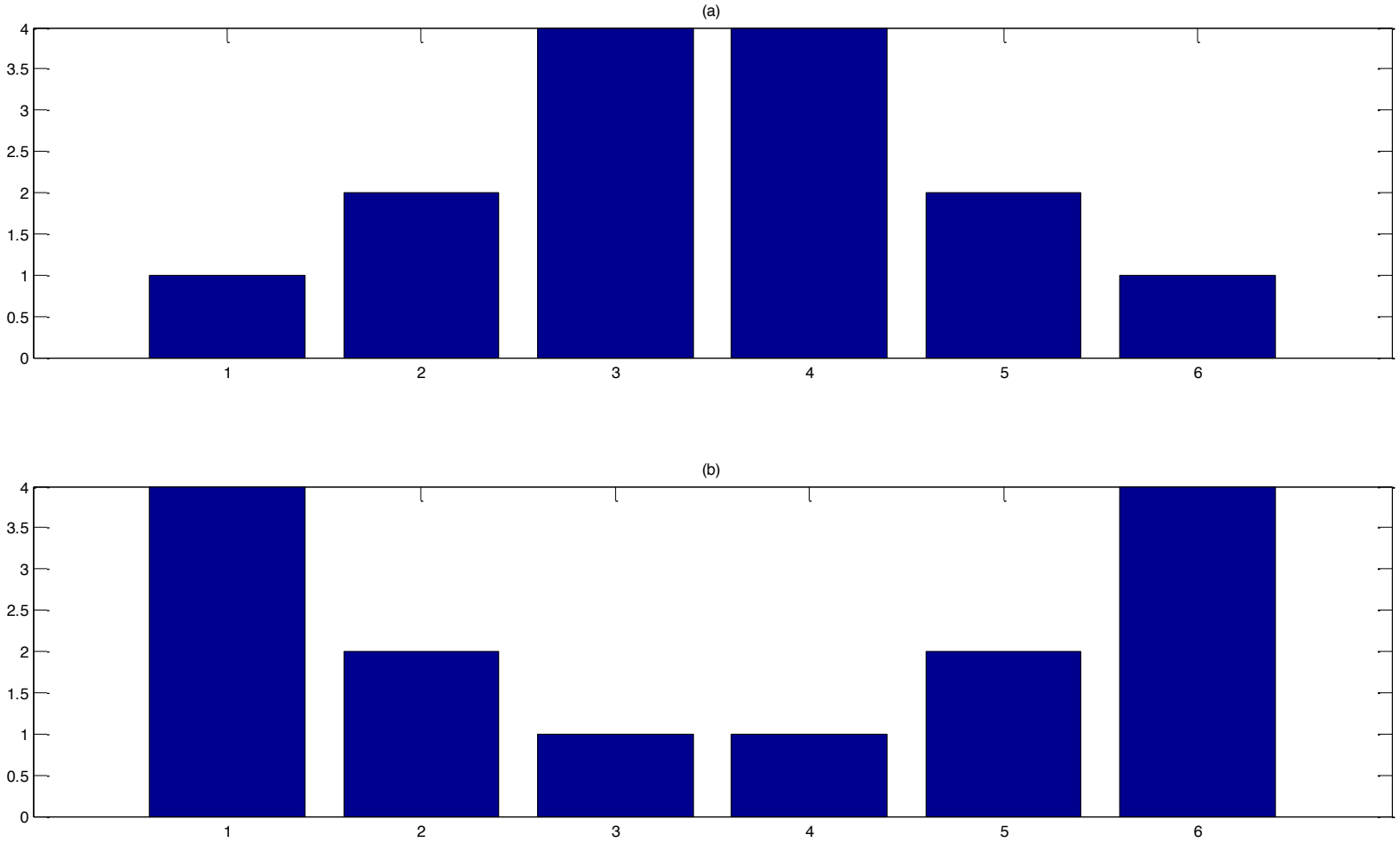
Oltre agli indici di tendenza centrale è rilevante definire altri indici che ci informino circa il livello di *dispersione dei dati*, ovvero il grado di *variabilità* che le osservazioni manifestano *rispetto* ai *valori centrali*.



Analisi Univariata: attributi numerici

Una prima misura di dispersione è rappresentata dal *Range* definito come segue

$$x^{range} = x^{max} - x^{min}$$



Ignora l'effettiva dispersione dei dati, egual range ma dispersioni molto differenti.

La *Deviazione* o *Scarto* di un valore è definita come la differenza con segno dalla media aritmetica campionaria

$$s_i = x_i - \bar{x}$$

Vale la relazione

$$\sum_{i=1}^{m} s_i = 0$$

La dispersione delle osservazioni intorno alla media campionaria viene espressa tramite la *Mean Absolute Deviation* (**MAD**), definita come segue

$$\text{MAD} = \frac{1}{m}\sum_{i=1}^{m}|s_i| = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$$

Un'altra misura importante è rappresentata dalla *Varianza Campionaria*:

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m} s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

e dalla sua radice, la *Deviazione Standard Campionaria*

$$S = \sqrt{S^2}$$

Nel caso particolare in cui la distribuzione dei valori dell'attributo considerato sia *Normale* è possibile ricordare che il

68.27% delle osservazioni apparterranno all'intervallo $\left[\bar{x} - S, \bar{x} + S\right]$

95.45% delle osservazioni apparterranno all'intervallo $\left[\bar{x} - 2S, \bar{x} + 2S\right]$

99.74% delle osservazioni apparterranno all'intervallo $\left[\bar{x} - 3S, \bar{x} + 3S\right]$

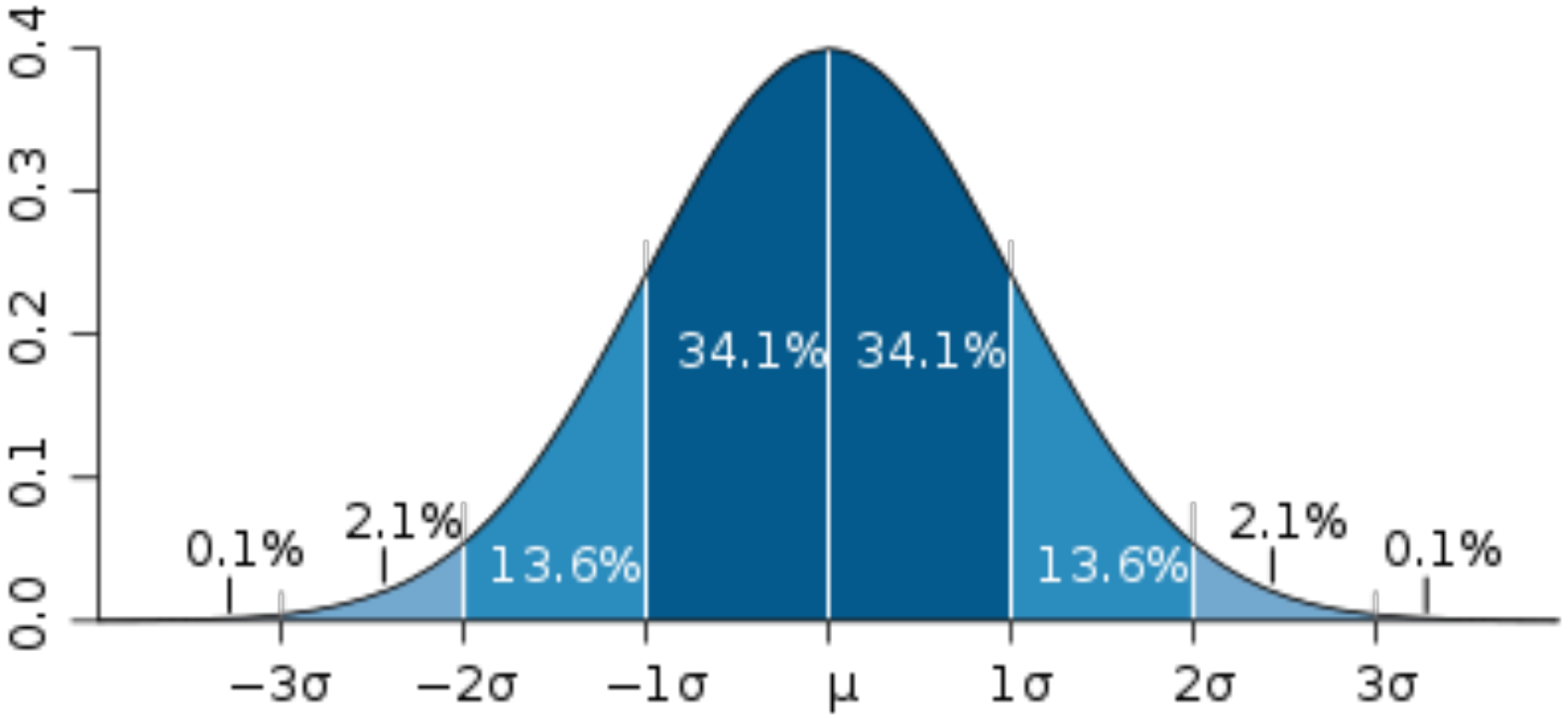
Un ulteriore indice di dispersione è rappresentato dal *Coefficiente di Variazione*

$$CV = 100 \frac{S}{\bar{x}}$$

I valori degli indici vengono riportati di seguito per la variabile **Weight**

range:	2590
variance:	2.8694e+005
mad:	433.7589
std:	535.6635
cv:	18.1112

Analisi Univariata: attributi numerici



Indici di posizionamento relativo per attributi numerici

Sono indici utilizzati per localizzare un valore rispetto ad altri valori del campione.

Quantili

Supponiamo di aver disposto gli "m" valori dell'attributo

$$\{x_1,\ldots, x_m\}$$

in ordine non decrescente

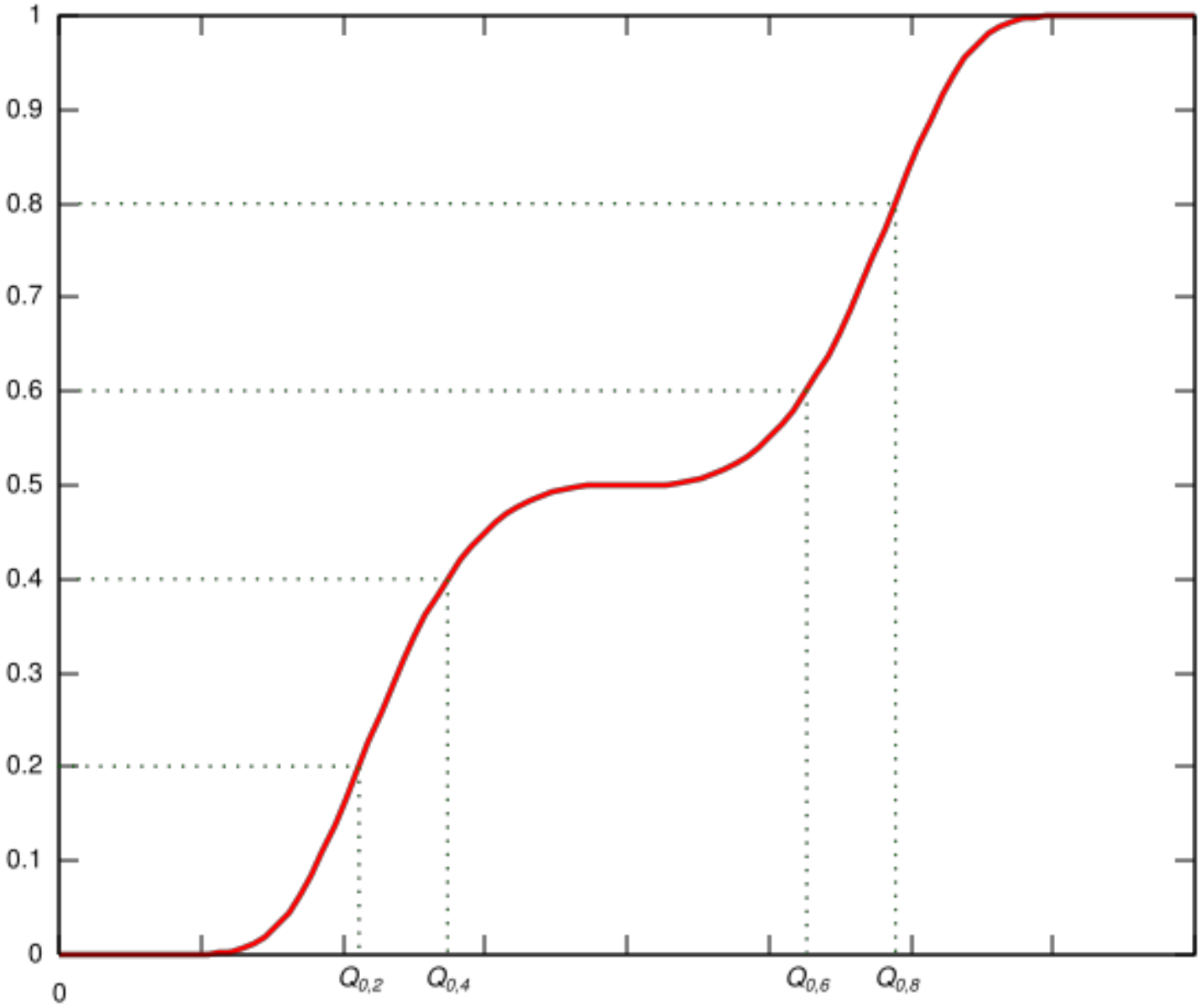
$$\{x_{(1)},\ldots, x_{(m)}\}$$

Dato un qualunque valore

$$p, \qquad 0 \le p \le 1$$

un *quantile di ordine "p"* è un valore "q_p" tale che "pm" osservazioni cadono alla sinistra di "q_p" e le rimanenti "$(1-p)m$" cadono alla sua destra.

Analisi Univariata: attributi numerici



I quantili sono anche noti come *percentili*.

Il *quantile per p=0.5* equivale alla *mediana*.

Il *quantile per p=0.25* viene denominato *quartile inferiore* (q_L) mentre il *quantile per p=0.75* è noto come *quartile superiore* (q_U).

In alcuni casi può essere utile conoscere il valore della *distanza interquantile*, definita come segue:

$$D_q = q_U - q_L = q_{0.75} - q_{0.25}$$

Per ridurre la dipendenza della media da osservazioni estreme è possibile ricorrere ad indicatori di posizionamento centrale basati sui quantili che di norma risultano più robusti di quelli introdotti in precedenza.

La *semi-media*, ottenuta calcolando la media dei valori dell'attributo compresi tra il quartile inferiore ed il quartile superiore.

La *media troncata*, costituisce una generalizzazione della semi-media, dal momento che utilizza per il calcolo della media solo i valori compresi tra i quantili di ordine "*p*" ed "*1-p*". Usualmente si utilizza *p=0.05*.

La *media winsorized*, non esclude dal calcolo i valori che appartengono alle code, ma li codifica secondo una regola intuitiva: i valori minori del quantile di ordine "*p*" vengono incrementati a "q_p", mentre i valori superiori al quantile di ordine "*1-p*" vengono decrementati a "q_{1-p}".

Lo *z-score*, è definito per la "i-esima" osservazione $\boldsymbol{x}_i$, come segue

$$z_i = \frac{x_i - \bar{x}}{S}$$

Identificazione degli outlier per attributi numerici

Si definisce intuitivamente come *outlier* un'osservazione che abbia caratteristica anomala rispetto all'insieme delle restanti osservazioni per l'attributo considerato.

Un modo per identificare un'outlier è offerto dallo z-score, nello specifico vengono considerati come *sospetti outlier* le osservazioni che mostrano un valore dello *z-score maggiore di 3 in valore assoluto* e come *fortemente sospetti outlier* quelle osservazioni per le quali tale valore è *molto maggiore di 3*, sempre in valore assoluto.

Un secondo modo per identificare gli outlier è offerto dalla rappresentazione grafica che va sotto il nome di *diagramma box-and-whisker* che si basa sulla rappresentazione in forma grafica della mediana e dei quartili inferiore e superiore. Un'osservazione viene identificata come outlier se cade all'esterno delle seguenti recinzioni

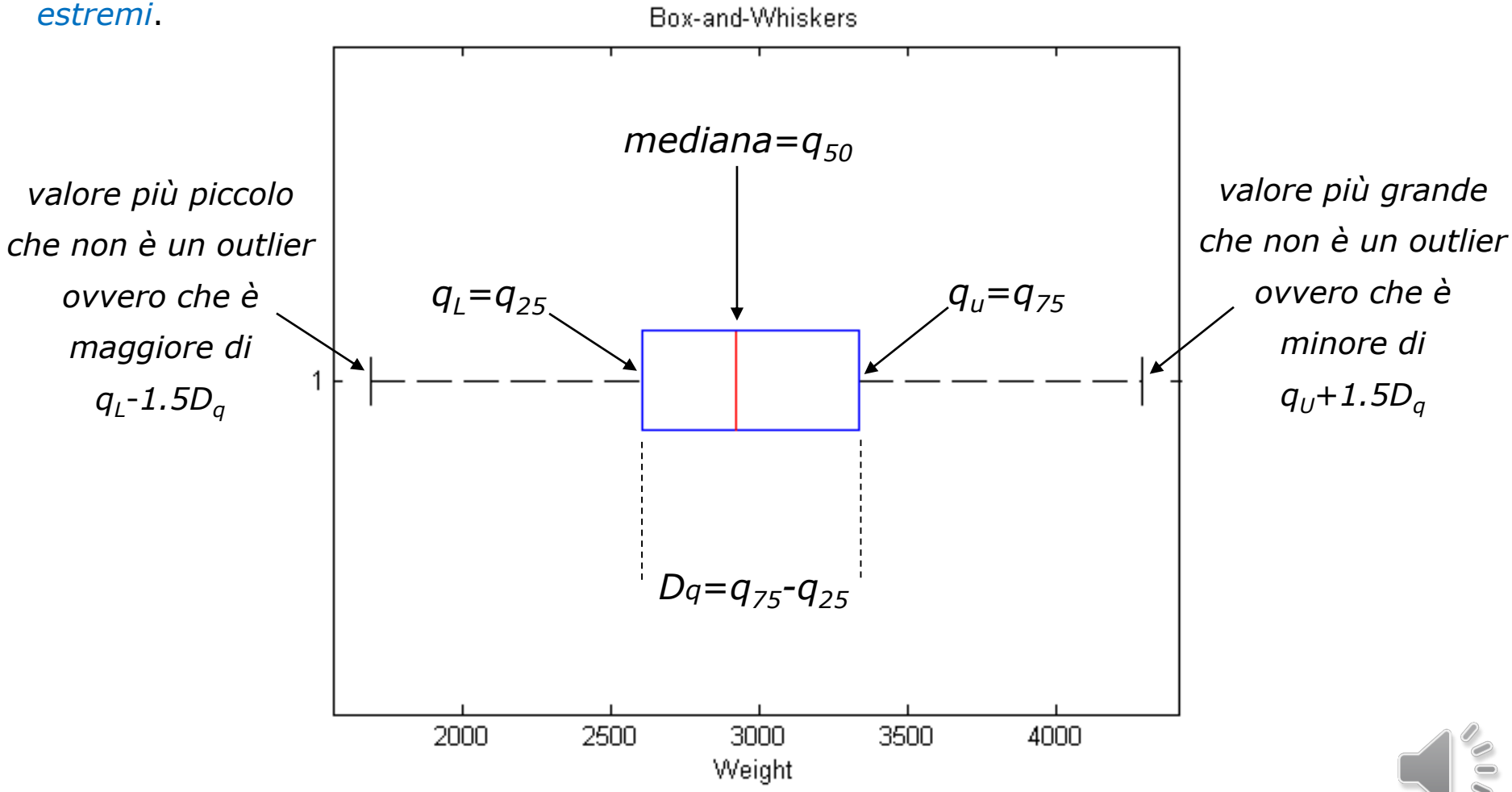
$$bordo\ superiore\ esterno\ \ q_L - 3D_q$$

