
Dispensa del corso di Machine Learning

Professor Fabio A. Stella (primo semestre, 6 cfu)

Indice

1	Dati	3
1.1	Tipologie di Dati	3
1.2	<i>Data Exploration</i>	3
1.3	<i>Missing Values</i>	5
1.4	<i>Pre-Processing</i>	6
2	Classificazione	9
2.1	Introduzione	9
2.2	<i>Performance Evaluation</i>	10
2.2.1	<i>Accuracy and fitting</i>	10
2.2.2	Performance measures	11
2.2.3	Estimate Accuracy	12
2.2.4	<i>Comparing Classifiers</i>	13
2.3	<i>Class imbalance problem</i>	16
2.3.1	Counting the cost	18
2.4	<i>Feature Selection</i>	22
2.4.1	<i>Filter e Wrapper</i>	22
3	Clustering	24
3.1	Introduzione	24
3.1.1	<i>Well-Separated Cluster</i>	26
3.1.2	<i>Prototype-Based Cluseter</i>	26
3.1.3	<i>Density-based Cluster</i>	27
3.1.4	<i>Graph-based Cluster</i>	27
3.1.5	<i>The clusetring "Cicle"</i>	28
3.2	<i>Proximity</i>	28
3.2.1	<i>Proximity Measures</i>	30
3.3	<i>Clustering Evaluation</i>	33
3.3.1	<i>External/Supervised measures</i>	33
3.3.2	<i>Internal/Unsupervised measures</i>	34
3.3.3	<i>Relative measures</i> e il Problema Fondamentale del <i>Clustering</i>	37
3.3.4	Validity Paradigm	38
4	Analisi e Regole di Associazione	39
4.1	<i>Rules Evaluation</i>	41
4.1.1	Criteri soggettivi	41
4.1.2	Criteri statistici	41

Dati

1.1 Tipologie di Dati

Esistono diverse tipologie di dati, la cui principale differenziazione viene riassunta nella tabella sottostante.

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

Un'altra possibile distinzione è tra:

- **Attributi Discreti:** sono caratterizzati da una cardinalità finita, o infinita ma numerabile. Fanno parte di tale categoria gli attributi categorici, numerici (discreti) e binari.
- **Attributi continui:** sono caratterizzati da valori appartenenti ai numeri reali ($X \in R$). Tipicamente ne fanno parte misure come la temperatura, il peso o l'età.

1.2 Data Exploration

Il processo di **Esplorazione dei Dati** consiste nell'insieme di metodi, tecniche ed analisi volta all'individuazione di caratteristiche strutturali dei dati a disposizione.

Una primo tipo di analisi può essere fornita dallo studio delle **statistiche descrittive**, ovvero i principali indicatori riguardanti le distribuzioni dei dati; tra le

più utilizzate vi sono la media, la mediana, la moda e la deviazione standard. Nonostante tale studio possa essere applicato all'intero set di dati, non tutte le statistiche possono essere calcolate su ogni tipologia di attributo; un esempio riguarda le variabili nominali, per le quali non possono essere calcolati (o comunque non ha senso calcolare) media e mediana, mentre la moda (*most frequent value*) può risultare un indicatore molto utile.

Per ciò che riguarda invece gli attributi numerici, esistono diversi indicatori utili:

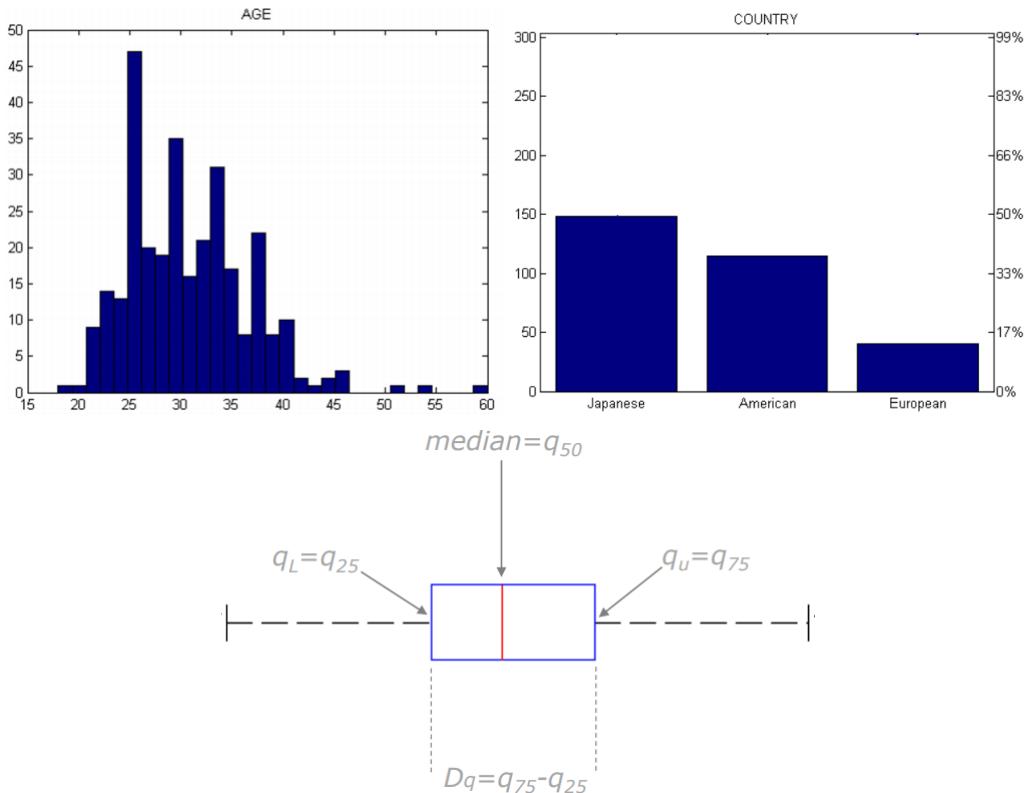
1. **Media** → $\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$ con N numerosità del campione/popolazione, x_i valore dell'attributo X per l' i -esima osservazione ($i = 1, \dots, N$). Un'alternativa, meno sensibile ai valori anomali, risulta la **Media Troncata**, per quale vengono esclusi dal calcolo il valore maggiore e minore.
2. **Quantili** → Rappresentano valori che dividono la distribuzione in un punto determinato; vengono individuati riordinando in maniera crescente la distribuzione di X , decidendo la quota che si vuole lasciare a sinistra del quantile (ordine del quantile) e individuando tale valore. Alcuni casi particolari riguardano i **Quartili** della distribuzione, i quali dividono la distribuzione lasciando a sinistra il 25% (**Primo**), il 50% (**Mediana**) o il 75% (**Terzo**). La mediana viene spesso impiegata al posto della media, come indicatore centrale della distribuzione, poiché meno sensibile ai valori anomali (*outliers*).
3. Un'altra famiglia di indicatori è quella riferita agli **Indici di Dispersione** (*Spread Indices*), ovvero misure che danno un'idea della dispersione dei dati nel piano. Un primo esempio è dato semplicemente dal **Range**, ovvero il campo di variazione calcolato come $Range(X) = X_{max} - X_{min}$; essendo tuttavia che spesso si può avere una grossa concentrazione di valori entro una certa fascia, il *range* può portare a conclusioni ingannevoli. Un indice più utilizzato è la **Varianza**, ovvero il momento centrale di secondo ordine: $Var(X) = \sigma_X^2 = \frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N}$ o in alternativa la sua radice (**Deviazione Standard**): $SD(X) = \sqrt{\sigma_X^2} = \sigma_X$. Entrambe sono sempre non negative ($\sigma^2, \sigma \in [0, +\infty)$), e indicano una distribuzione con poca dispersione intorno alla media in caso di valori vicini a 0. Esattamente come la media, anche la varianza risulta influenzata dagli *outliers*; per tale ragione spesso si utilizzano delle varianti più robuste:

- **Deviazione Media Assoluta** $AAD(X) = \frac{\sum_{i=1}^N |x_i - \bar{X}|}{N}$
- **Deviazione Mediana assoluta** ($MAD(X) = med(|x_1 - \bar{X}|, \dots, |x_{N-1} - \bar{X}|, |x_N - \bar{X}|)$, con $x_1 < \dots < x_{N-1} < x_N$)
- **Differenza interquartile** (IQR): $x_{0.75} - x_{0.25}$ (con $x_{0.75}, x_{0.25}$ terzo e primo quartile.)

In caso di attributi multipli, bisogna modificare l'approccio e studiare la dispersione delle due variabili contemporaneamente; solitamente si utilizza la **covarianza**: $cov(X, Y) = \sigma_{XY} = \frac{\sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})}{N - 1}$.

Se infine si vuole studiare la relazione associativa tra due variabili, viene impiegata il **Coefficiente di Correlazione Lineare** (solitamente di Pearson):
 $cor(X, Y) = \rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_X \cdot \sigma_Y}}$

Una seconda tipologia di analisi esplorative riguarda i metodi di *Data Visualization*, come **Istogrammi** (variabili continue), **Bar Chart** (variabili categoriche), **Box Plot** (variabili continue).



1.3 Missing Values

Con **Valori Mancanti** si intendono tutti quei dati per i quali, per casualità o per determinate ragioni, non si hanno informazioni. Esistono diverse motivazioni per il quale un dato può risultare assente: il valore potrebbe essere non osservabile, considerato irrilevante, non registrato per errore, etc. Esistono diverse soluzioni:

- **Eliminazione dell'osservazione:** solitamente si applica quando determinate osservazioni presentano numerosi dati mancanti, e la numerosità del campione consente un'eliminazione.
- **Eliminazione della variabile:** anche in questo caso può essere applicata in caso un determinato attributo presenti numerose osservazioni mancanti e non venga ritenuto necessario all'analisi.
- **Imputazione del valore mancante:** consiste nel sostituire il dato mancante con una costante, oppure con un valore rappresentativo dell'attributo, come la moda (variabili nominali), oppure media (media condizionata) e mediana (mediana condizionata) (per variabili numeriche). Esistono inoltre metodi più complessi, basati su modelli statistici o probabilistici.

1.4 *Pre-Processing*

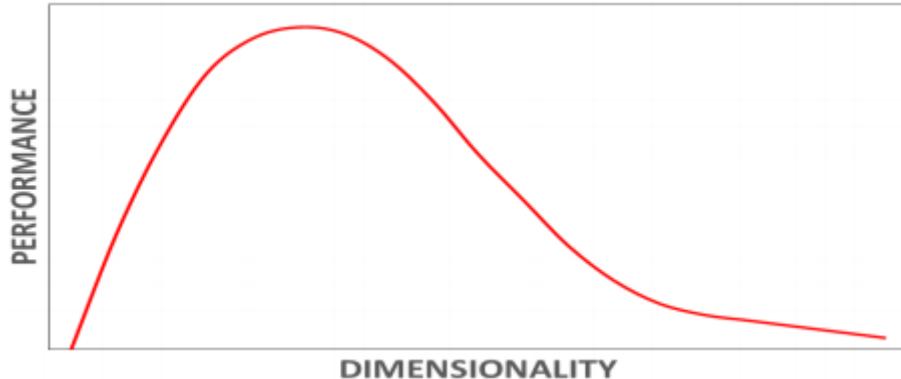
Durante uno studio di *Data Mining*, spesso risulta necessario svolgere alcune operazioni volte a preparare i dati per l'analisi: tale fase viene denominata **Pre-Processing**.

Quando si ha a che fare con set di dati contenenti un numero elevato di osservazioni, risulta molto utile al fine dell'analisi applicare un **campionamento** (*Sampling*) per estrarre una porzione di dati. Il principale scopo di tale procedimento è il risparmio in termini di onere computazionale (tempo e memoria), in particolare quando vengono applicati gli algoritmi di *machine-learning* più potenti (i quali sono anche i più onerosi computazionalmente). Esistono diversi metodi di campionamento, tuttavia tutti hanno uno scopo comune: ridurre la dimensionalità mantenendo la rappresentatività del campione (rispetto alla popolazione). I passaggi per mettere in atto un campionamento sono due, ovvero la definizione della numerosità e della tecnica; tra i metodi più diffusi troviamo:

- **Campionamento Casuale Semplice:** ogni osservazione della popolazione viene estratta (con reinserimento o senza) con la stessa probabilità. In presenza di attributi categoriali non risulta ottimale.
- **Campionamento Stratificato:** utile in presenza di variabili categoriali, poiché consente un'estrazione casuale, la quale mantiene tuttavia la distribuzione di frequenza relativa della variabile stratificante.