$$bordo\ inferiore\ interno\ \ q_L - 1.5D_q$$

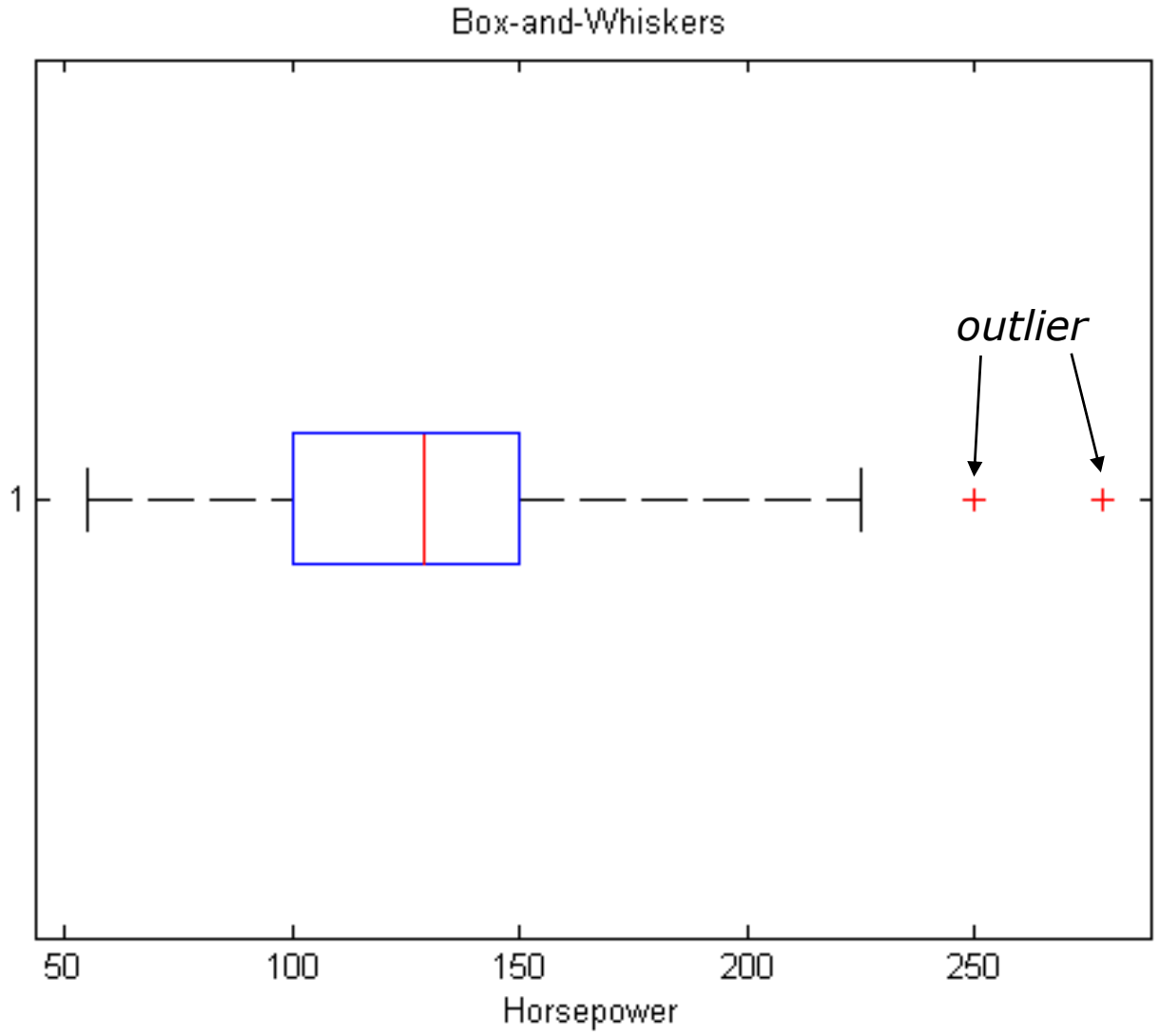
$$bordo\ superiore\ interno\ \ q_U + 1.5D_q$$

$$bordo\ superiore\ esterno\ q_U + 3D_q$$

Il vantaggio rispetto allo z-score è rappresentato dal fatto che la procedura identifica gli outlier indipendentemente dai valori estremi delle osservazioni, infatti di tale indipendenza godono mediana e quartili. Al contrario lo *z-score* è *fortemente influenzato da valori estremi*.



Due outliers, identificati tramite il simbolo "+" di colore rosso.



Indici di eterogeneità per attributi categorici

Nel caso di attributi categorici gli indicatori di posizionamento centrale e relativo, insieme alle misure di dispersione che abbiamo introdotto in precedenza non sono applicabili.

Per un attributo categorico si preferisce di norma ricorrere ad indicatori che esprimono il grado di regolarità con cui i dati

$$\{x_1, \ldots, x_m\}$$

si dispongono all'interno dell'insieme di "*H*" valori distinti.

La *massima eterogeneità* si ottiene nel caso in cui le *frequenze empiriche relative* "f_h" sono *uguali per tutte le classi* "*h*".

La *minima eterogeneità* si ottiene se "$f_g=1$" per una classe "*g*" e "$f_h=0$" per le restanti classi "*h*".

Descriviamo di seguito due indici di eterogeneità per attributi categorici, l'indice di Gini e l'Entropia.

Indice di Gini

Definito come

$$G = 1 - \sum_{h=1}^{H} f_h^2$$

Assume *valore nullo* nel caso di *minima eterogeneità*, ovvero quando una sola classe assume valore con frequenza pari a 1 e tutte le altre classi hanno frequenza pari a zero.

Se tutte le classi assumono egual valore della frequenza empirica relativa ($1/H$), l'indice di Gini assume il suo *valore massimo* pari a

$$G = 1 - \sum_{h=1}^{H} f_h^2 = 1 - \sum_{h=1}^{H} \left(\frac{1}{H}\right)^2 = 1 - \frac{H}{H^2} = \frac{H-1}{H}$$

È possibile normalizzare l'indice in modo tale che assuma valori nell'intervallo [0,1]:

$$G_{rel} = \frac{G}{\left(\frac{H-1}{H}\right)}$$

Indice di Entropia

Definito come

$$E = -\sum_{h=1}^{H} f_h log_2 f_h$$

Assume *valore nullo* nel caso di *minima eterogeneità*, ovvero quando una sola classe assume valore con frequenza pari a 1 e tutte le altre classi hanno frequenza pari a zero.

Se tutte le classi assumono egual valore della frequenza empirica relativa ($1/H$), l'indice di Entropia assume il suo *valore massimo* pari a

$$E = -\sum_{h=1}^{H} f_h log_2 f_h = -\sum_{h=1}^{H} \frac{1}{H} log_2 \frac{1}{H} = -\frac{1}{H} \sum_{h=1}^{H} log_2 \frac{1}{H} = log_2 H$$

È possibile normalizzare l'indice in modo tale che assuma valori nell'intervallo [0,1]:

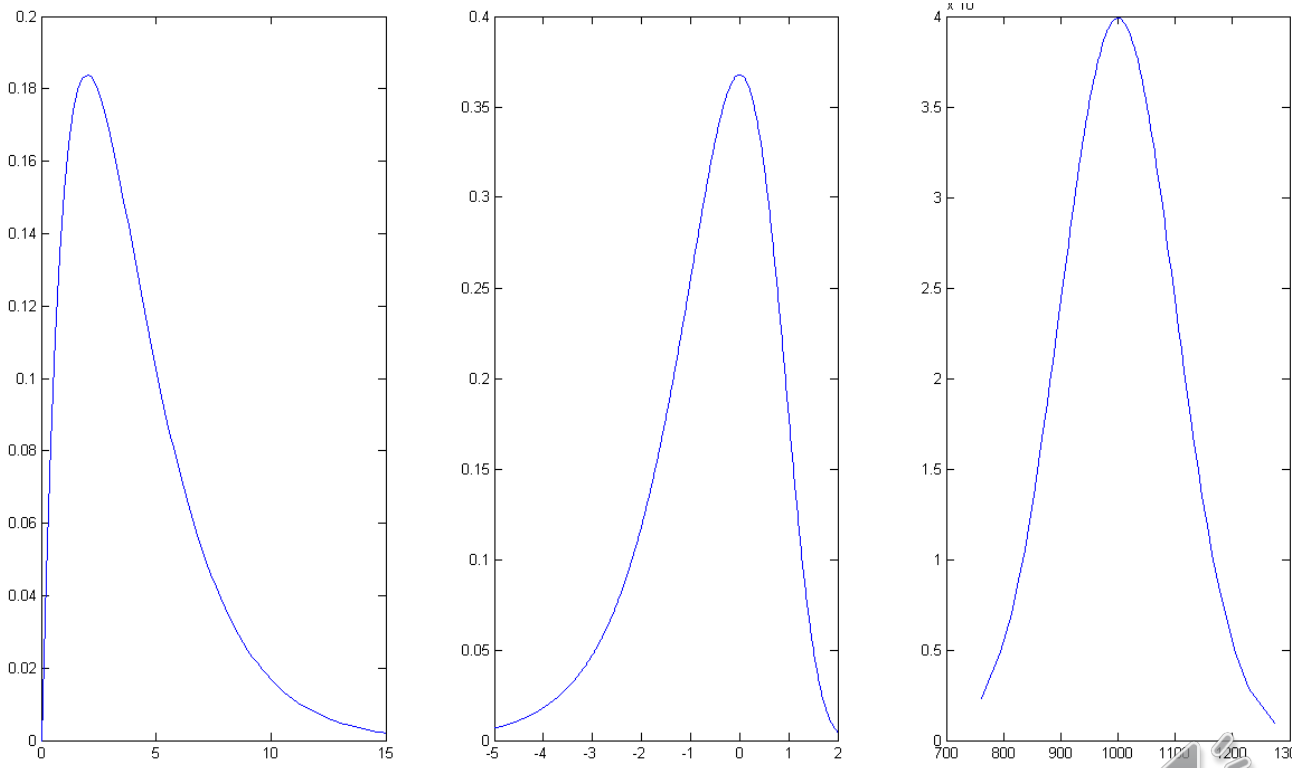
$$E_{rel} = \frac{E}{log_2 H}$$

Analisi della densità empirica

L'istogramma delle frequenze empiriche relative costituisce un importante strumento per l'analisi grafica di attributi sia categorici che numerici. È pertanto fondamentale disporre di indicatori di sintesi che consentano di studiare le proprietà della curva di densità empirica.

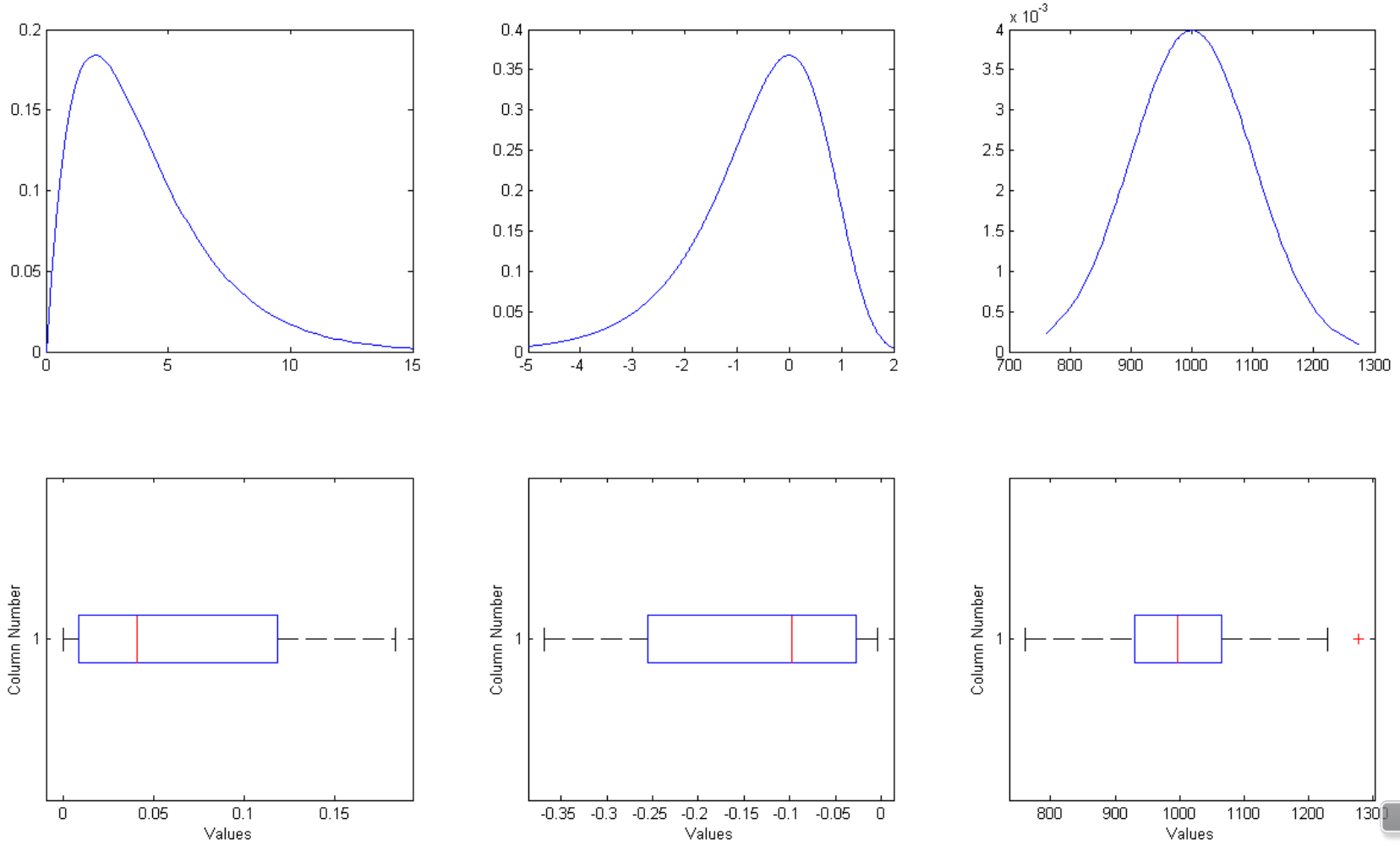
Asimmetria della curva di densità

Una *curva di densità* si dice *simmetrica* se la *media coincide con la mediana*, in caso contrario si dice asimmetrica.



Analisi Univariata: attributi numerici

Il diagramma Box-and-Whisker consente di identificare in modo particolarmente intuitivo la natura ed il grado di asimmetria di una densità empirica, come viene illustrato nella figura sottostante.



Nel caso di un attributo numerico, accanto all'analisi grafica è possibile introdurre un indice di asimmetria basato sul *momento terzo campionario*:

$$\bar{x}^3 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^3$$

L'*indice di asimmetria* è definito come segue:

$$I_{as} = \frac{\bar{x}^3}{S^3}$$

ed è tale per cui se la curva è

- simmetrica $I_{as} = 0$
- asimmetrica a destra $I_{as} > 0$
- asimmetrica a sinistra $I_{as} < 0$

Curtosi della curva di densità

Un altro problema rilevante relativo all'istogramma di densità riguarda la natura della distribuzione di probabilità teorica, solitamente ignota a priori, da cui le osservazioni sono estratte.

Si tratta di un problema di stima della distribuzione incognita a partire da dati, che risulta complesso nella sua forma generale e che pertanto non affronteremo.

Esiste comunque una distribuzione che ricorre con una certa frequenza per la quale sono disponibili elementari criteri grafici e di sintesi per valutare il grado di approssimazione a una densità empirica assegnata.

Il primo criterio, di natura grafica, si basa sul confronto visivo dell'istogramma della densità empirica con una curva normale avente

$$media\ \bar{x} \quad e \quad deviazione\ standard\ S$$

coincidenti con quelle della densità assegnata.