Un secondo obiettivo del *pre-processing* riguarda senza dubbio la **riduzione della dimensionalità**, la quale consiste nella selezione delle variabili utili (*Feature Selection*). Vi sono diverse motivazioni che giustifichino il ricorso ad una riduzione degli attributi, tra le quali l'aumento della capacità operativa degli algoritmi, la maggiore interpretabilità dei risultati (e dell'influenza di ciascun attributo), una maggiore possibilità di rappresentare il fenomeno mediante grafici e infine una diminuzione del costo computazionale (memoria/tempo). Un concetto che riassume tali punti è la cosiddetta *Curse of dimensionality*: l'aggiunta di attributi consente

di migliorare la *performance* del modello, ma solo fino a che non vi sia un eccedenza di variabili



Tra le principali tecniche per la riduzione della dimensionalità troviamo l'**Analisi delle Componenti Principali** (PCA); è una tecnica basata sull'utilizzo dell'algebra lineare al fine di proiettare i dati in uno spazio con dimensionalità minore. Si compongono come combinazioni lineari degli attributi originali, sono tra loro ortogonali e catturano la massima variazione nei dati. Solitamente è necessario specificare il numero di PCA da o la percentuale di variazione che si vogliono mantenere. Strettamente legata alla PCA troviamo la **Scomposizione a valori singolari** (SVD).

Quando si ha a che fare con variabili categoriali, le quali assumano più di due valori, può risultare conveniente applicare una **Binarizzazione** dell'attributo, come mostra la tabella sottostante.

Taste	Integer Value	X_1	X_2	X_3	X_4	X_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
ok	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Quando si ha invece a che fare con variabili numeriche continue, si può applicare la **Discretizzazione** per riportare l'attributo ad una variabile discreta (per esempio età → classe d'età); solitamente si definisce un intervallo, basato sulle frequenze assolute/relative, al fine di ottenere classi di ampiezza simile (o uguale se possibile). La discretizzazione può essere:

- Non supervisionata: si divide la variabile solamente sulla base delle informazioni contenute nella variabile stessa.
- Supervisionata: si divide la variabile sulla base di una determinata "misura di purezza" delle classi risultanti, la quale tipicamente deve risultare massimizzata quando si è in situazione di divisione ottimale. Solitamente si utilizza l'**Entropia** come misura di purezza, la quale viene definita come segue:

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

con e_i entropia dell'intervallo i -esimo, K il numero di classi della variabile di classe e p_{ki} la probabilità della classe k -esima associata all'intervallo i -esimo. Nel caso l'intervallo i contenga solo osservazioni della classe k allora $e_i = 0$, e si è in condizione di massima purezza. Lo scopo della discretizzazione supervisionata basata sull'entropia ha quindi come scopo la minimizzazione della stessa (con la conseguente massimizzazione della purezza). L'entropia totale si calcola come $E = \sum_{i=1}^n w_i \cdot e_i$, con n il numero d'intervalli, mentre $w_i = \frac{m_i}{m}$ rappresenta un peso (m_i numerosità dell'intervallo i -esimo, m numerosità totale).

La discretizzazione può essere anche applicata alle variabili qualitative ordinali e nominali; per le prime il procedimento risulta identico a quello per le variabili continue, mentre per le nominali servono altri tipi di procedimenti.

L'ultima componente del *pre-processing* da analizzare riguarda le **Trasformazioni di variabili**; ne esistono di due tipi:

- Funzioni semplici: vengono applicate delle semplici funzioni matematiche alla variabile intera, come logaritmi, radici o funzioni trigonometriche.
- Normalizzazione/Standardizzazione: si utilizzano per riportare tutte le variabili quantitative ad un *range* di variazione comune. Data una variabile quantitativa X , con media μ_X e deviazione standard σ_X , una delle standardizzazioni più usuali è data da $X \rightarrow Z$, con

$$Z = \frac{X - \mu_X}{\sigma_X}$$

la quale viene definita *Z-normalization* (con $Z \sim N(0, 1)$). Un'alternativa che riduce il *bias* in caso di valori anomali è sostituire la media con la mediana e/o la deviazione standard con la deviazione standard assoluta (o con la deviazione assoluta media)

Classificazione

2.1 Introduzione

Con classificazione si intende una particolare forma di *Supervised Learning*, ovvero un apprendimento guidato da conoscenza a priori di una determinata variabile di classe (*target*). La variabile target/output rappresenta il fenomeno che si ha intenzione di studiare, mentre le tecniche di classificazione rappresentano lo strumento per predire il suo valore a partire da una serie di variabili esplicative/input (x_1, \dots, x_k). Il tipo di classificazione dipende dal tipo di variabile target cui si fa riferimento, e può essere sia binaria (target dicotomica), sia multivariata (target *multiclass*). Il problema di classificazione viene definito *classification problem/task*.

Un sistema di classificazione si serve di due tipologie di strumenti:

- *Descriptive Modelling*: serve (a partire dalle variabili esplicative) a distinguere le osservazioni nelle varie classi
- *Predictive Modelling*: utilizzato per prevedere il valore della classe per un set di osservazioni per le quali risulti ignota (può essere considerata una *Black-Box*)

Una tecnica di classificazione, o classificatore (*classifier*) definisce un approccio sistematico volto alla costruzione di modelli di classificazione partendo da un set di dati; un modello di classificazione si serve di due *dataset*:

- Il primo viene definito **Training-Set**; il training consiste in un insieme di osservazioni per il quale sono note sia le variabili esplicative, sia la variabile *target*. Serve a istruire, in termini di apprendimento, il modello di classificazione.
- Il secondo è il **Test-Set**, ovvero un set per il quale sono note le variabili input, ma non il valore della *target* (o perlomeno si assume sia ignota). Rappresenta la porzione di dati su cui si applicano le previsioni del classificatore istruito sul *training*.

Dato un set di dati D (m osserv.), un *training-set* D_t (t osserv.) e un *test-set* D_{ts} (v osserv.) si avrà:

$$D = D_t \bigcup D_{ts}; D_t \bigcap D_{ts} = \emptyset; m = t + v$$

In altre parole il *Classification Model* si compone di un **Learner**, istruito sul *training-set*, e di un **Inducer**, ovvero l'output del modello di classificazione (le previsioni sul *test-set*).

Per valutare la *performance* di un classificatore si analizzano le previsioni corrette, ovvero si analizza la *Confusion-Matrix*, una tabella che mette in relazione i valori attuali con i valori predetti dall'*inducer* (TP, TN, FP, FN). La matrice di confusione permette un confronto tra le performance di diversi modelli di classificazione, uno degli indicatori più utilizzati è l'**Accuracy**, ovvero un indicatore che misura la quota di osservazioni predette correttamente sul totale delle previsioni ($\frac{TP+TN}{TP+TN+FP+FN}$). Una visione alternativa è fornita dallo studio dell'**Error**, che misura la quota di previsioni errate ($\frac{FP+FN}{TP+TN+FP+FN}$); come si può notare i due sono complementari.