L'*indice di curtosi* esprime in modo sintetico il grado di approssimazione di una densità empirica alla curva normale, esso utilizza il *momento quarto campionario*:

$$\bar{x}^4 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^4$$

ed è definito come segue:

$$I_{curt} = \frac{\bar{x}^4}{S^2} - 3$$

Se la *frequenza empirica corrisponde* perfettamente a quella di una *curva normale* si ha

$$I_{curt} = 0$$

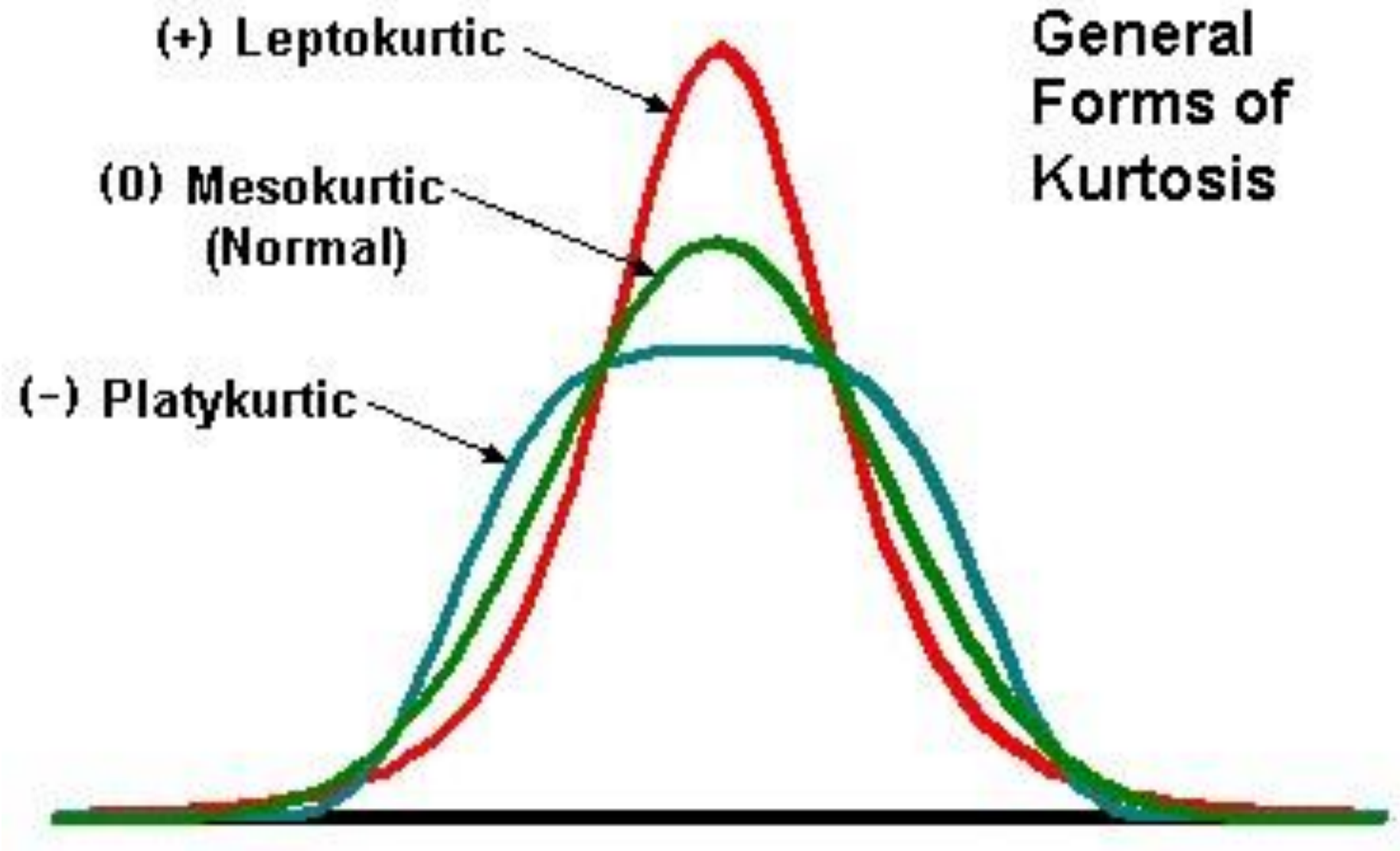
nel caso in cui

$$I_{curt} < 0$$

si parla di *distribuzione iponormale*, mentre se

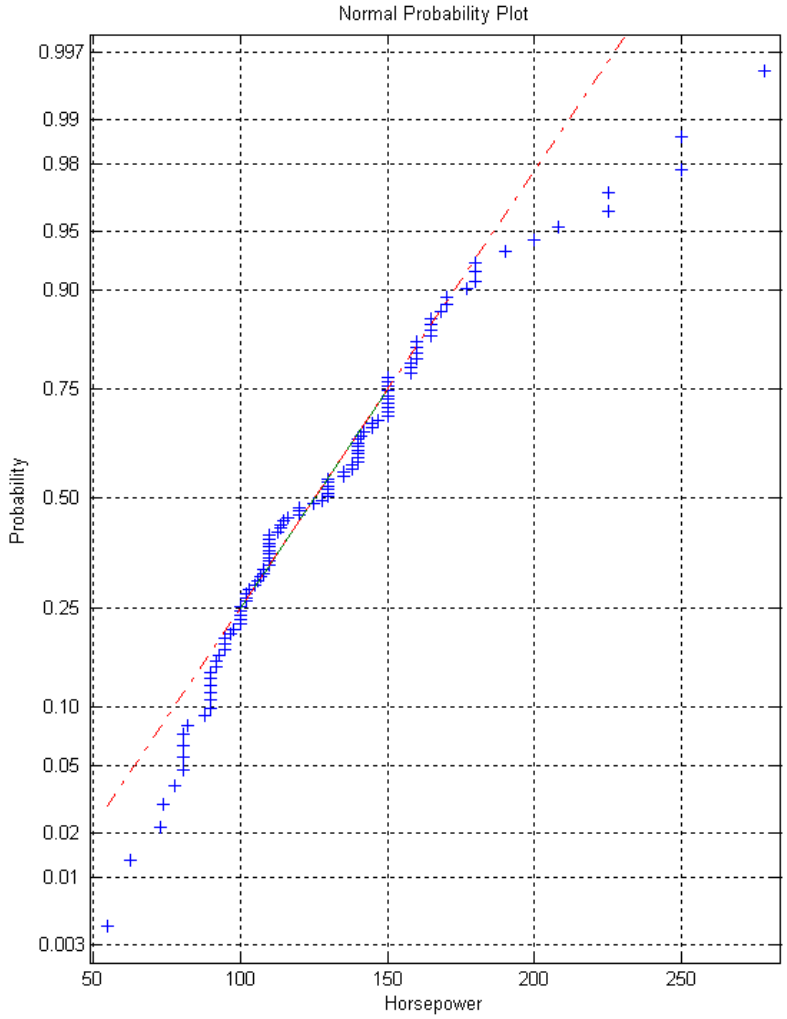
$$I_{curt} > 0$$

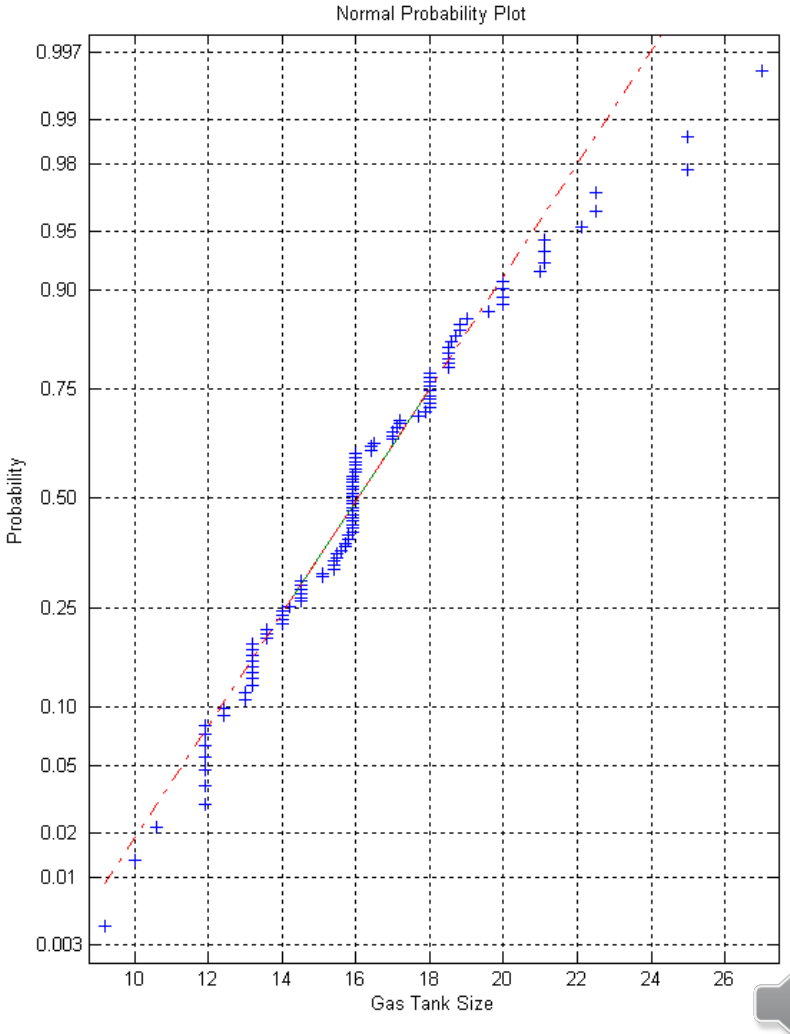
si parla di *distribuzione ipernormale*.


(+) Leptokurtic
(0) Mesokurtic
(Normal)
(-) Platykurtic
General
Forms of
Kurtosis

Analisi Univariata: attributi numerici

È possibile ricorrere anche al *normal probability plot*, che rappresenta un caso particolare di *quantile-quantile plot*.





Analisi Bivariata

Analizziamo ora le relazioni tra coppie di attributi che indicheremo genericamente con X^j e X^k. È possibile distinguere tre casi che possono presentarsi:

- *entrambi gli attributi sono numerici*
- *un attributo numerico e l'altro categorico*
- *entrambi gli attributi sono categorici*

Nel corso delle prossime slide utilizzeremo la seguente notazione:

$$\underline{X}^j = \{x_1^j, \ldots, x_m^j\} \qquad \underline{X}^k = \{x_1^k, \ldots, x_m^k\}$$

Indicando le "*m*" osservazioni per la coppia di attributi considerati ovvero l'attributo X^j e l'attributo X^k.

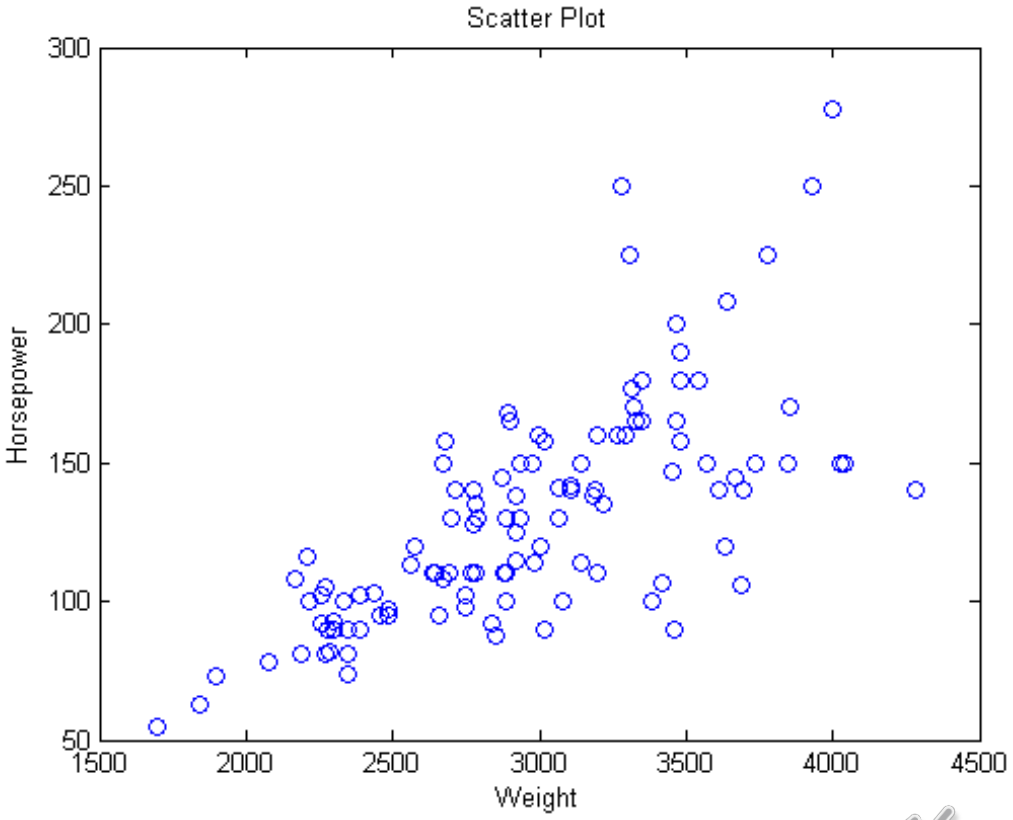
Analisi grafica

Esistono diversi tipi di visualizzazione grafica che permettono di studiare la relazione tra due attributi.

Diagrammi di dispersione

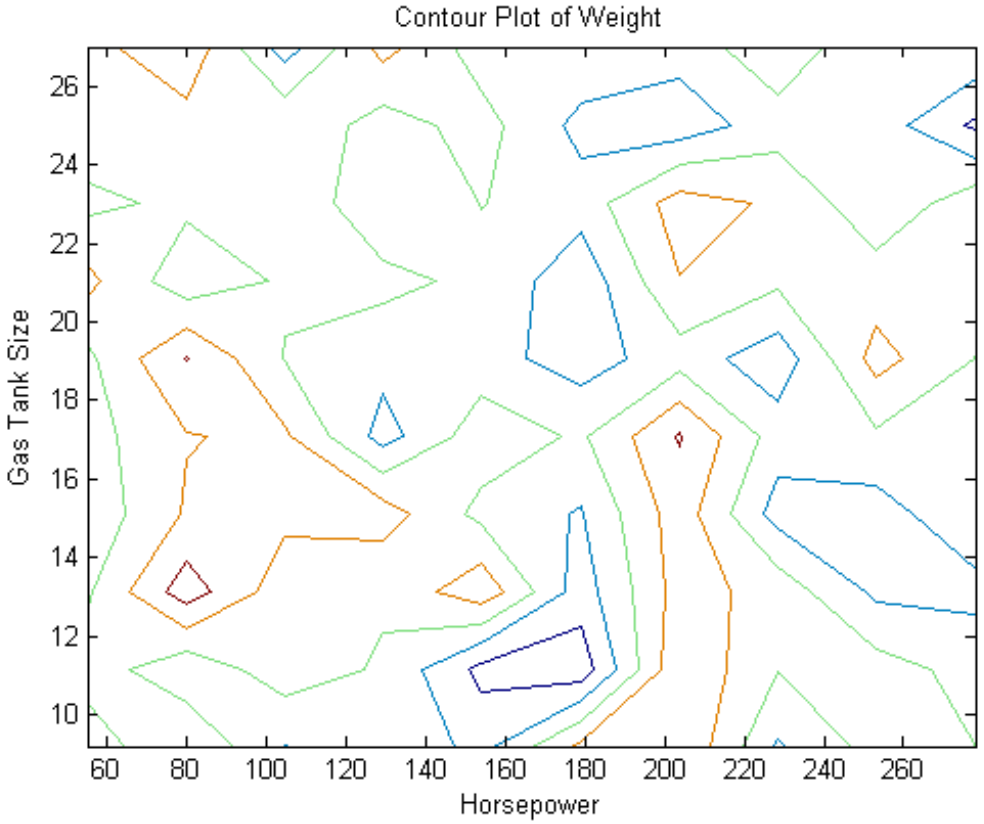
Un *diagramma di dispersione* (*scatterplot*) è la rappresentazione grafica più intuitiva del legame esistente tra due attributi numerici.

Nella figura a destra vengono presi in considerazione l'attributo **Horsepower** e l'attributo **Weight**.



Curve di livello

Costituiscono un'evoluzione dei diagrammi di dispersione e sono pertanto applicabili solo ad attributi numerici. Vengono di norma utilizzate per evidenziare il valore di un terzo attributo numerico X^z al variare dei due attributi X^j e X^k collocati sugli assi del diagramma.



Analisi Bivariata: analisi grafica

Diagrammi quantili-quantili

Consentono di confrontare tra loro le distribuzioni del medesimo attributo in corrispondenza di due diverse caratteristiche della popolazione o di campioni estratti da differenti popolazioni (i campioni devono avere egual cardinalità).

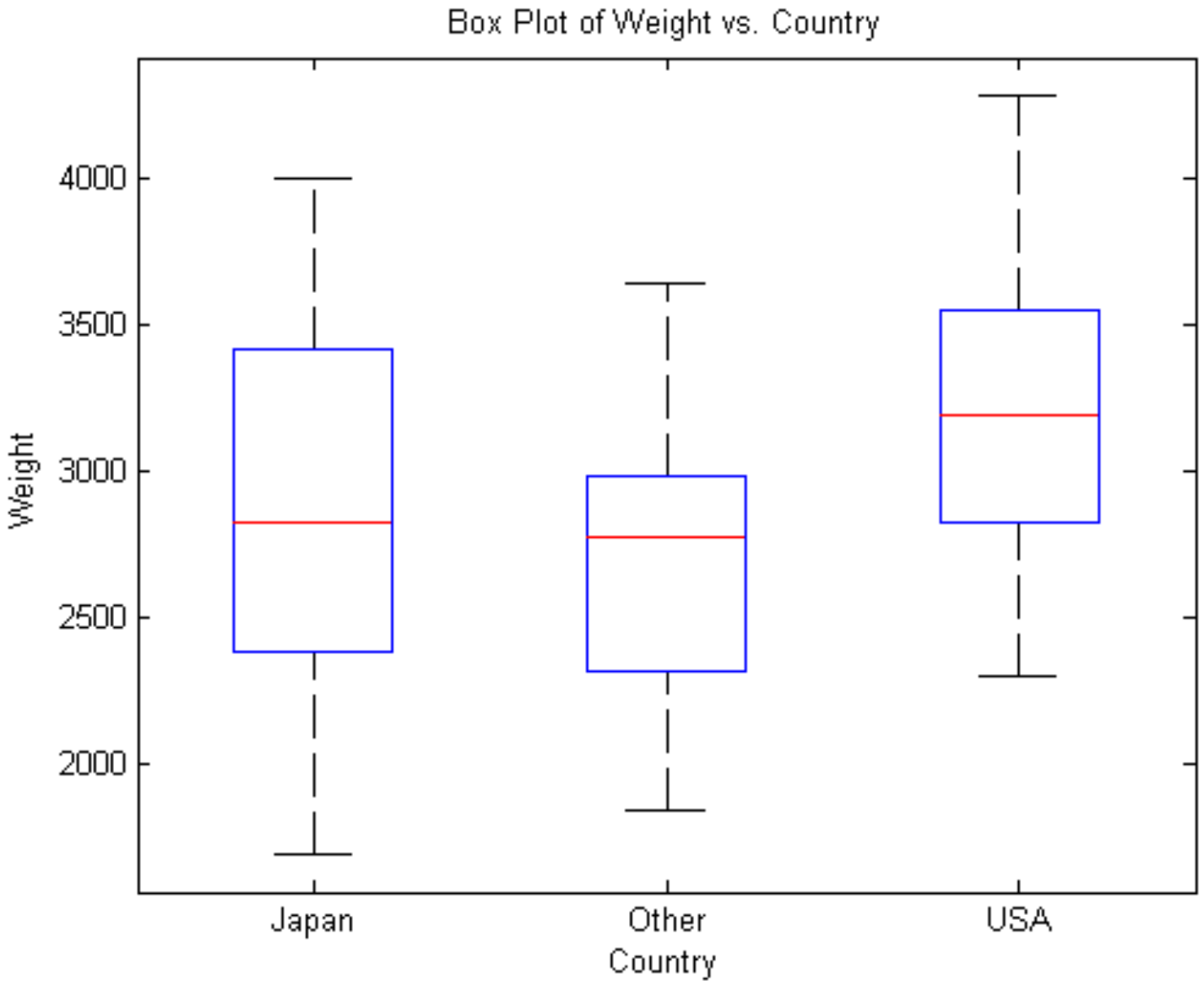
Se i punti si dispongono lungo la linea a 45° allora c'è ragione di pensare che i due campioni analizzati provengano dalla medesima distribuzione.

Diagrammi Box-and-Whisker

Un ulteriore utilizzo di tali diagrammi consiste nel confrontare le distribuzioni della medesima variabile per osservazioni appartenenti a gruppi distinti.

Tecnica applicabile ad una coppia di attributi formata da un attributo continuo ed un attributo categorico.

Analisi Bivariata: analisi grafica



Indici di associazione per attributi numerici

Come nel caso di analisi univariate, è utile introdurre accanto ai metodi grafici anche indicatori sintetici che esprimono la natura e l'intensità del legame tra attributi numerici.