		INDUCER PREDICTION (IP)	
		-1	+1
ACTUAL CLASS (AC)	-1	TN	FP
	+1	FN	TP

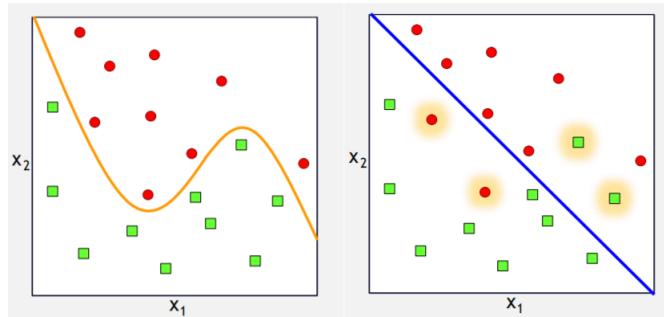
2.2 Performance Evaluation

Rappresenta la fase di valutazione dei risultati ottenuti.

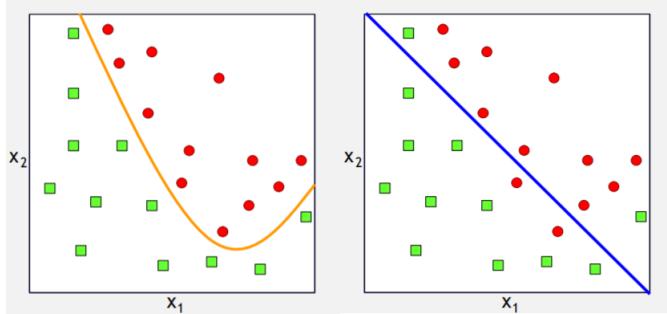
2.2.1 Accuracy and fitting

Non sempre l'*Accuracy* risulta l'indicatore ideale per la valutazione della *performance* di un classificatore; bisogna infatti tenere conto di due possibili fenomeni derivanti dallo sviluppo di un modello di classificazione:

- **Overfitting**: si ha quando il modello di classificazione funziona così bene sul *training-set* che non riesce a generallizzare, e quindi a predire correttamente sul *test-set*.



- **Underfitting:** il modello ha una capacità di classificare le osservazioni troppo bassa.



Tali fenomeni risultano strettamente legati al concetto di errore; possiamo identificare due tipi:

- **Training-Error:** riguarda le osservazioni *misclassified* sul *training-set*.
- **Generalization Error:** rappresenta l'errore atteso nella classificazione delle osservazioni ignote (*test-set*).

Un buon modello di classificazione deve quindi avere entrambi gli errori bassi, il primo per evitare l'*underfitting*, il secondo per evitare l'*overfitting*. Di conseguenza un modello che riesce ad avere un'*Accuracy* molto elevata sul *training-set*, ha una buona probabilità di peccare di generalizzazione, e rischia quindi di incorrere nell'*overfitting*. Al contrario un modello con un'eccessiva generalizzazione, ovvero troppo semplice o rigido, rischia di ottenere errore elevato sia sul *trainig* che sul *test-set* (*underfitting*).

2.2.2 Performance measures

Poiché è spesso necessario utilizzare più di un classificatore durante lo studio di un modello di classificazione, la fase di *performance comparison* risulta fondamentale; è infatti necessario determinare quale sia il classificatore che performa meglio tenendo conto dei dati disponibili per l'analisi. Solitamente si valuta la *performance* di un classificatore in termini di:

- **Accuracy:** misura la capacità del classificatore di prevedere correttamente la classe per le varie osservazioni (*test-set*). Permette inoltre un confronto tra quale sia il classificatore con la miglior *performance* previsiva. Si può definire una *Loss-Function*: $L(y_i, f(\mathbf{x}_i)) = 0$ se $f(\mathbf{x}_i) = y_i$, $= 1$ se $f(\mathbf{x}_i) \neq y_i$; l'*accuracy* può essere quindi calcolata come:

$$acc(D_{ts}) = 1 - \frac{1}{v} \sum_{i=1}^v L(y_i, f(\mathbf{x}_i))$$

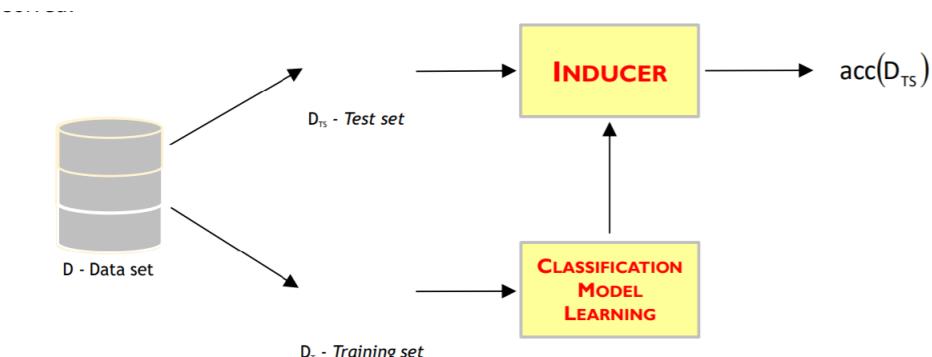
L'*Error* avrà invece una definizione complementare $(1 - acc(D_{ts}))$.

- **Speed:** può essere valutata sia per ciò che concerne la velocità di esecuzione dell'algoritmo (*Learning Time*), sia per lo spazio di memoria occupata. Spesso per ottimizzare un algoritmo secondo tali fattori si può utilizzare solo una parte dei dati, per esempio partizionando le osservazioni o selezionando le variabili più rilevanti.
- **Robustness:** la robustezza di un modello di classificazione si valuta in termini di *outliers*, valori mancanti e variazioni tra *training* e *test-set*.
- **Scalability:** un modello si dice scalabile se è in grado di gestire/apprendere da una grossa quantità di dati; è fortemente legato al concetto di *speed*.
- **Interpretability:** è una condizione necessaria nel caso si voglia studiare il fenomeno in profondità, per esempio analizzando i fattori influenti. Un modello interpretabile può essere di sostegno agli esperti del settore di riferimento (*Domain Experts*).

2.2.3 Estimate Accuracy

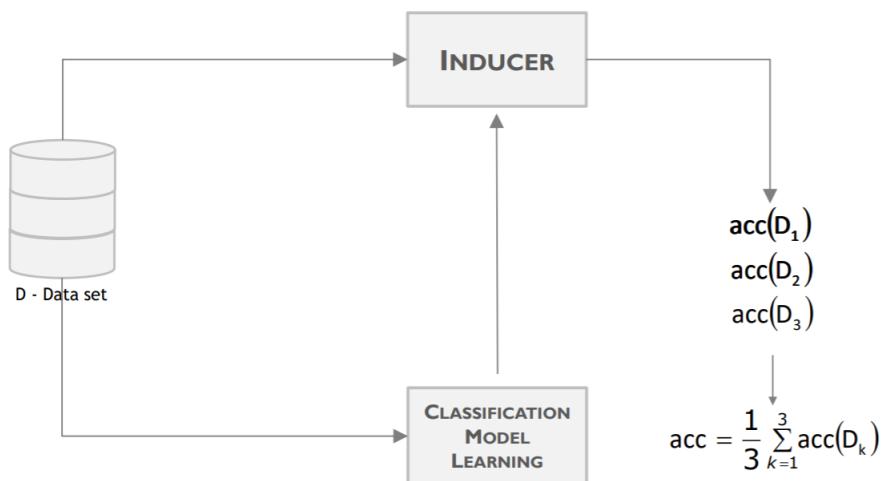
La costruzione di un modello che non rischi di ricadere in situazioni di *overfitting* risulta fondamentale affinché sia riutilizzabile per prevedere le classi di osservazioni ignote. Vi sono alcuni metodi per ovviare a tale situazione:

- **Holdout:** è tra i metodi più semplici, e consiste nella divisione del set di dati in *training* e *test-set*. L'*accuracy* viene calcolata tramite la *Loss-Function* come descritto in precedenza. La divisione consente di separare il set per il *learning* dal set utilizzato per la previsione. Ovviamente la criticità risulta la dipendenza del risultato dalla scelta dei due set.



- **Iterated Holdout:** viene iterato il metodo appena descritto R volte. L'*Accuracy* è data dalla media sugli R diversi D_{ts} ; tale tecnica ha un *Bias* minore rispetto al metodo semplice. Il fatto di non poter controllare quante volte un'osservazione viene inserita all'interno degli R *test-set* può generare distorsioni.
- **K-Fold Cross Validation:** è un metodo leggermente più avanzato rispetto al precedente, infatti è capace di ridurre l'influenza dei valori anomali (*outliers*) durante le fasi di apprendimento e previsione. Consente di assicurarsi

che ciascun osservazione venga inserita lo stesso numero di volte, ovvero una. Il *dataset* viene diviso in K *subset* con un numero costante di osservazioni (D_1, \dots, D_K); vengono effettuati k *learning-testing*, nel quale all'iterazione k si ha $D_T(k)$ composto da tutte le $K - 1$ partizioni (meno la k -esima), mentre $D_{Ts}(k)$ è la k -esima partizione. Attraverso la media aritmetica delle K *accuracies* si può stimare: $acc_{est} = \frac{1}{K} \sum_{k=1}^K acc(D_k)$. Si può richiedere all'algoritmo di contenere in ogni partizione lo stesso rapporto per la variabile *target*, in particolare rispettando la distribuzione dei dati originali: in questo caso si parla di **K-Fold Cross Validation Stratified**. Tale tecnica utilizza un campionamento stratificato al posto di quello casuale semplice.



2.2.4 Comparing Classifiers

Spesso, per risolvere la *classification task*, risulta necessario l'impiego di diversi modelli di classificazione; in tale situazione risulta fondamentale il processo nel quale si determina quale tecnica risulti la migliore in termini di risoluzione della problematica e di *performance*. Vi sono tuttavia situazioni, soprattutto quando i dati a disposizione sono pochi, in cui le misure di accuratezza possono non essere statisticamente significative; quando ci si trova in tali condizioni la non significatività dei risultati può generare diverse criticità nel caso si voglia determinare il miglior classificatore (in termini di accuratezza).

Quando si compara la *performance* di due, o più, classificatori è necessario porsi due domande

- Quanta **Confidence** si può dare all'accuratezza di ciascun classificatore?
- La **differenza** nell'*accuracy* può derivare da una variazione nella **composizione** del *test-set*?

Per rispondere a tali domande si utilizzano due strumenti, rispettivamente lo studio dell'intervallo di confidenza per l'*accuracy* dei classificatori e il test della significatività statistica della deviazione (differenza tra *accuracies*) osservata. Dati D_N dataset di N osservazioni, X il numero di record correttamente classificati e p l'*accuracy* reale, ma ignota, si può costruire un esperimento binomiale:

- $X \sim \text{Binomiale}(N \cdot p, N \cdot p \cdot (1 - p))$
- L'*accuracy* empirica ha forma $\frac{X}{N}$, la quale ha distribuzione binomiale con media pari a p e varianza pari a $\frac{p \cdot (1-p)}{N}$.

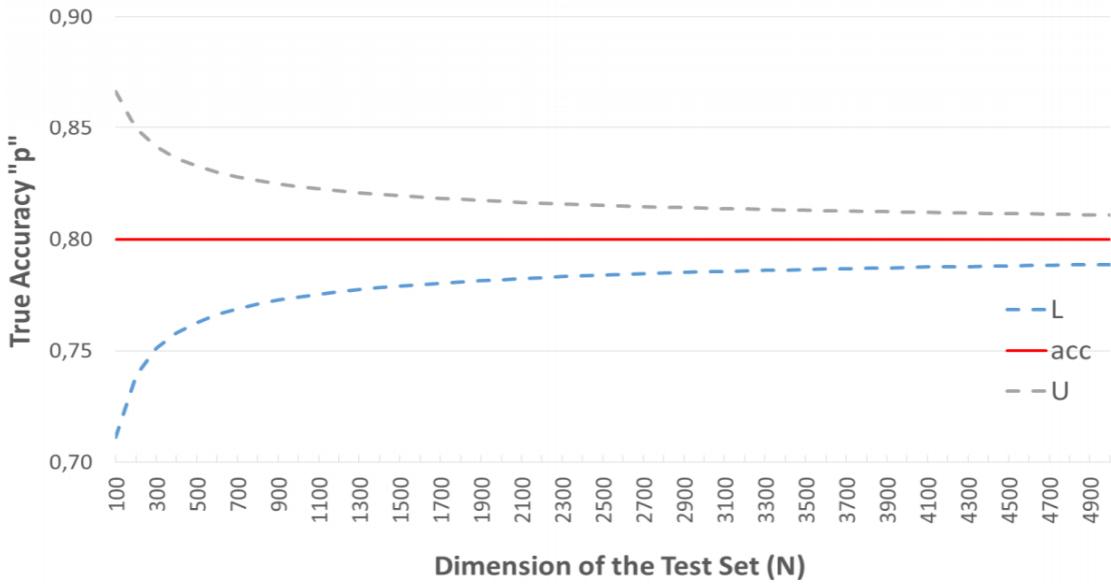
Nonostante la distribuzione binomiale sia utilizzabile al fine di costruire gli I.C per l'*accuracy* empirica, solitamente è conveniente ricondursi ad una distribuzione normale: tale procedimento è tuttavia consigliato solo in caso di *test-set* sufficientemente ampi. Basandosi sulla Normale si può quindi definire la **confidenza** per l'*accuracy* empirica come segue:

$$P(-z_{1-\frac{\alpha}{2}} < \frac{acc - p}{\sqrt{p \cdot (1 - p) / N}} < z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Si possono quindi definire il limite superiore e inferiore, per l'ignota accuratezza del modello di classificazione, come segue:

$$\left[\frac{acc + \frac{Z^2_{1-\frac{\alpha}{2}} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{acc \cdot \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4 \cdot N^2}}}{2 \cdot N}, acc + \frac{Z^2_{1-\frac{\alpha}{2}} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{acc \cdot \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4 \cdot N^2}}}{2 \cdot N}}{1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}}, \frac{Z^2_{1-\frac{\alpha}{2}}}{1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}} \right]$$