Covarianza

Data la coppia di attributi X^j e X^k è definita come segue:

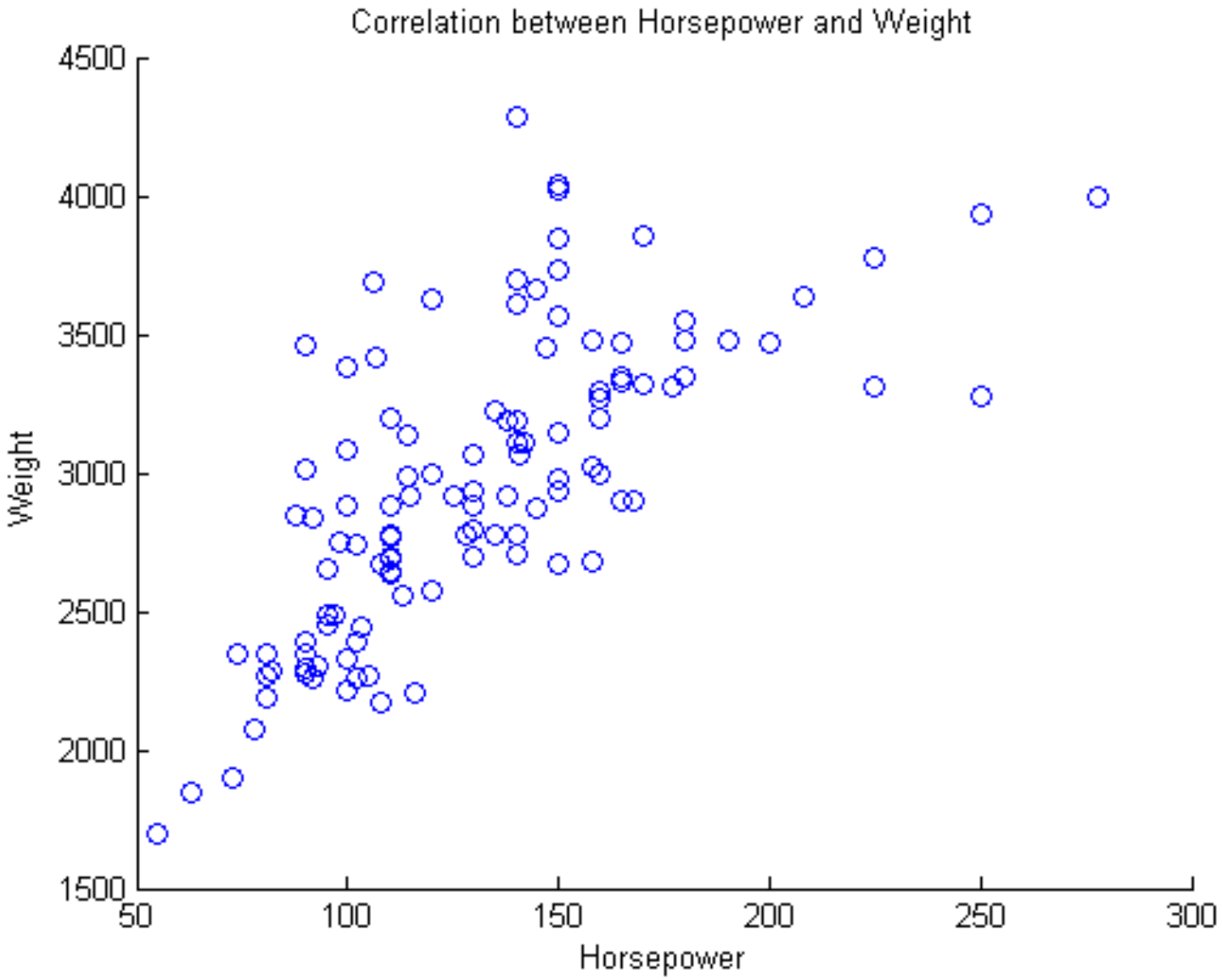
$$v^{jk} = \mathrm{cov}\left(X^j, X^k\right) = \frac{1}{m-2}\sum_{i=1}^{m}\left(x_i^j - \bar{x}^j\right)\left(x_i^k - \bar{x}^k\right)$$

Correlazione

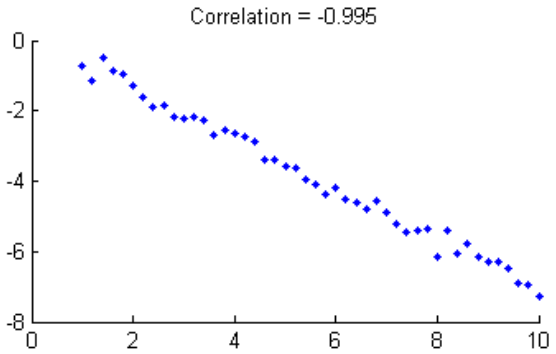
Data la coppia di attributi X^j e X^k è definita come segue:

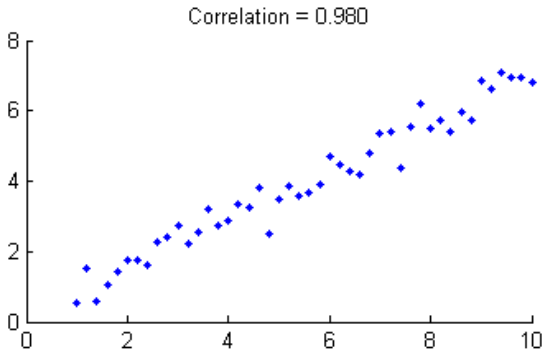
$$r^{jk} = \mathrm{corr}\left(X^j, X^k\right) = \frac{v^{jk}}{S^j S^k}$$

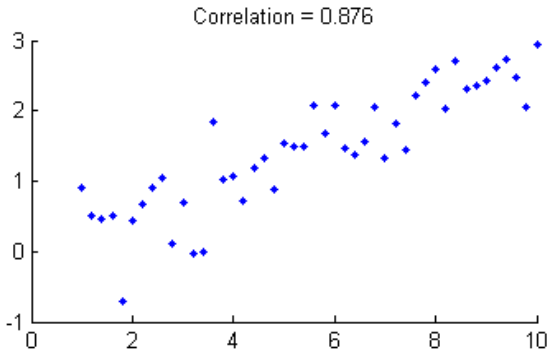
Analisi Bivariata: indici di associazione

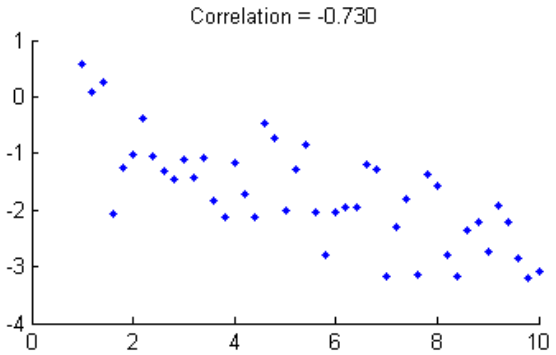


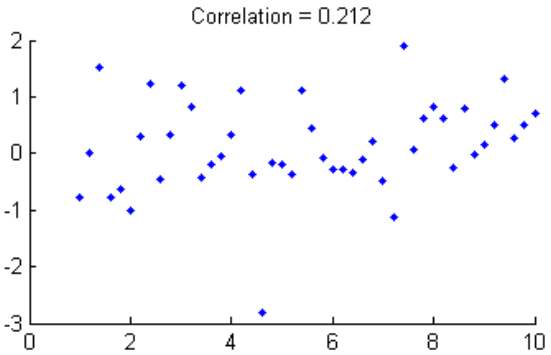
Analisi Bivariata: indici di associazione

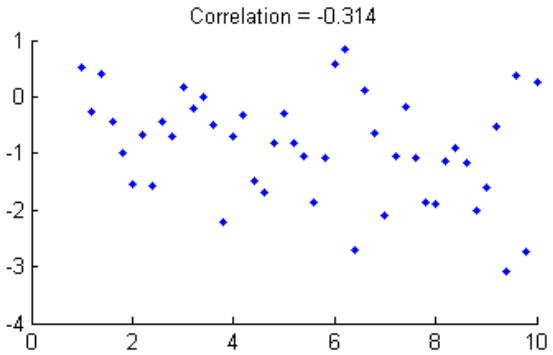


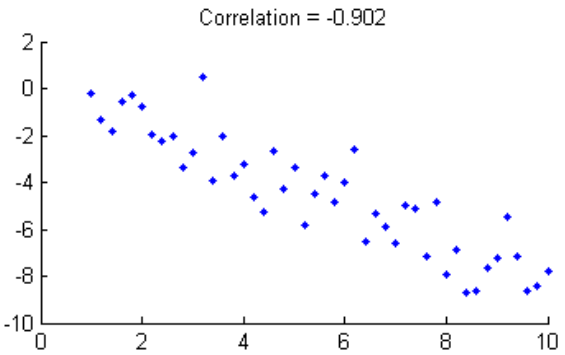


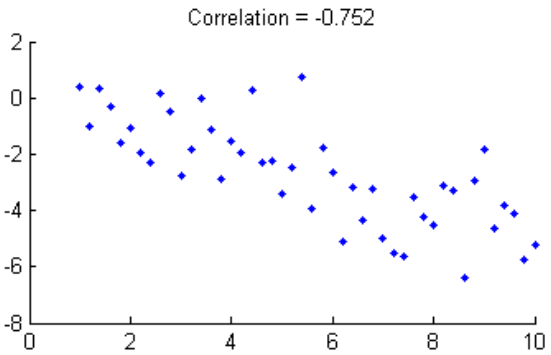












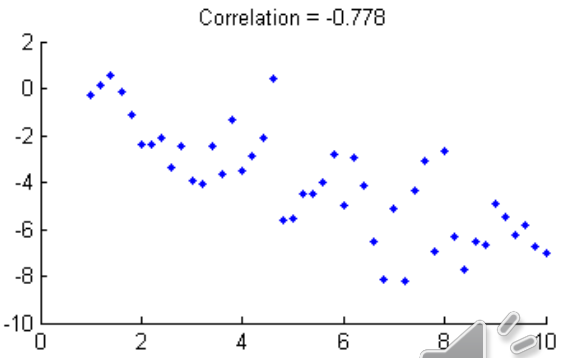


Tabelle di contingenza per attributi categorici

Indichiamo con

$$V = \{v_1, \ldots, v_J\} \qquad U = \{u_1, \ldots, u_K\}$$

gli insiemi di valori distinti assunti dagli attributi categorici $\mathbf{X^j}$ e $\mathbf{X^k}$.

È possibile costruire una *tabella di contingenza*, costituita da una matrice "**T**" il cui generico elemento "t_{rs}" indica la frequenza con cui è presente tra i record del dataset la coppia di valori:

$$\{x_i^j = v_r\} \qquad \{x_i^k = u_s\}$$

		Type				
		small	*medium*	*compact*	*large*	*sporty*
Country	*Japan*	7	6	3	4	10
	Other	12	8	12	1	4
	USA	3	16	7	12	11

È possibile calcolare la somma dei valori per ogni riga e per ogni colonna ottenendo le *frequenze marginali*:

$$f_r = \sum_{s=1}^{K} t_{rs} \qquad g_s = \sum_{r=1}^{J} t_{rs}$$

		Type					
		small	*medium*	*compact*	*large*	*sporty*	
Country	*Japan*	7	6	3	4	10	**30**
	Other	12	8	12	1	4	**37**
	USA	3	16	7	12	11	**49**
		22	**30**	**22**	**17**	**25**	**116**

Gli attributi $\mathbf{X^j}$ e $\mathbf{X^k}$ vengono detti *stocasticamente indipendenti* se

$$\frac{t_{r1}}{g_1} = \frac{t_{r2}}{g_2} = \ldots = \frac{t_{rK}}{g_K}, \qquad r = 1,2,\ldots, J$$

NOTA: non è esattamente così !!!

È possibile mostrare che la condizione precedente equivale alla seguente

$$\frac{t_{1s}}{f_1} = \frac{t_{2s}}{f_2} = \ldots = \frac{t_{Js}}{f_J}, \qquad s = 1,2,\ldots, K$$

In termini intuitivi ed informali *due attributi sono stocasticamente indipendenti se l'analisi condotta su uno dei due attributi, disponendo della conoscenza dei valori assunti dall'alltro attributo, porta ad ottenere gli stessi risultati ricavati senza disporre della conoscenza dei valori assunti dall'altro attributo*.

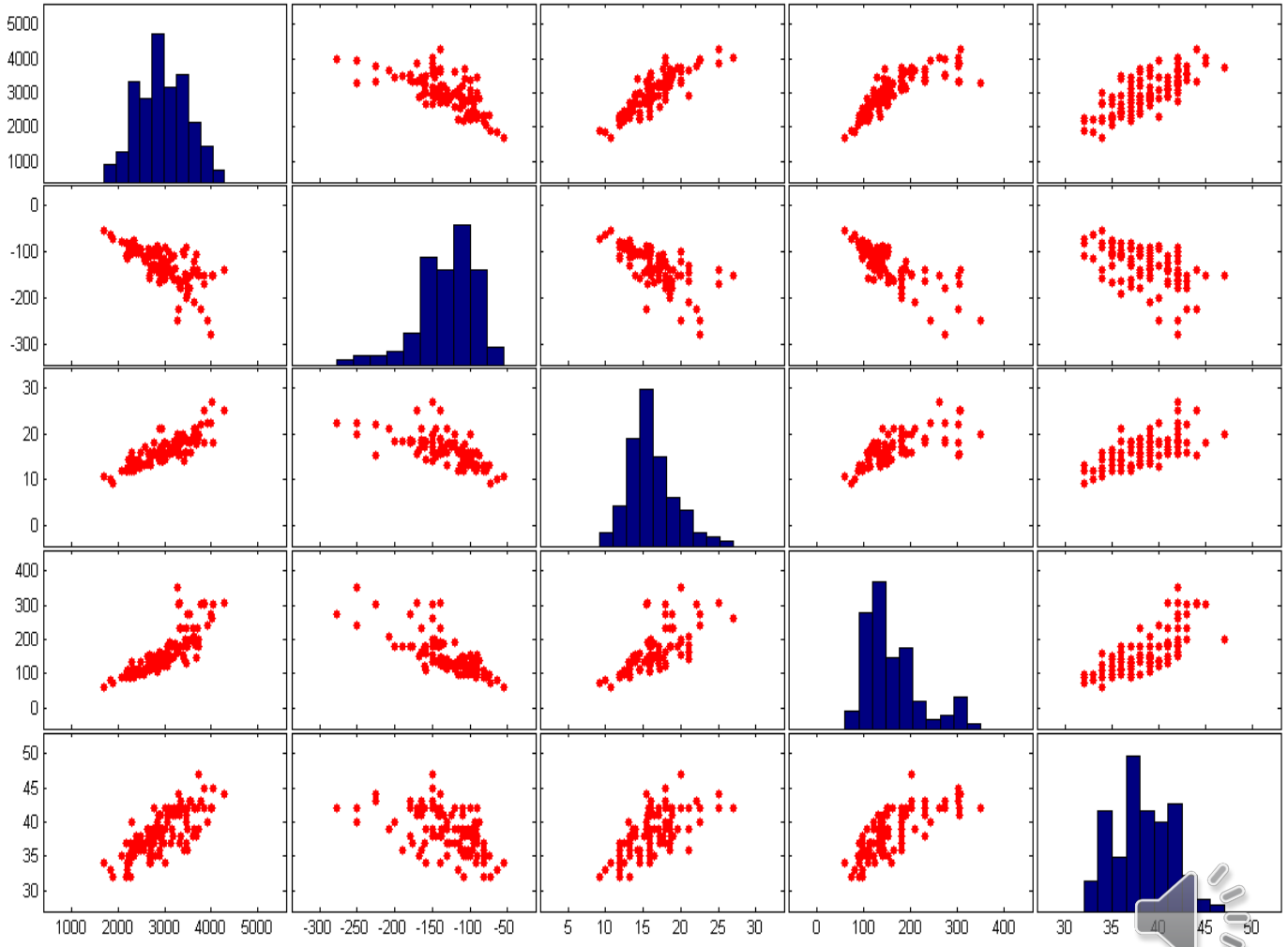
Analisi Multivariata

Si propone di estendere le nozioni introdotte nel caso bivariato per valutare le relazioni che sussistono tra molteplici attributi di un dataset.

Analisi grafica

Tutti i metodi grafici che presentiamo di seguito si applicano solo ad attributi numerici.

Matrice di diagrammi di dispersione

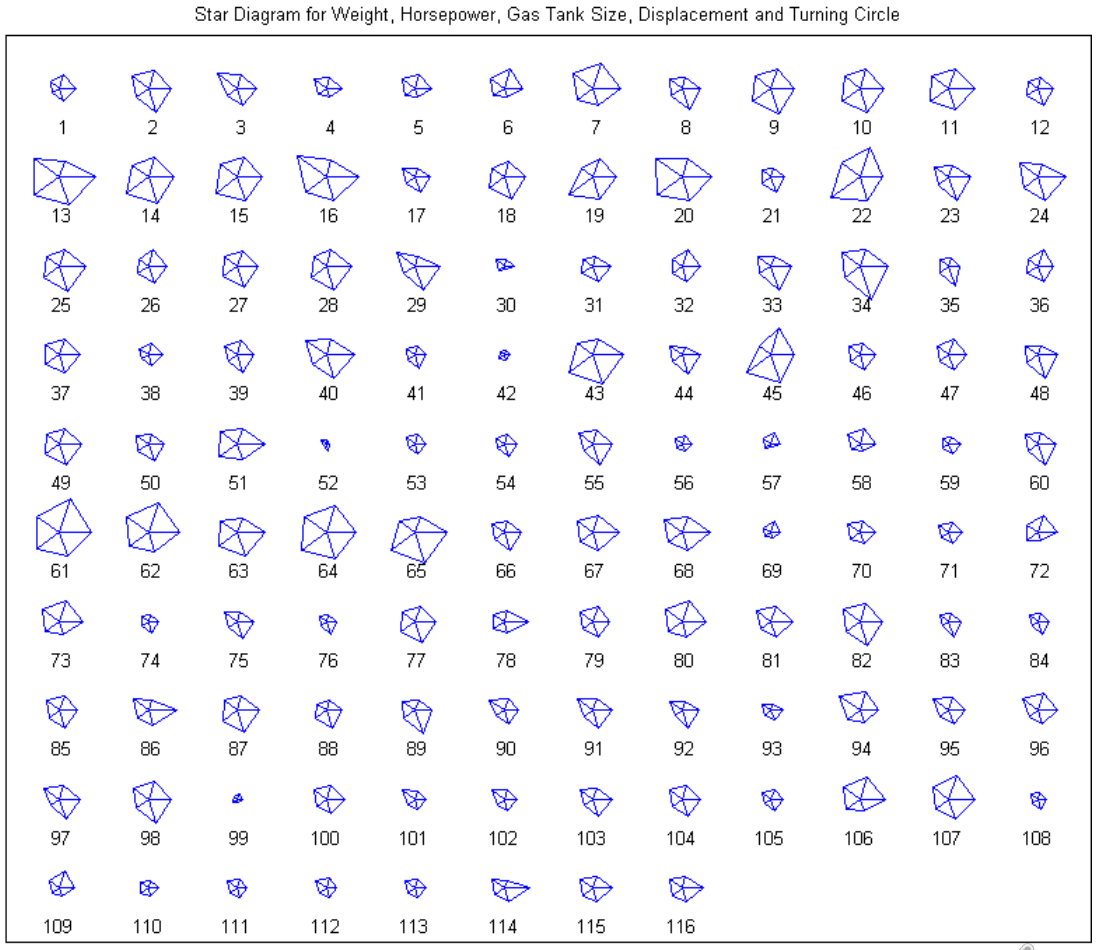


Diagrammi a stella

Appartengono alla più ampia classe dei diagrammi basati su icone. Consentono di evidenziare in modo intuitivo le differenti caratteristiche di ogni specifica osservazione del dataset.