Nel caso di un *accuracy* empirica di 0.8 e $N = 10$ (dato $Z_{0.975} = 1.96$), si avrebbe un intervallo di confidenza: $I.C(acc = 0.8) : (0.711, 0.867)$ Di seguito viene mostrato un esempio di come possa variare l'intervallo di confidenza dell'*accuracy* al variare della dimensione del *Test-Set*; si può subito notare che come, al crescere di N , vi sia una tendenziale convergenza dell'I.C verso l'*empirical accuracy* (acc).



Quando si vogliono comparare le *performance* di due modelli di classificazione (per esempio in termini di *accuracy*), risulta importante definire se la differenza nei risultati sia statisticamente significativa. Dati M_1 e M_2 modelli, D_1 e D_2 *Test-Set*, e_1 e e_2 errori di classificazione, assumiamo che le dimensionalità dei *test* sia sufficientemente ampia, e che gli errori siano approssimabili con una Normale; definita $d = e_1 - e_2 \sim N(d_t, \sigma_d^2)$ (con $\sigma_d^2 \simeq \hat{\sigma}_d^2 = \frac{e_1 \cdot (1-e_1)}{n_1} + \frac{e_2 \cdot (1-e_2)}{n_2}$) la differenza tra gli errori, l'I.C per d risulta:

$$IC(d) : (d \pm z_{0.975})$$

Definito $IC(d)$ possiamo interpretare il test sulla *performance* distinguendo tra 3 casi:

- $0 \in IC(d)$: la differenza tra la *performance* dei due classificatori non è significativamente (a livello $\frac{\alpha}{2}$) diversa da 0. Ciò implica che non si possa affermare che vi sia un miglioramento nel passaggio da un classificatore all'altro.
- $Sup_d (d + z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_d)$ risulta negativo: si può concludere che M_1 è migliore di M_2 a livello $\frac{\alpha}{2}$.
- $Inf_d (d - z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_d)$ risulta positivo: si può concludere che M_2 è migliore di M_1 a livello $\frac{\alpha}{2}$.

Quando si utilizza il metodo *K-Fold Cross Validation* (KF-CV) si applicano M_1 e M_2 per K volte, dove all'iterazione k – esima si utilizzano $k-1$ partizioni per il *training-set* e la k – esima partizione per il *test-set*; seguendo tale schema si possono definire M_{1k} e M_{2k} i due modelli alla k – esima iterazione ($K = 1, \dots, K$). Il test statistico si può ora applicare ad ogni iterazione sulla coppia di modelli M_{1k} e M_{2k} , con e_{1k} e e_{2k} errori del modello, $d_k = e_{1k} - e_{2k}$; assumendo K sufficientemente ampio $d_k \sim (d_{cv}, \sigma_{cv})$ con varianza stimata $\hat{\sigma}_{d_{cv}}^2 = \frac{\sum_{k=1}^K (d_k - \bar{d})^2}{K \cdot (K-1)}$ e $\bar{d} = \frac{1}{K} \sum_{k=1}^K d_k$, si utilizza la distribuzione *T-student* per costruire $IC(d_{cv})$ (varianza ignota ma stimata):

$$IC : (\bar{d} \pm t_{1-\frac{\alpha}{2}}^{K-1} \cdot \hat{\sigma}_{dCV})$$

con $t_{1-\frac{\alpha}{2}}^{K-1}$ quantile con confidenza $1 - \alpha$ della *T-student* con $K - 1$ gradi di libertà. Possono essere applicate le stesse 3 considerazioni fatte in precedenza per M_1 e M_2 .

2.3 Class imbalance problem

Il problema dello sbilanciamento dei dati si ha quando la distribuzione della variabile *target* vede una significativa asimmetria tra le possibili classi: in tale situazione si avrà una classe maggioritaria e una minoritaria. Vi sono diverse criticità legate alla *class imbalance*, tra le quali la ridotta significatività dell'*accuracy* la quale potrebbe risultare fuorviante; tale misura infatti tratta tutte le osservazioni egualmente, e in una condizione di sbilanciamento ciò può favorire la classe maggioritaria (un modello che classifica tutte le osservazioni come appartenenti a tale classe potrebbe comunque avere un'*accuracy* molto elevata). La **ZeroR Rule** si riferisce proprio a tale situazione, ovvero ai problemi derivanti dalla minoranza della classe di interesse (*rare class*). Quando si ha un problema di classificazione binaria, la classe minoritaria viene generalmente denominata *positive class*, mentre la maggioritaria *negative class*.

Utilizzando la *Confusion-Matrix* vi sono alcuni indicatori che possono essere considerati in alternativa all'*accuracy*:

- **Sensitivity/Recall:** rappresentano la capacità del classificatore di prevedere i valori positivi (*True Positive Rate*, TPR):

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

- **Specificity:** rappresenta la capacità del classificatore di prevedere correttamente i valori negativi (*True Negative Rate*, TNR):

$$Specificity = \frac{TN}{TN + FP}$$

- **False Positive Rate (FPR):** quota di FP sul totale delle osservazioni (*actual*) negative:

$$FPR = \frac{FP}{TN + FP}$$

- **False Negative Rate (FNR):** quota di FN sul totale delle osservazioni (*actual*) positive:

$$FNR = \frac{FN}{TP + FN}$$

- *Precision*: rappresenta la quota di valori positivi previsti correttamente sul totale delle previsioni positive (indicatore dell'affidabilità nella previsione dei positivi):

$$\boxed{Precision = \frac{TP}{TP + FP}}$$

- *F-Measure*: è un indicatore composto il quale fornisce una valutazione generale della qualità del classificatore. Si calcola come media armonica tra *Recall* (R) e *Precision* (P):

$$\boxed{F = \frac{2 \cdot R \cdot P}{R + P}}$$

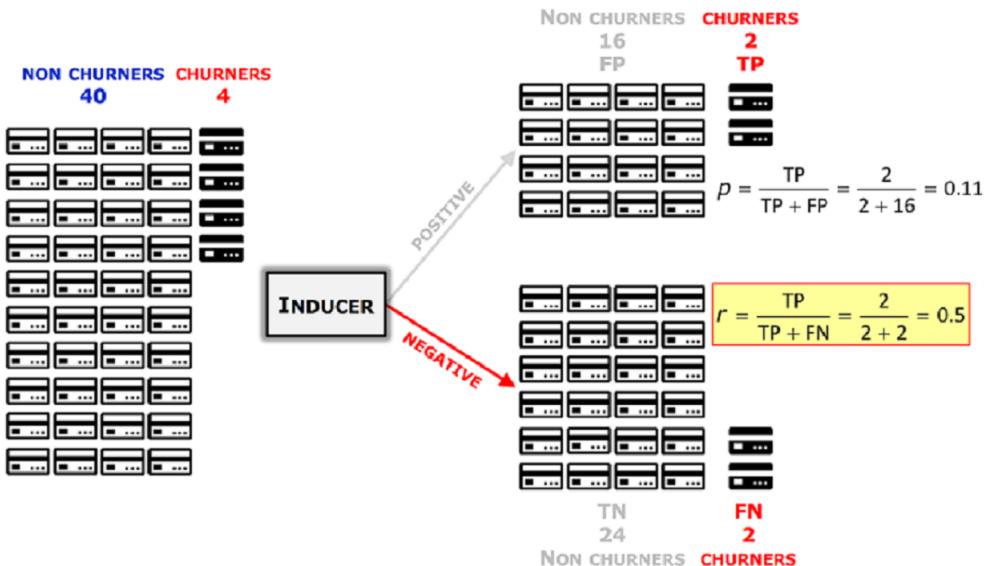
- *F₂-Measure*: poiché si è optato per privilegiare la *Recall* è stata utilizzata anche una trasformata *F_β* (con $\beta = 2$) che ricalcola la *F-Measure* come segue:

$$\boxed{F = \frac{(\beta^2 + 1) \cdot R \cdot P}{R + \beta^2 \cdot P}}$$

Gli indicatori *Recall* e *Precision* risultano molto utili e indicativi quando si vuole privilegiare il corretto riconoscimento di una classe in particolare:

- Un'elevata *Recall* indica che il modello classifica, erroneamente, poche osservazioni appartenenti alla classe positiva come negative.
- Un'elevata *Precision* indica che il modello classifica, erroneamente, poche osservazioni appartenenti alla classe negativa come positive.

Di seguito viene mostrato un esempio in cui il *Classification Task* riguarda una compagnia telefonica, la quale vuole prevedere se i suoi clienti abbandoneranno (*churners*) o no



Solitamente le due misure vengono riassunte nella *F-Measure* e nelle sue trasformate *F_β* ($\beta = 1, 2$).

2.3.1 Counting the cost

Un concetto che si lega molto bene con ciò che è appena stato descritto è quello di matrice dei costi (*Cost-Matrix*): tale strumento consente di conoscere l'eventuale costo (materiale o teorico) che si dovrebbe sostenere in caso di FP e FN. Di seguito viene mostrato un esempio di *Confusion-Matrices* di due diversi modelli di classificazione (riferiti al precedente task: *curners/non-churners*); la tabella sottostante mostra invece la matrice dei costi.

SMP	PREDICTED CLASS		
	-1	+1	
ACTUAL CLASS	-1	830	111
	+1	45	114

ACCURACY = 0.858

M	PREDICTED CLASS		
	-1	+1	
ACTUAL CLASS	-1	931	10
	+1	57	102

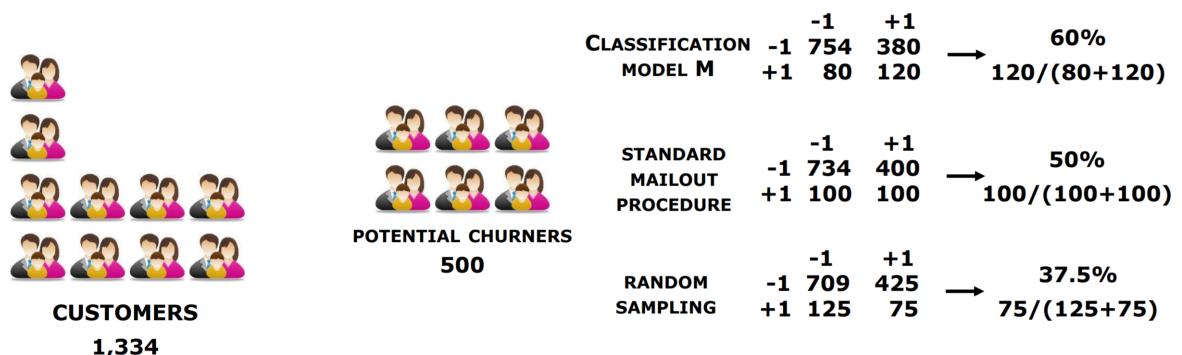
ACCURACY = 0.939

	PREDICTED CLASS		
	-1	+1	
ACTUAL CLASS	-1	0	1
	+1	100	-1

Si può notare come, in questo caso, l'azienda sostiene un costo molto più elevato in caso di un'osservazione erroneamente classificata come negativa (ovvero un *churner* non identificato); il costo di un FN infatti equivale alla perdita di un cliente, il costo di un FP invece all'invio di un eventuale promozione ad un cliente che non avrebbe abbandonato. Il costo totale può essere calcolato come segue:

$$Cost_{TOT} = C_{TN} \cdot TN + C_{FN} \cdot FN + C_{TP} \cdot TP + C_{FP} \cdot FP$$

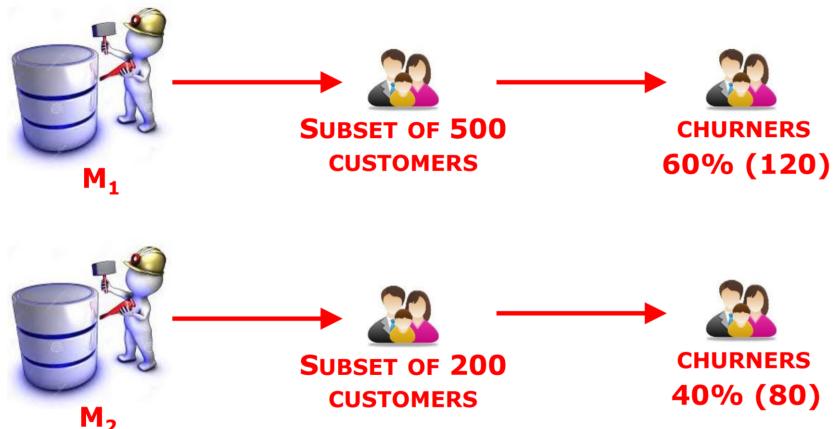
Nel caso specifico, nonostante l'accuracy del secondo modello sia maggiore, si avrebbe un costo di 4,497 contro 5,608. Quando i costi sono speculari ($C_{TP} = C_{TN} = q$ e $C_{FP} = C_{FN} = p$) il costo totale è proporzionale all'accuracy.