Ad ogni osservazione corrisponde un'icona a forma di stella dal cui centro dipartono *tanti raggi quanti* sono gli *attributi*.

La *lunghezza* di ogni *raggio* è *pari al valore del corrispondente attributo, normalizzato nell'intervallo* [0,1] in modo da avere una rappresentazione omogenea dei diversi attributi.



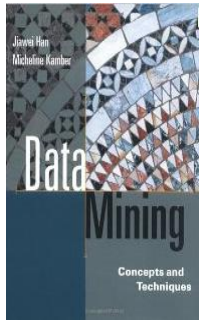
Analisi Multivariata: analisi grafica

Star Diagram for Weight, Horsepower, Gas Tank Size, Displacement and Turning Circle

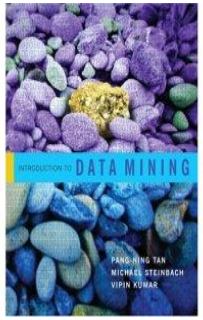
PREPROCESSING

Preprocessing

Parte dei contenuti della presente lezione sono tratti dai testi elencati di seguito.



Jiawei Han and Micheline Kamber (2001). *Data Mining: Concepts and Techniques*, Academic Press.



Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2006). *Introduction to Data Mining*, Pearson International.

Preprocessing

Il preprocessing è costituito da un insieme di strategie e tecniche che stanno in complesse relazioni tra loro. Presenteremo di seguito alcune delle idee e degli approcci più importanti, cercando di chiarire di volta in volta le relazioni che le legano.

Presenteremo i seguenti argomenti

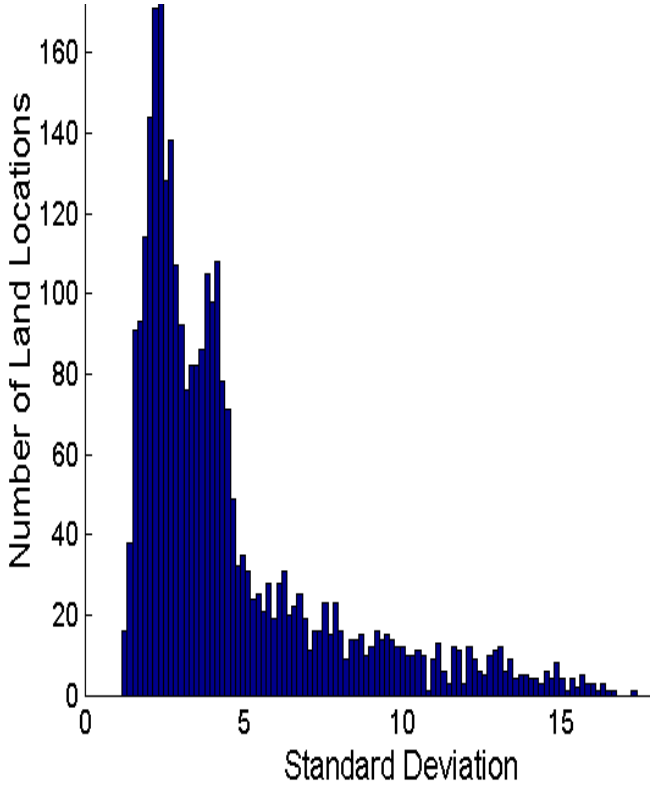
- *Aggregazione*
- *Campionamento*
- *Riduzione Dimensionalità*
- *Feature Subset Selection*
- *Feature Creation (construction)*
- *Discretizzazione e Binarizzazione*
- *Trasformazione delle variabili*

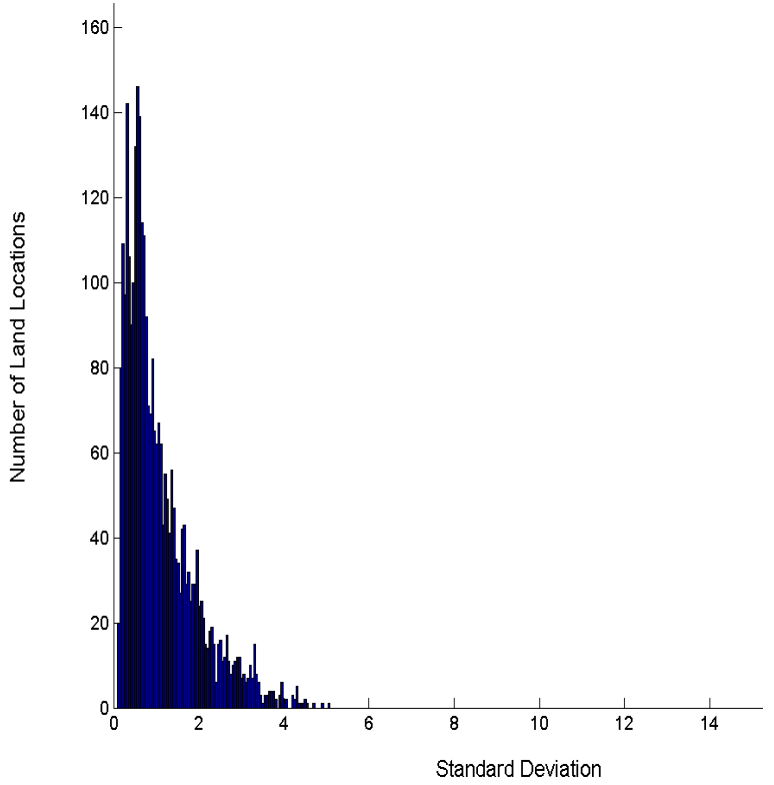
Preprocessing: aggregazione

Aggregazione: combinare due o più oggetti (o attributi) in un singolo oggetto (o attributo)

Obiettivi

- *Data reduction*; ridurre il numero di attributi associati agli oggetti
- *Cambio di scala*; città aggregate in regioni, stati, …
- *Dati più stabili*; dati aggregati tendono ad avere minor variabilità





Preprocessing: campionamento

Campionamento: tecnica principale per selezionare le osservazioni, impiegata sia nelle fasi preliminari che nelle fasi conclusive di un'analisi di Data Mining.

Obiettivi

- *Ridurre sforzo economico*; si campiona per risparmiare soldi
- *Ridurre sforzo computazionale*; si campiona per risparmiare tempo

Il principio base sul quale si fonda il campionamento è il seguente:

- *Utilizzare un campione, in luogo dell'intero insieme di dati, porta ad ottenere risultati comparabili con quelli che si otterrebbero se venisse utilizzato l'intero insieme dei dati, a patto che il campione estratto sia rappresentativo dell'intero insieme di dati.*
- *Un campione è rappresentativo se gode approssimativamente delle stesse proprietà (di interesse) delle quali gode l'intero insieme di dati.*

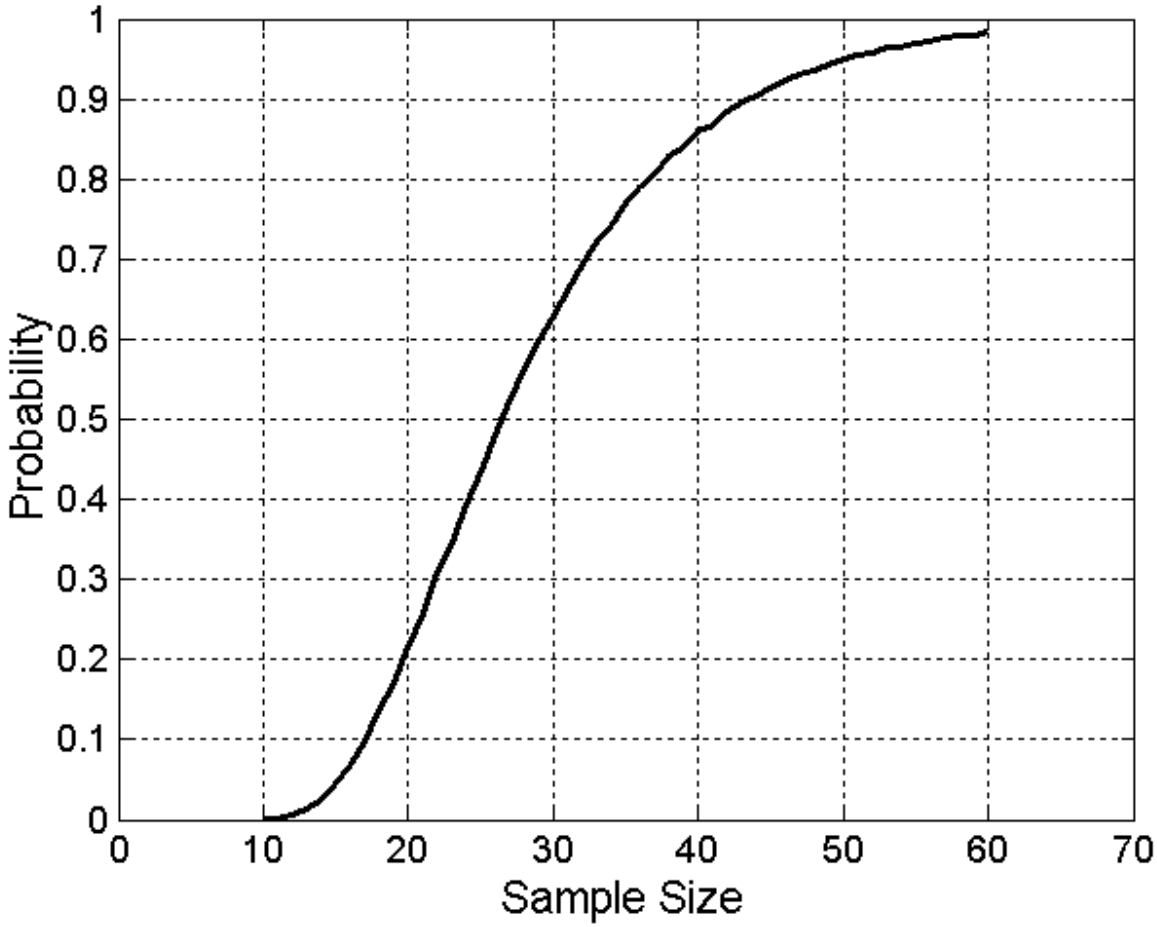
Preprocessing: campionamento

Tipologie di Campionamento

- ***campionamento casuale semplice***; *egual probabilità di selezionare ogni osservazione dell'intero insieme di dati.*
- ***campionamento senza reimbussolamento***; *ogni osservazione selezionata viene rimossa dall'intero insieme di dati, non può essere selezionata una seconda volta.*
- ***campionamento con reimbussolamento***; *ogni osservazione può essere selezionata più volte.*
- ***campionamento stratificato***; *l'intero insieme dei dati viene partizionato in sottoinsiemi disgiunti, si estraggono campioni da ogni sottoinsieme della partizione.*

Preprocessing: campionamento

Quale deve essere la dimensione di un campione affinchè si sia in grado di ottenere almeno un oggetto proveniente da ognuno di 10 gruppi?



Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

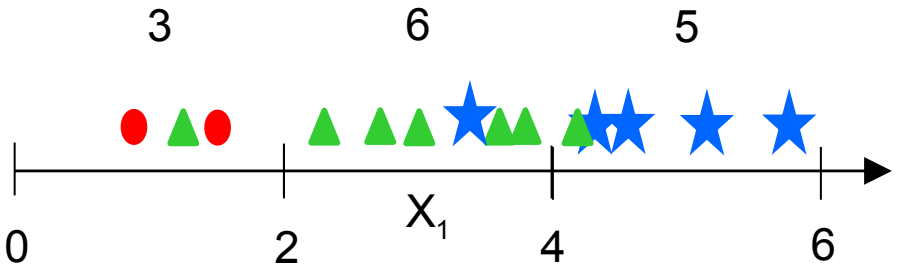
Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

Problema di classificazione con una classe che può assumere tre valori differenti rappresentati da cerchio rosso, triangolo verde e stella blu.

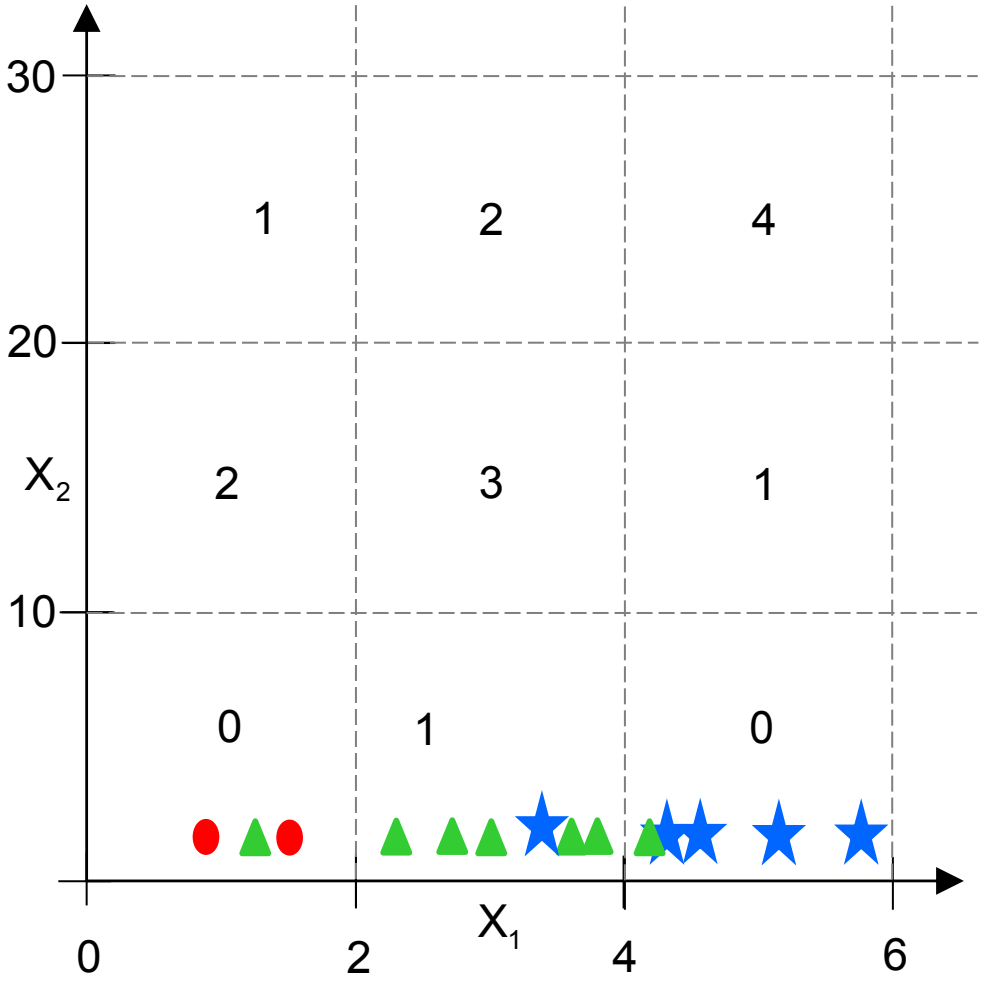


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

Considerando anche l'attributo X2 si può riscontrare una diminuzione della densità delle osservazioni, densità nelle relative celle.

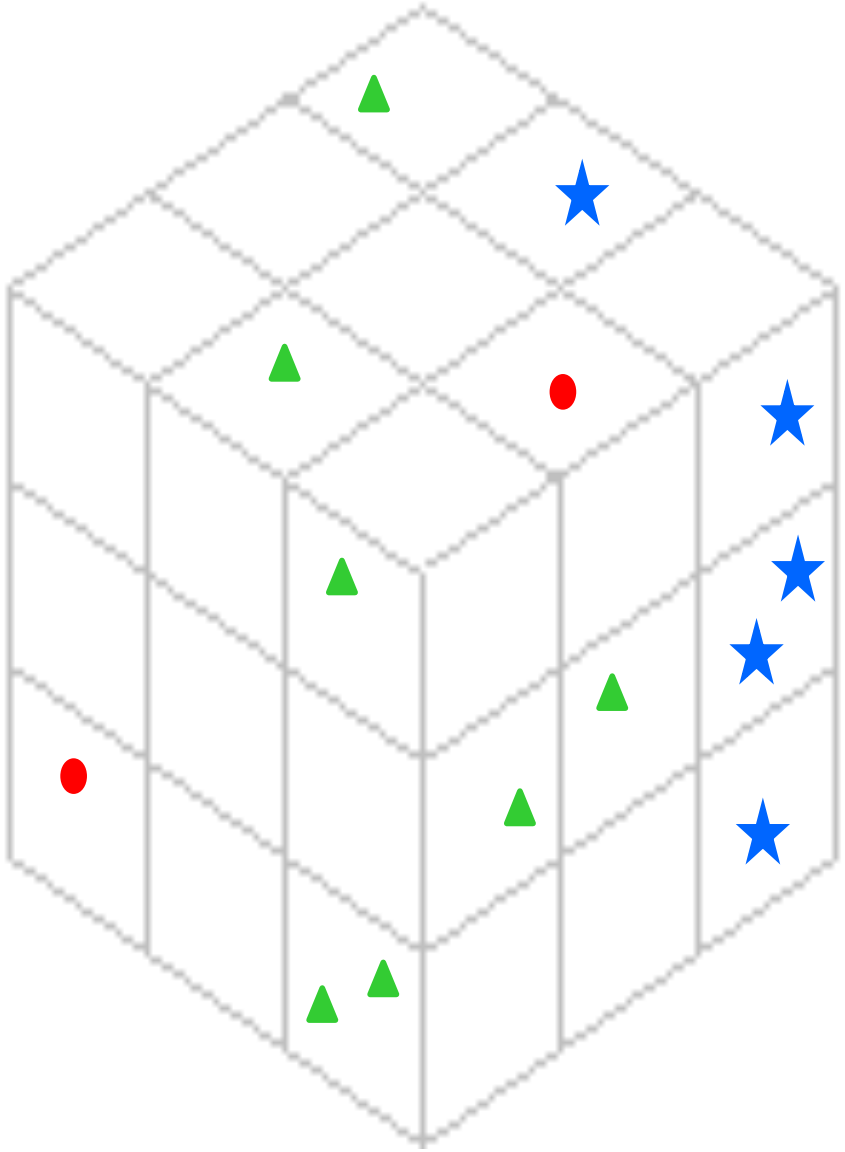


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

Considerare un ulteriore attributo X_3 non fa altro che esacerbare il problema della bassa densità delle celle risultanti dal processo di partizionamento dello spazio di input.

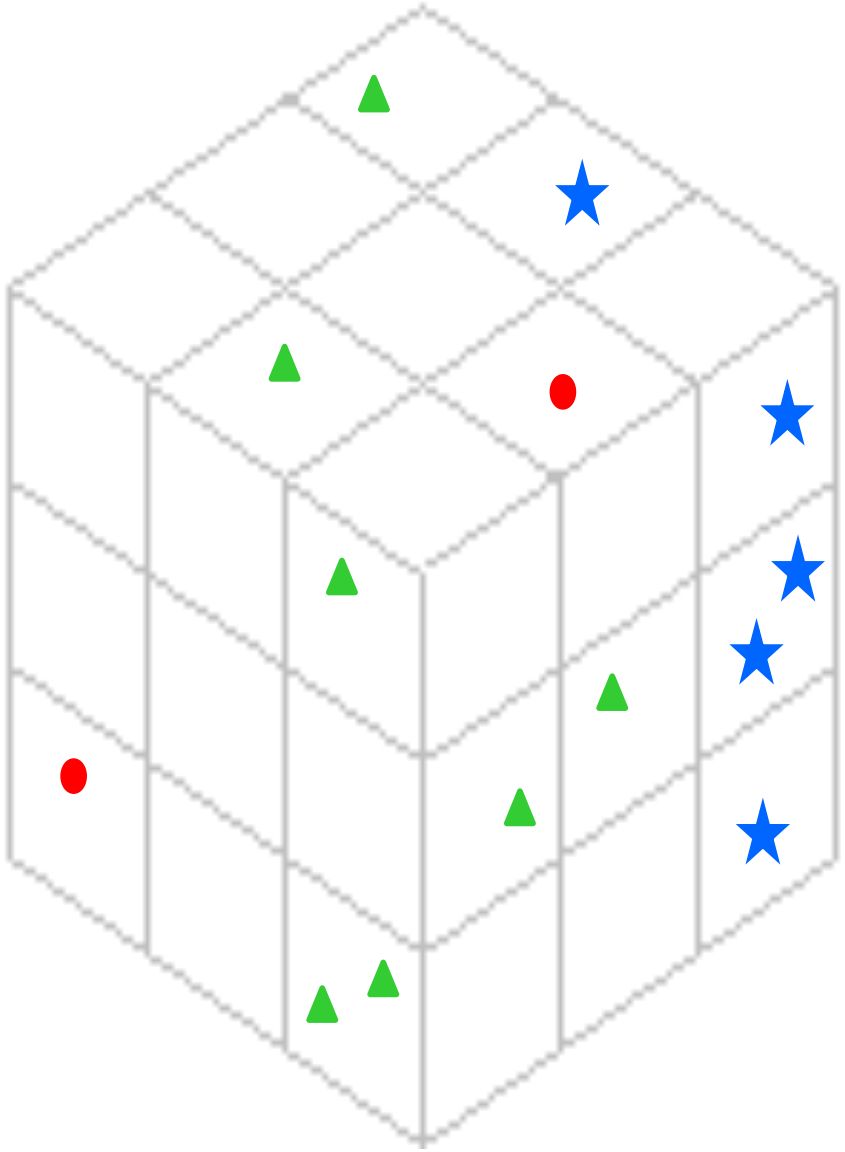


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D **3 bins**



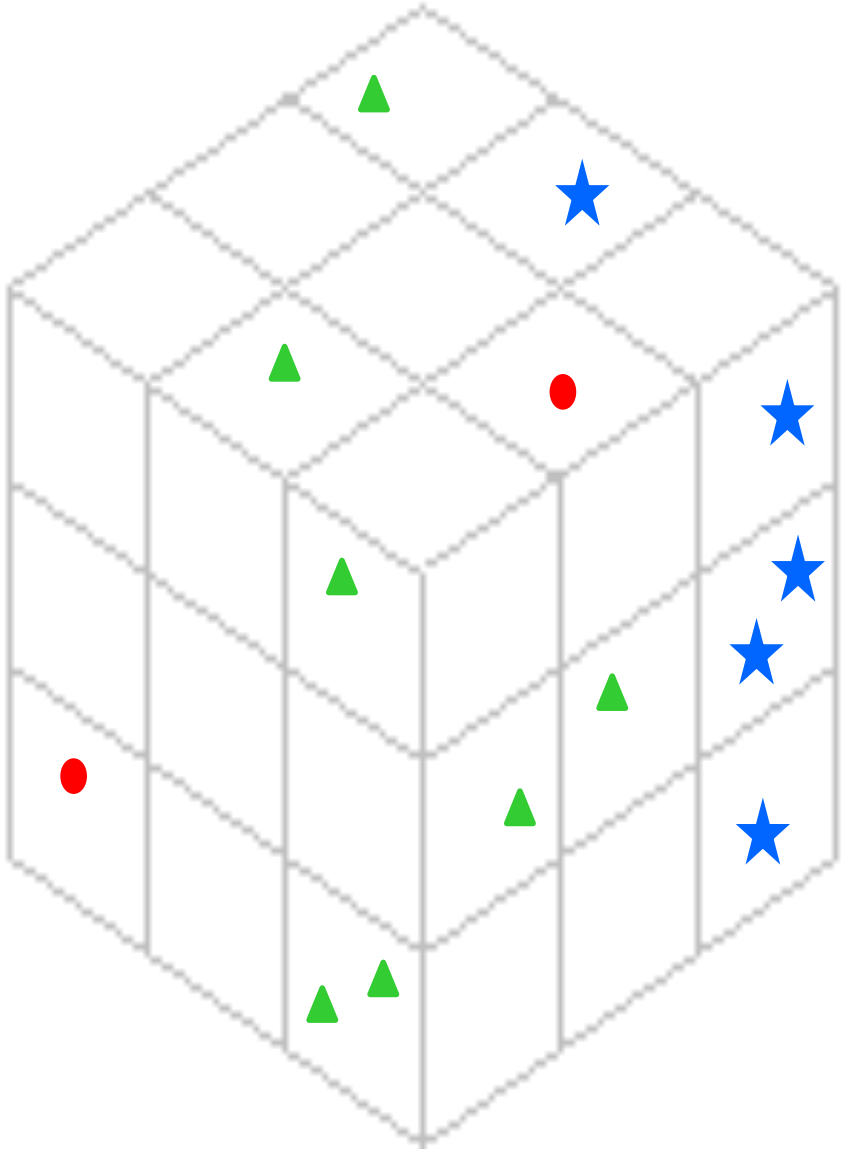
Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D **3 bins**

2D **9 bins**



Preprocessing: curse of dimensionality

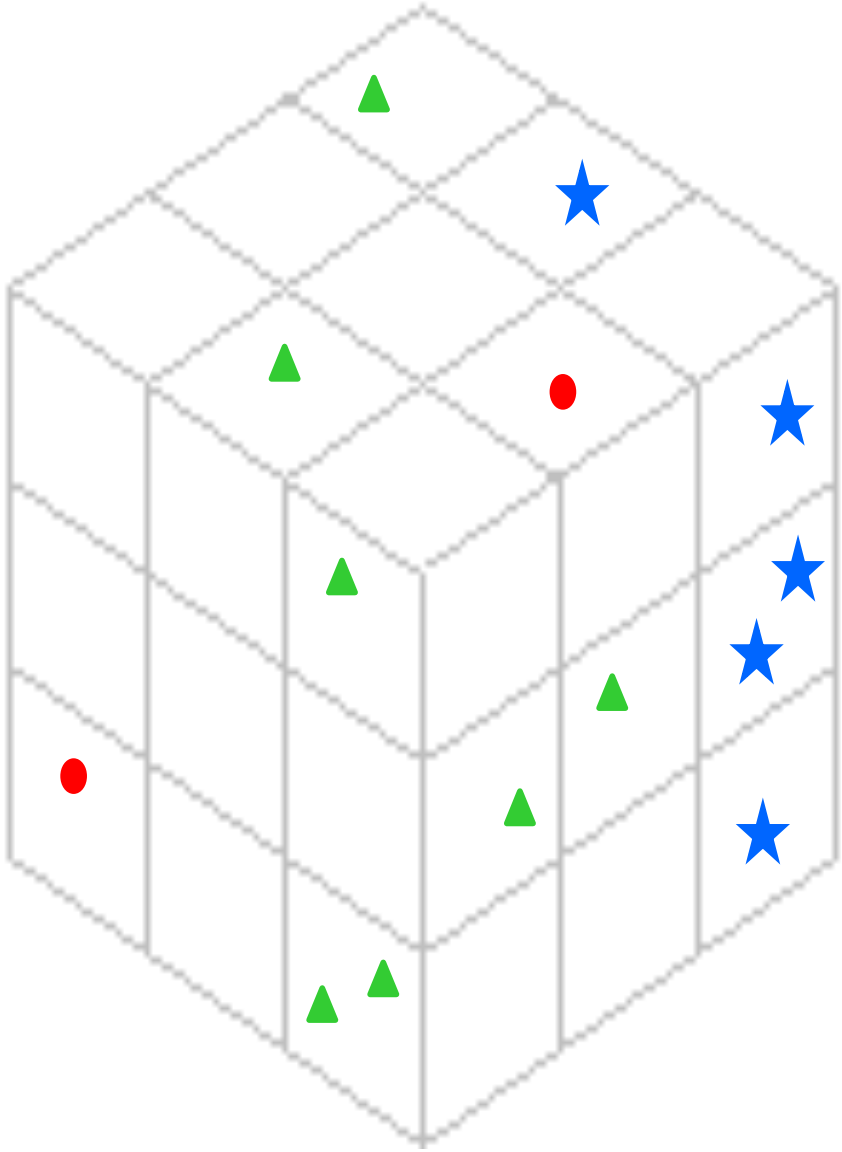
All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D **3 bins**

2D **9 bins**

3D **27 bins**

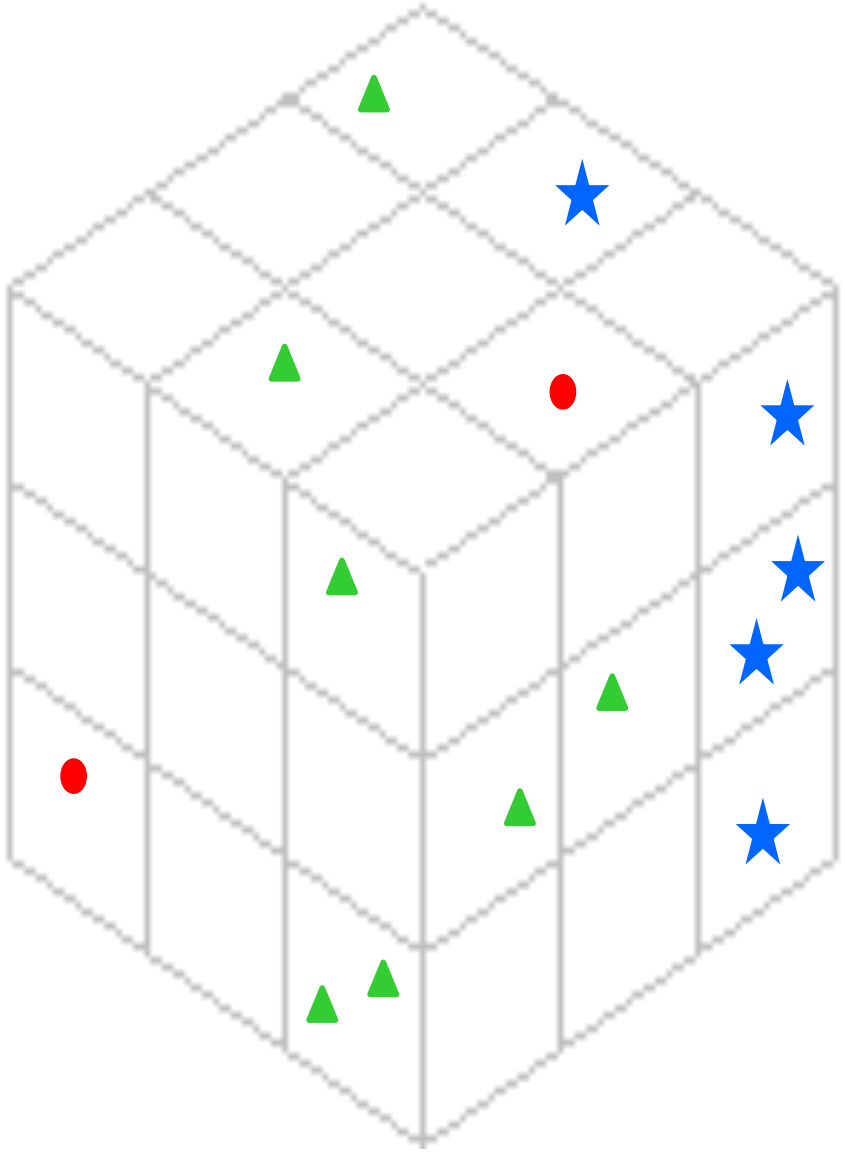


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D **3 bins** **9 osservazioni**



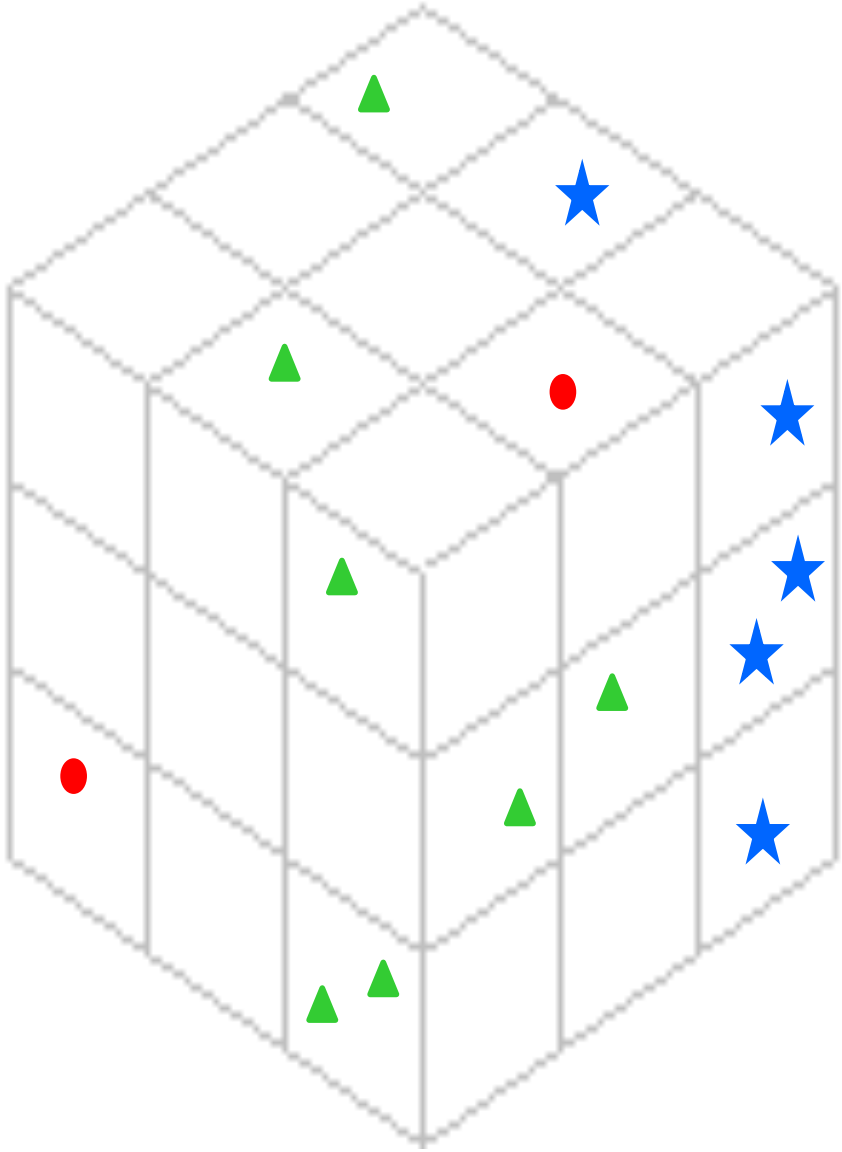
Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D **3 bins** **9 osservazioni**

2D **9 bins** **27 osservazioni**

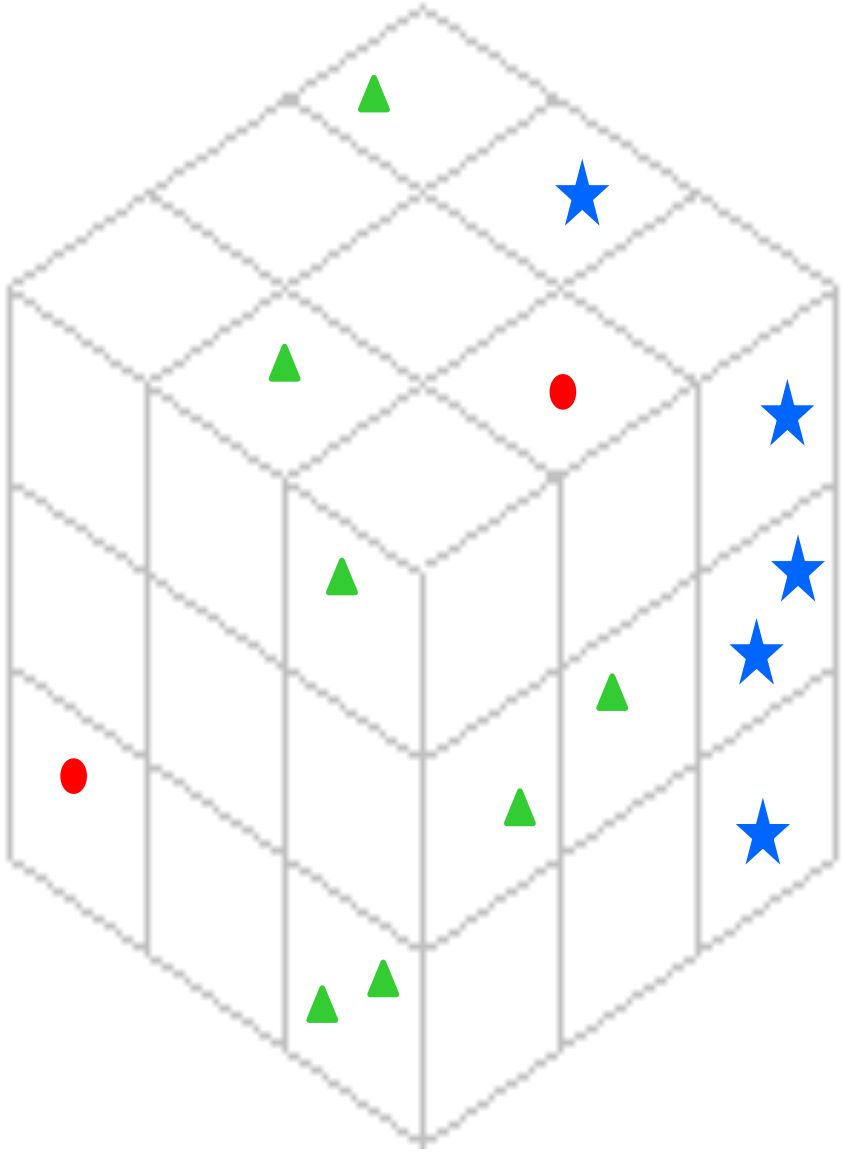


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

1D	**3 bins**	**9 osservazioni**
2D	**9 bins**	**27 osservazioni**
3D	**27 bins**	**81 osservazioni**

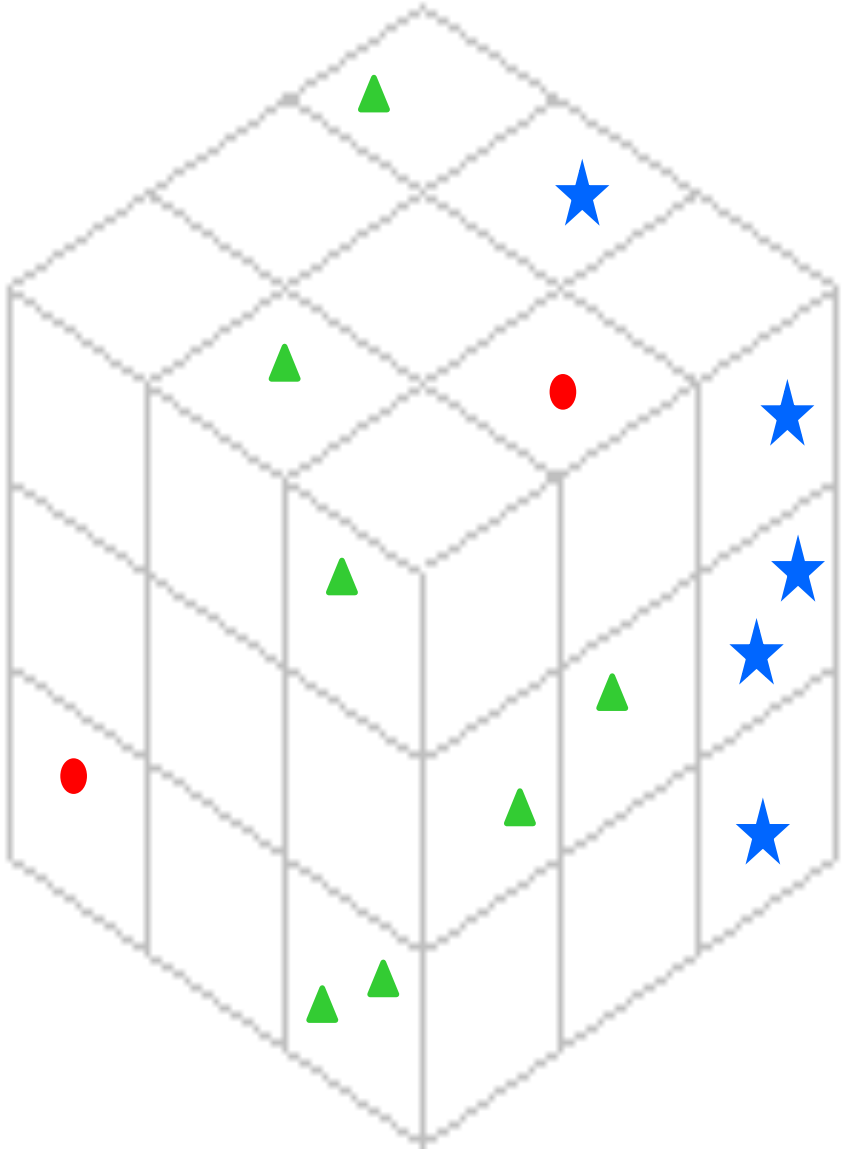


Preprocessing: curse of dimensionality

All'aumentare della dimensionalità, i dati sono sempre più sparsi se valutati in termini dello spazio di input nel quale si trovano.

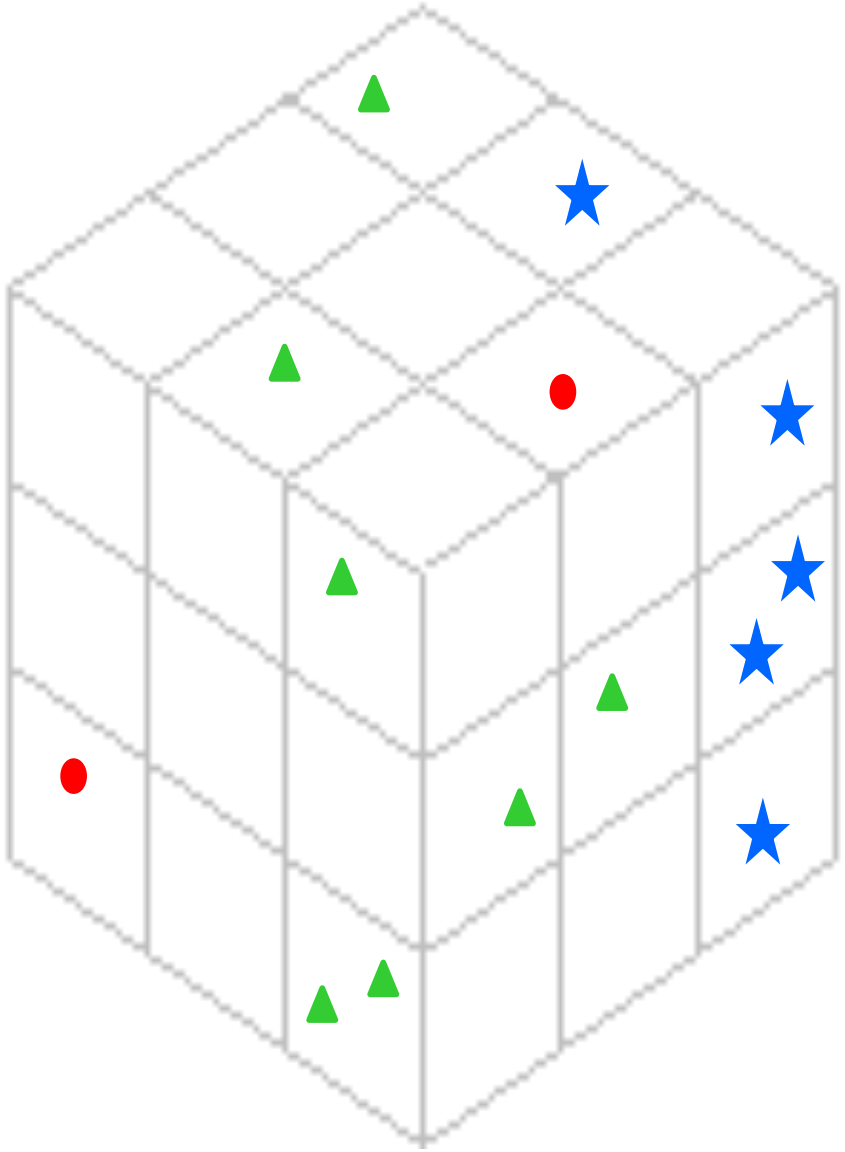
Le definizioni di densità e distanza tra punti, critiche per i task di clustering e identificazione degli outlier, perdono di significato.

Disponendo di 9 osservazioni ci si attende che nel caso vengano utilizzati 3 attributi, circa 2/3 dello spazio (celle) sia vuoto, non contiene alcuna osservazione.



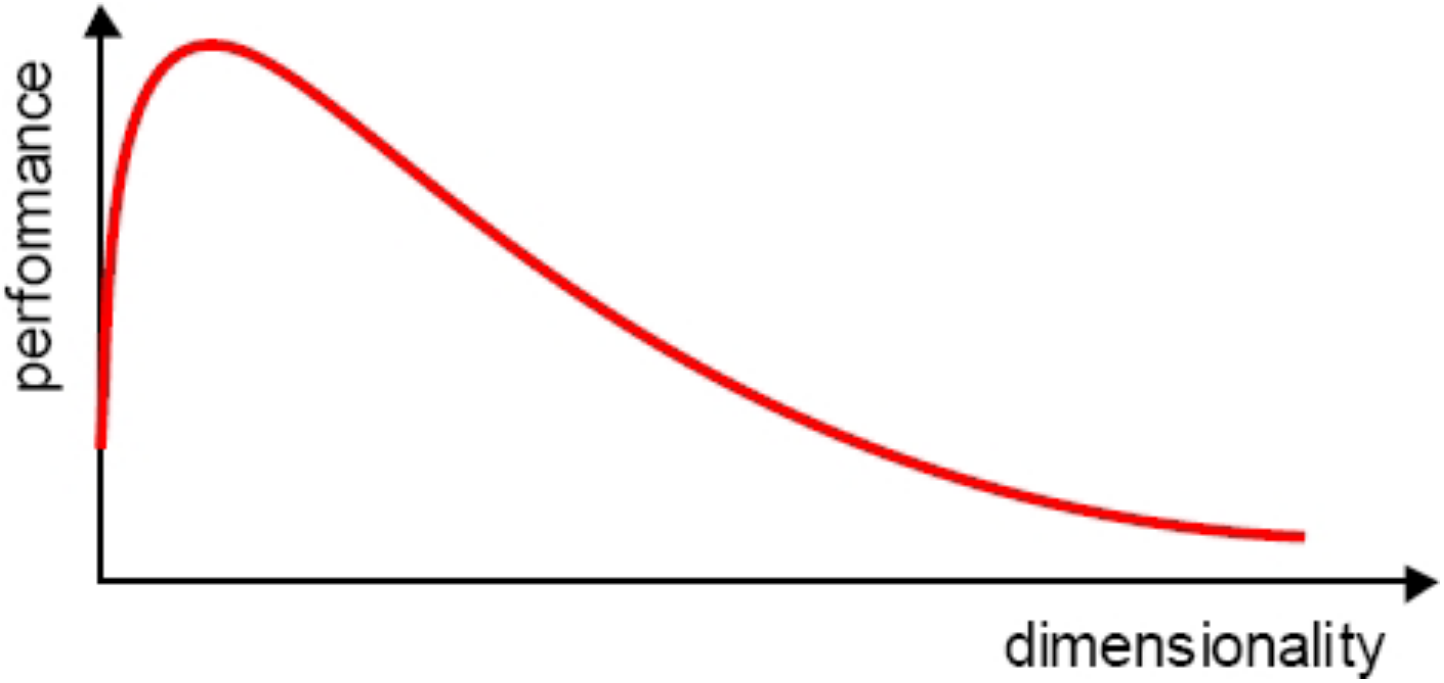
Alternative

- *Usare informazione a priori per escludere porzioni dello spazio di input*
- *Ridurre il numero di intervalli, utilizzare intervalli più ampi*
- *Ridurre il numero di attributi utilizzati per descrivere il dataset che deve essere analizzato*



Riassumendo

Il fenomeno del curse of dimensionality significa che dato un campione con dimensione fissata (numerosità del campione) esiste un numero di attributi che possono essere presi in considerazione in modo congiunto oltre al quale le prestazioni ottenibili da parte di un modello di classificazione iniziano inevitabilmente *a degradare.*



Preprocessing: riduzione dimensionalità

Obiettivi

- Evitare di incorrere nel fenomeno del "*curse of dimensionality*"
- Ridurre tempo e memoria necessarie per gli algoritmi di Data Mining
- Consentire una più agevole visualizzazione dei dati
- Identificare attributi non rilevanti o contribuire alla riduzione del livello di rumore presente nei dati

Tecniche

- *Principal Component Analysis (PCA)*
- *Singular Value Decomposition (SVD)*
- *Altre: tecniche supervisionate e non supervisionate*

Preprocessing: riduzione dimensionalità

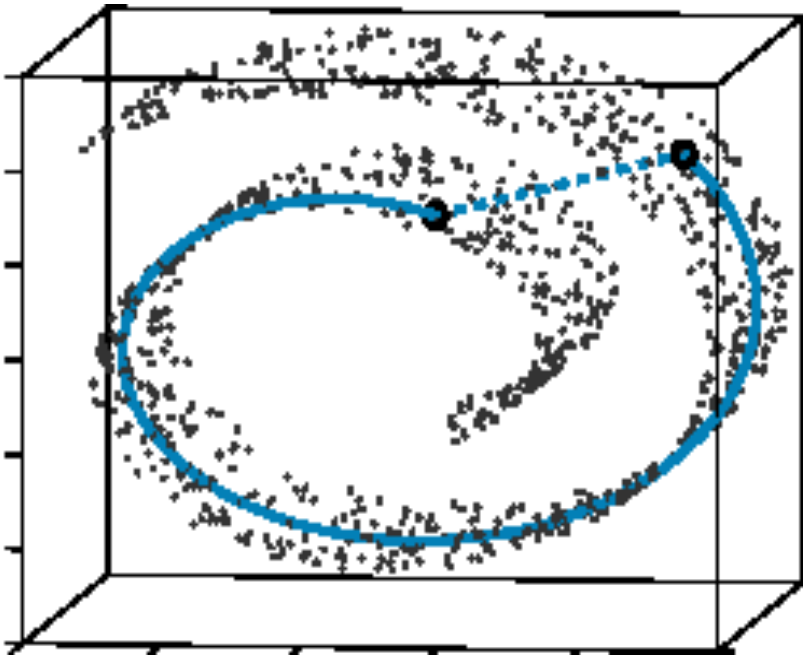
La *Principal Component Analysis* (**PCA**) ha lo scopo di identificare patterns nei dati, di esprimere i dati in modo tale da evidenziarne similarità e differenze.

Dato che i patterns nei dati possono essere particolarmente difficili da trovare, in spazi di elevata dimensionalità nei quali il lusso della rappresentazione grafica non è concesso, la PCA costituisce uno strumento molto potente per l'analisi dei dati.

Un'altro dei vantaggi primari della PCA è rappresentato dal fatto che una volta identificati i patterns nei dati, i dati vengono compressi, riducendo il numero di dimensioni, senza un'eccessiva perdita di informazione.

Un'altra possibilità è offerta dall'algoritmo denominato **ISOMAP** che prevede i seguenti passi principali:

- Costruire un grafo di vicinanza delle osservazioni del dataset
- Computazione, per ogni coppia di punti del grafo di vicinanza, del relativo cammino a costo minimo (distanza geodetica)



Preprocessing: features subset selection

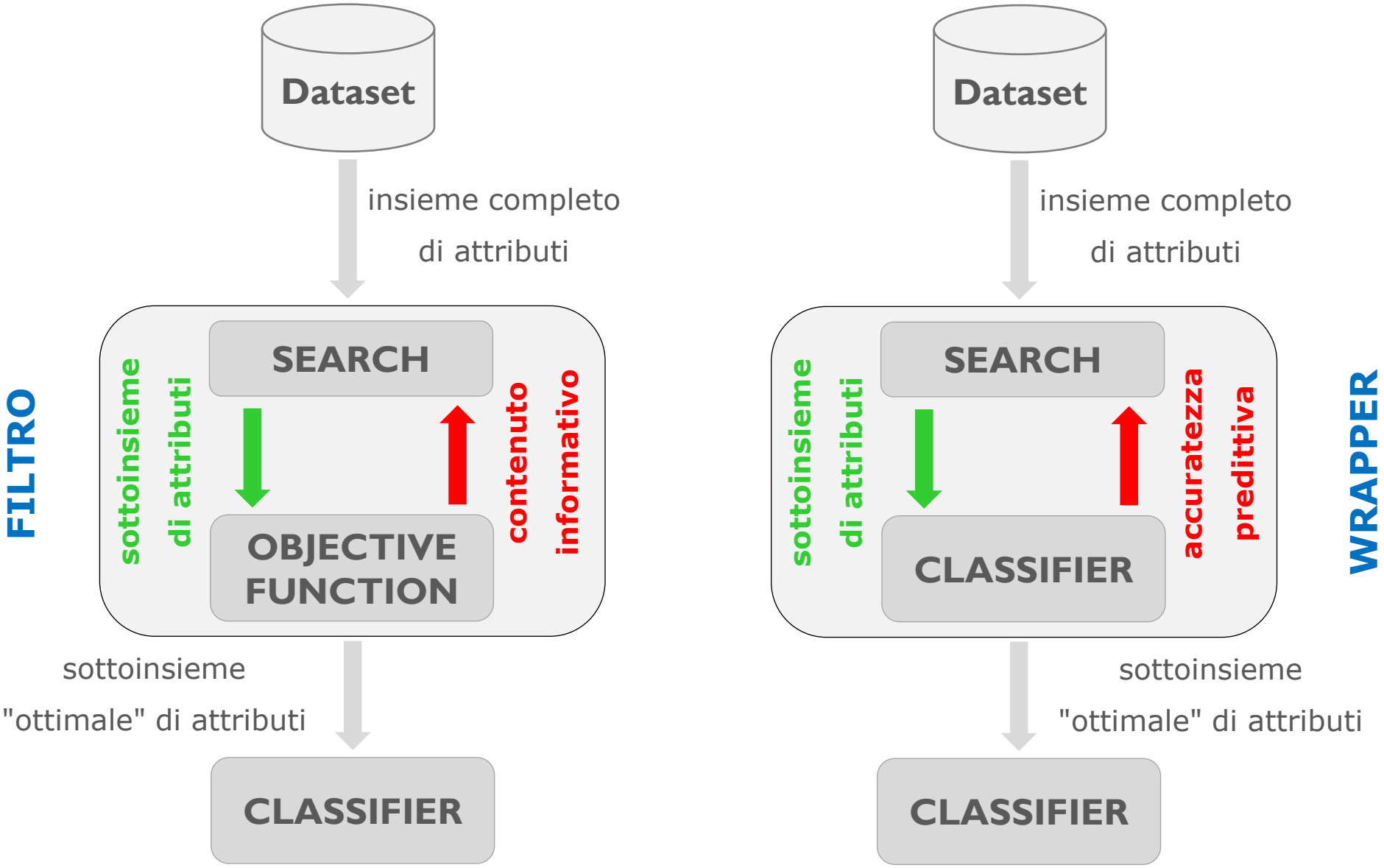
Offre un altro modo di ridurre la dimensionalità dei dati da analizzare. Gli attributi (features) che descrivono le osservazioni potrebbero essere:

- **Ridondanti**; portano informazione duplicata (ridondante) in quanto contenuta in uno o più dei restanti attributi, (prezzo di acquisto di un bene e costo della relativa tassa, …)
- **Irrilevanti**; non contengono informazione utile per il task di Data Mining che deve essere affrontato, (identificatore di un cliente, …)

Approcci disponibili

- **Brute-force**: provare tutti i possibili sottoinsiemi di attributi come input agli algoritmi di DM
- **Embedded**: la FSS si ottiene come parte dell'esecuzione dell'algoritmo di DM
- **Filtri**: gli attributi vengono selezionati prima che l'algoritmo di DM venga utilizzato
- **Wrapper**: utilizzare l'algoritmo di DM come black-box per trovare il miglior sottoinsieme di attributi

Preprocessing: features subset selection



Preprocessing: features subset selection

Principali Vantaggi

- *Riduzione del costo di acquisizione dei dati*
- *Diminuzione del tempo necessario per l'inferenza*
- *Aumentata comprensibilità*
- *Aumentata accuratezza*

Obiettivi

- *Evitare il fenomeno dell'overfitting*
- *Ottenere un modello più veloce e cost-effective*
- *Comprensione approfondita del data generating process*

Preprocessing: features subset selection

FILTRI

- **Univariati**
  - *Si definisce una misura di associazione tra l'attributo X e la classe Y, Mutual Information*
  - *Si ordinano gli attributi in base alla misura di associazione scelta*
  - *Si selezionano i primi "r" attributi dell'ordinamento*
  - *Si rilevano gli attributi rilevanti ma non si eliminano i ridondanti*
- **Multivariati**
  - *Tentativo di identificare contemporaneamente attributi rilevanti e ridondanti*
  - *Selezionare un sottoinsieme di attributi associati alla classe ma incorrelati tra loro*
  - *Misure simmetriche di incertezza o misure correlation-based*

Comparazione tra approccio Univariato e Multivariato

	Vantaggi	Svantaggi
Univariati	velocità scalabilità indipendenti dal classificatore	ignora la dipendenza tra attributi ignora interazioni con il classificatore
Multivariati	modella la dipendenza tra gli attributi indipendente dal classificatore costo computazionale favorevole rispetto a wrapper	più lento delle tecniche univariate meno scalabile delle tecniche univariate ignora interazioni con il classificatore

Univariati	Multivariati
Parametrici	Correlation Feature Selection (CFS)
t-test	Relief
ANOVA	Blanket
Informazione Mutua	
Non Parametrici	
Mann-Whitney	
Kruskal-Wallis	
Permutation test	

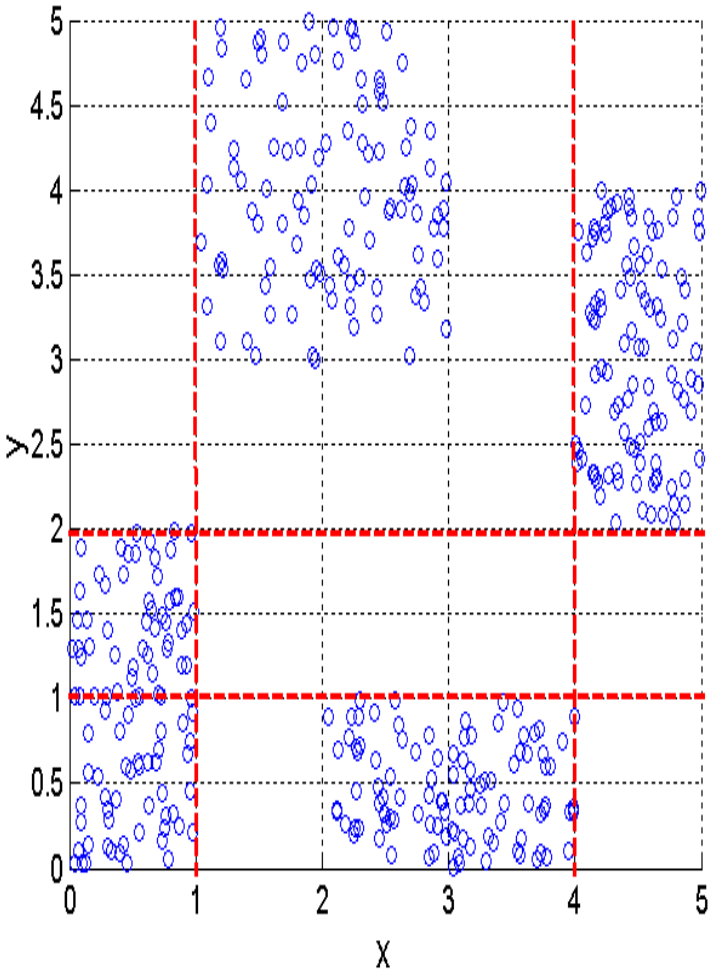
Preprocessing: features creation

Generare nuovi attributi che siano in grado di rappresentare in modo efficace informazione importante presente nel dataset, in grado di fornire informazione in modo più efficiente di quanto non siano in grado di fare gli attributi originali.

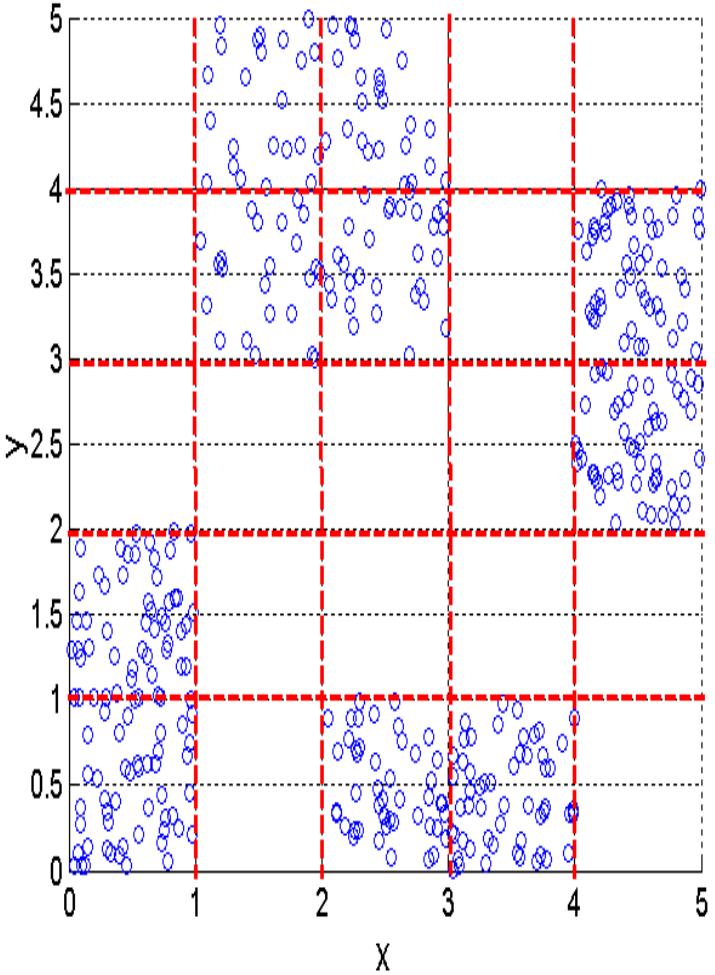
METODOLOGIE

- **Feature Extraction**; specifica del particolare dominio che viene analizzato
- **Mappare i dati in un nuovo spazio**
- **Feature Construction**; combinare attributi esistenti tra loro

Preprocessing: discretizzazione



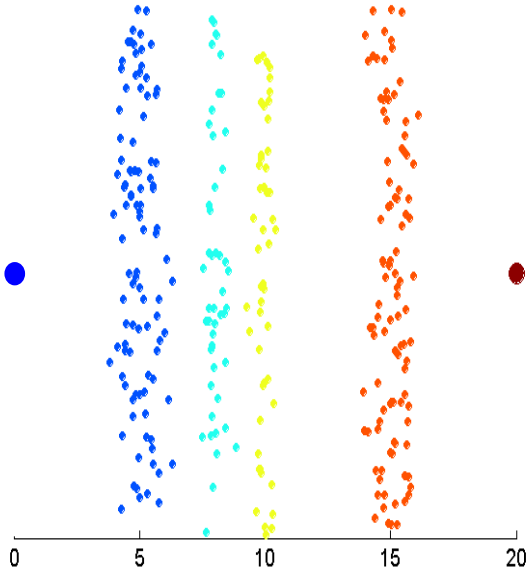
3 bin per x e y

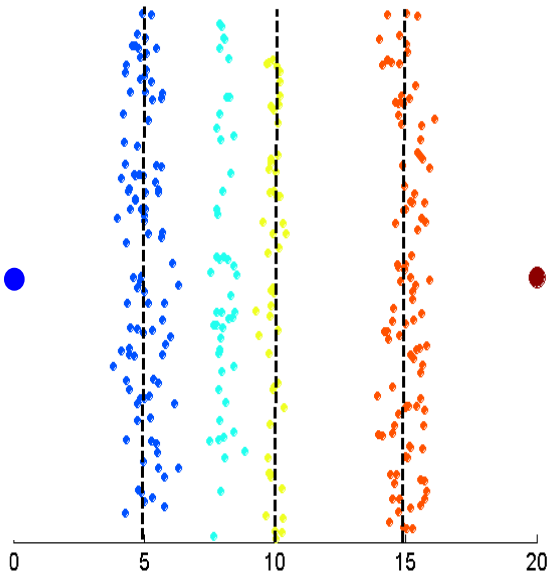


5 bin per x e y

Preprocessing: discretizzazione

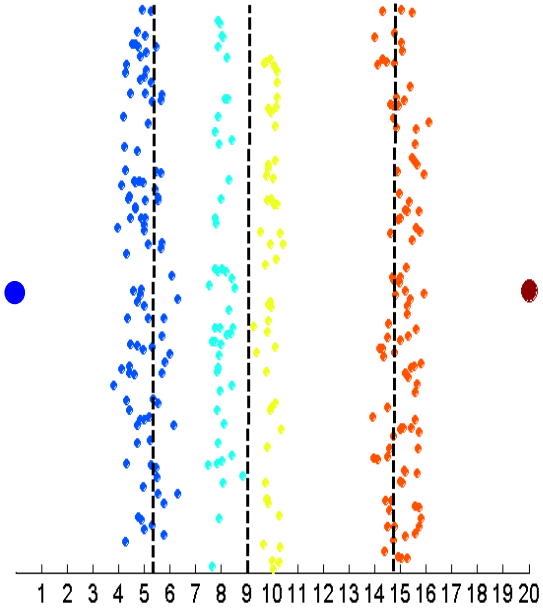
data

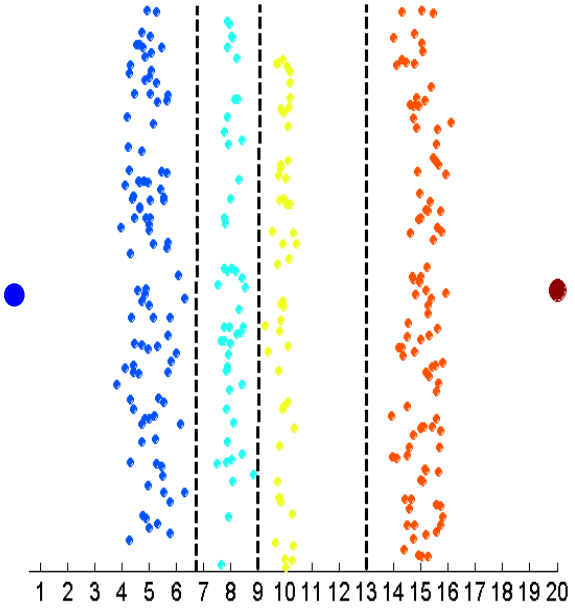




equal interval width

equal frequency





K-means

Preprocessing: binarizzazione

Attributo che possa assumere "k" valori distinti, si associa ad ogni valore del supporto un intero nell'intervallo [0, k-1].

Gusto				
categorie	interi	*X1*	*X2*	*X3*
pessimo	0	0	0	0
cattivo	1	0	0	1
discreto	2	0	1	0
buono	3	0	1	1
ottimo	4	1	0	0

Attenzione a non indurre relazioni d'ordine fittizie e/o associazioni tra gli attributi trasformati. È necessario ricorrere ad attributi binari asimmetrici (analisi di associazione).

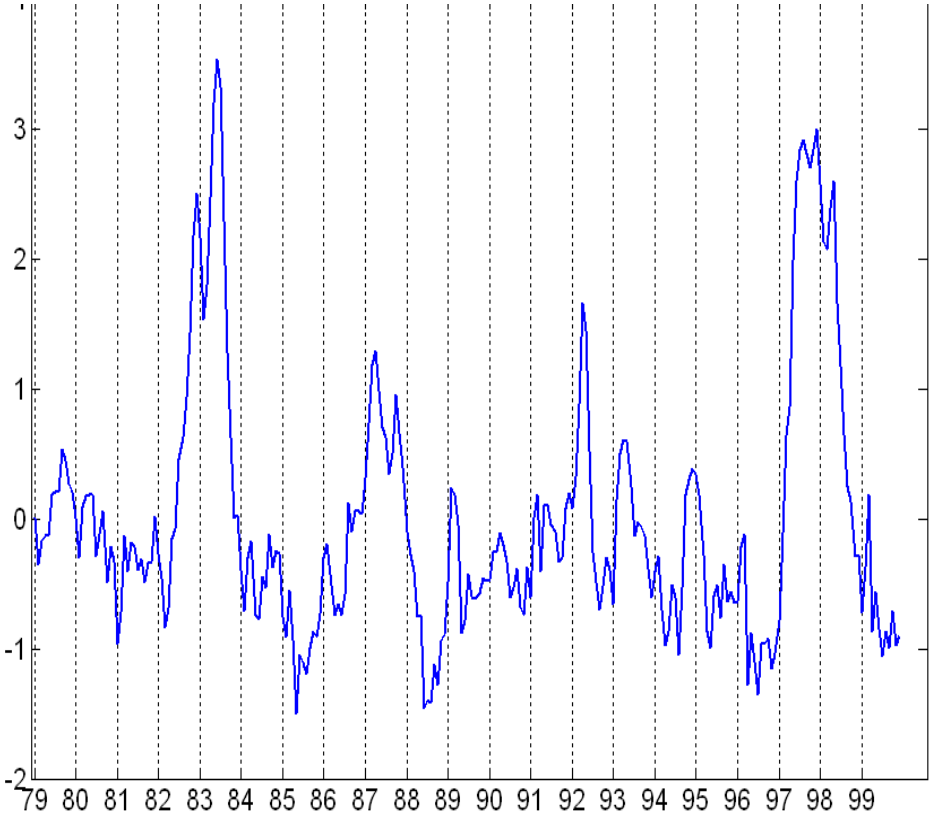
Gusto						
categorie	interi	*X1*	*X2*	*X3*	*X4*	*X5*
pessimo	0	1	0	0	0	0
cattivo	1	0	1	0	0	0
discreto	2	0	0	1	0	0
buono	3	0	0	0	1	0
ottimo	4	0	0	0	0	1

Preprocessing: trasformazione

Una funzione che mappi l'intero insieme di valori di un dato attributo su un nuovo insieme di valori in modo tale che i valori originali possano essere identificati tramite uno dei nuovi valori, come ad esempio funzioni semplici del tipo:

$$X^k \quad \log(X) \quad \exp(X) \quad |X|$$

- *Standardizzazione*
- *Normalizzazione*



Preprocessing: dati mancanti

Il problema delle *osservazioni incomplete* o dei *dati mancanti* (missing data) è molto rilevante soprattutto quando si abbia a che fare con database di elevata dimensionalità (elevato numero di attributi).

Alcune cause del fenomeno del dato mancante sono:

- *il valore dell'attributo non è sempre osservabile/misurabile*
- *l'attributo non era ritenuto rilevante per cui si è iniziato a misurarlo e a registrarne i valori solo a partire da un certo istante temporale*
- *banali incomprensioni sulle direttive di memorizzazione*
- *malfunzionamento dei dispositivi di misurazione o di memorizzazione*
- *rimozione di valori inconsistenti rispetto a valori di altri attributi*

Il problema dei *dati mancanti* deve essere preso in seria considerazione prima di avviare un qualsiasi studio di Data Mining.

Preprocessing: dati mancanti

Consideriamo una porzione di un database nel quale vengono registrati valori degli esami del sangue di pazienti ospedalizzati.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Preprocessing: dati mancanti

Consideriamo una porzione di un database nel quale vengono registrati valori degli esami del sangue di pazienti ospedalizzati.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Preprocessing: dati mancanti

I principali approcci alla trattazione dei dati mancanti sono:

- **Ignorare l'intera osservazione (record):** applicato tipicamente se il valore mancante è relativo alla variabile di classe
- **Riempimento manuale:** approccio molto time consuming
- **Utilizzo di una costante globale:** si rimpiazza il valore mancante tramite un valore dummy tipo "Sconosciuto", "999", ...
- **Mean replacement:** si rimpiazza il valore mancante con la media dell'attributo calcolata sulle osservazioni non mancanti
- **Conditional mean replacement:** si rimpiazza il valore mancante con la media dell'attributo calcolata sulle osservazioni non mancanti e solo per quelle osservazioni che appartengono alla stessa classe.
- **Most probable replacement:** si determina il valore più probabile tramite regressione o approcci Bayesiani o alberi di decisione.

Preprocessing: ignorare il record

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

La porzione di database visualizzata consiste di 14 osservazioni, casi o record.

Dopo aver ignorato i record con dati mancanti rimangono 8 osservazioni, 6/14 = 0.43.

Preprocessing: riempimento manuale

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Estremamente pesante dal punto di vista del tempo necessario per completare un task in modo affidabile.

Preprocessing: costante globale

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Si sceglie una variabile globale, eventualmente dipendente dall'attributo, con la quale vengono riempite le celle associate a dati mancanti.

Preprocessing: costante globale

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0	999	16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	999
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio	999	91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11	999	9	63	66
193649	28-Sep-47	maschio	174	999	15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48	sconosciuto	170	61	13-May-11	16.1	13	999	189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Si sceglie una variabile globale, eventualmente dipendente dall'attributo, con la quale vengono riempite le celle associate a dati mancanti.

Preprocessing: mean replacement

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11	**12**	9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Consideriamo l'attributo "*acido folico*", il record con **ID = 99372** ha valore mancante che tramite mean replacement puo' essere ricavato calcolando la media dei valori di "*acido folico*" che è pari a **12**, valore che viene usato per riempire il dato mancante.

Preprocessing: conditional mean repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Supponiamo che la variabile di classe sia "*genere*", lo schema conditional mean replacement computa la media condizionata rispetto ai vari valori della variabile di classe.

Preprocessing: conditional mean repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0	**25.7**	16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

La media della variabile "*ALT*" per "*genere = maschio*" è uguale a 25.7. Questo valore viene usato per riempire il dato mancante per il record con **ID=122345**.

Preprocessing: most probable repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Il record con **ID=113904** ha valore mancante per l'attributo "*altezza*".

Si desidera sfruttare la relazione tra l'attributo "*altezza*" e l'attributo "*peso*" dipendentemente dal valore dell'attributo "*genere*".

Preprocessing: most probable repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Si costruisce un modello di *regressione lineare semplice* del tipo

$$altezza = a_0 + a_1 * peso$$

Preprocessing: most probable repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Si costruisce un modello di *regressione lineare semplice* del tipo

$$altezza = 158 + 0.222 * peso$$

Preprocessing: most probable repl.

ID	data di nascita	genere	altezza	peso	data prelievo	acido folico	ALT	ferritina	eosinofili
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio	**178.5**	91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Si costruisce un modello di *regressione lineare semplice* del tipo

$$\mathbf{178.5} = 158 + 0.222 * \mathit{91}$$

Preprocessing: dati mancanti

I metodi:

- *Utilizzo di una costante globale*
- *Mean replacement*
- *Conditional mean replacement*
- *Most probable replacement*

introducono un bias (distorsione) nei dati.

I valori imputati in questi modi possono non essere corretti.

Comunque il metodo ***Most Probable Replacement*** viene molto spesso utilizzato.

Utilizza l'informazione sui dati disponibili per prevedere valori mancanti.