Nell'immagine precedente viene mostrato un esempio di *Classification task*, nel quale una compagnia telefonica richiede di identificare dei potenziali *churners* (clienti

che abbandonano) al fine di mettere in atto una campagna promozionale: vi è un vincolo sul numero di promo che possono essere inviate, in particolare un massimo di 500. Considerando sempre la precedente matrice dei costi, e tenendo conto che i *churners* effettivi sono 200 su 1334, vengono sviluppati 3 modelli: un campionamento casuale, un modello SMP e un modello M. Nell'immagine vengono mostrate le *Confusion-Matrx* per ciascuno dei tre modelli, mettendo in risalto la percentuale di *churners* effettivamente identificati. Come si può notare nel passaggio da una metodologia di *random sampling* al modello M vi è un aumento nella percentuale di *churners* correttamente individuati che segue il moltiplicatore 1.6 ($\frac{60\%}{37.5\%} = 1.6$) ; tale coefficiente viene denominato **Lift Factor** del classificatore M rispetto al *random sampling*.

Cambiando la struttura del modello di classificazione (per esempio il *subset*) si possono tuttavia riscontrare delle situazioni differenti, come mostra la figura sottostante.



In una situazione simile, per valutare il miglior *target* per il modello, risulta necessario conoscere la *Profitability*, la quale tuttavia richiede la conoscenza dei costi: il problema è che spesso i costi non sono noti.

Per ovviare a tale problematiche bisogna fare in modo che il subset considerato abbia un'elevata proporzione di osservazioni positive, e che possibilmente sia maggiore di quella del set totale. Immaginiamo ora che il nostro modello M calcoli la probabilità che ciascuna osservazione appartenente al *subset* possa essere un *churner*, come nell'esempio sottostante (10 osservazioni). In tale esempio si passa dalla prima tabella alla seconda riordinando in senso decrescente secondo la probabilità $P(y = 1)$, ovvero quella di *churning*; come viene evidenziato dei 3 clienti con P maggiore vi sono 2 *churners* su 3. Successivamente viene invece considerato un subset di 5 osservazioni, sempre considerando quelle con probabilità maggiore; vengono messe in risalto per entrambe le situazioni la porzione di record positivi correttamente identificati ($\frac{2}{3}$ per entrambi) e le rispettive lift su un classificatore casuale (con *positive rate* pari a $\frac{3}{10}$):

- $Lift_{s3} = \frac{0.66}{0.3} = 2.22$
- $Lift_{s5} = \frac{0.66}{0.5} = 1.33$

Rank	RowID	Prob(y)
1	134	0.95
2	221	0.88
3	18,923	0.86
4	90,034	0.73
5	874,823	0.64
6	67	0.47
7	98,324	0.32
8	2,553	0.15
9	289	0.10
10	2,349	0.03

→

Rank	RowID	Prob(y)	Churn
1	134	0.95	y
2	221	0.88	y
3	18,923	0.86	n
4	90,034	0.73	n
5	874,823	0.64	n
6	67	0.47	n
7	98,324	0.32	n
8	2,553	0.15	y
9	289	0.10	n
10	2,349	0.03	n

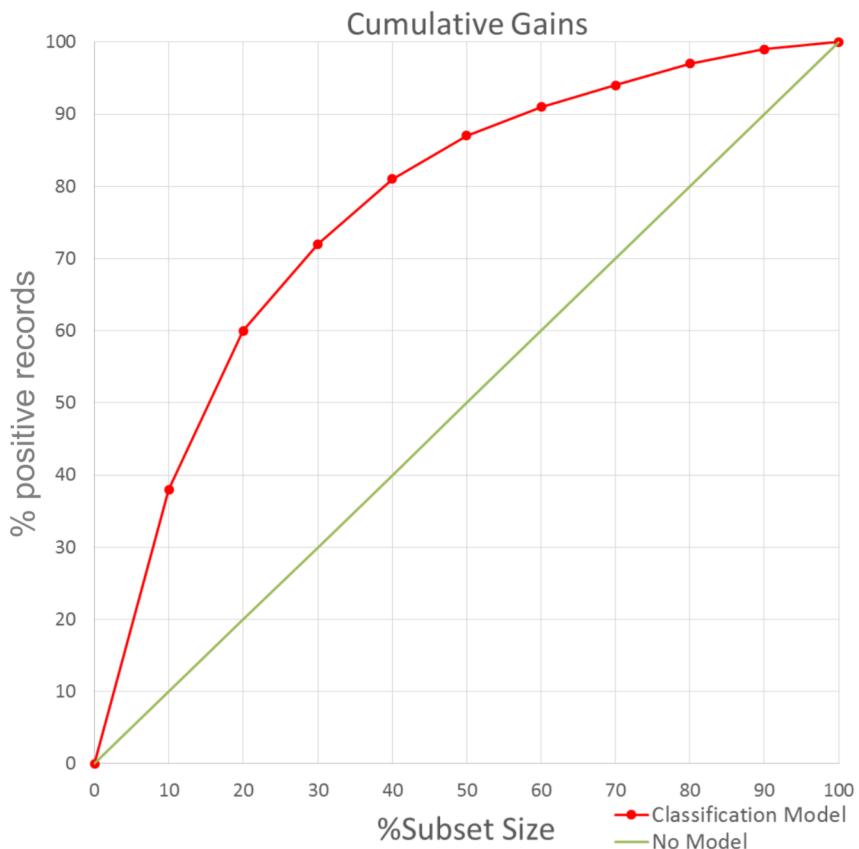
subset consisting of 3 records
2 records out of the 3 positive ones
are correctly identified
 $2/3 = 0.66$ of positive records
Lift = $0.66/0.3 = 2.22$

→

Rank	RowID	Prob(y)	Churn
1	134	0.95	y
2	221	0.88	y
3	18,923	0.86	n
4	90,034	0.73	n
5	874,823	0.64	n
6	67	0.47	n
7	98,324	0.32	n
8	2,553	0.15	y
9	289	0.10	n
10	2,349	0.03	n

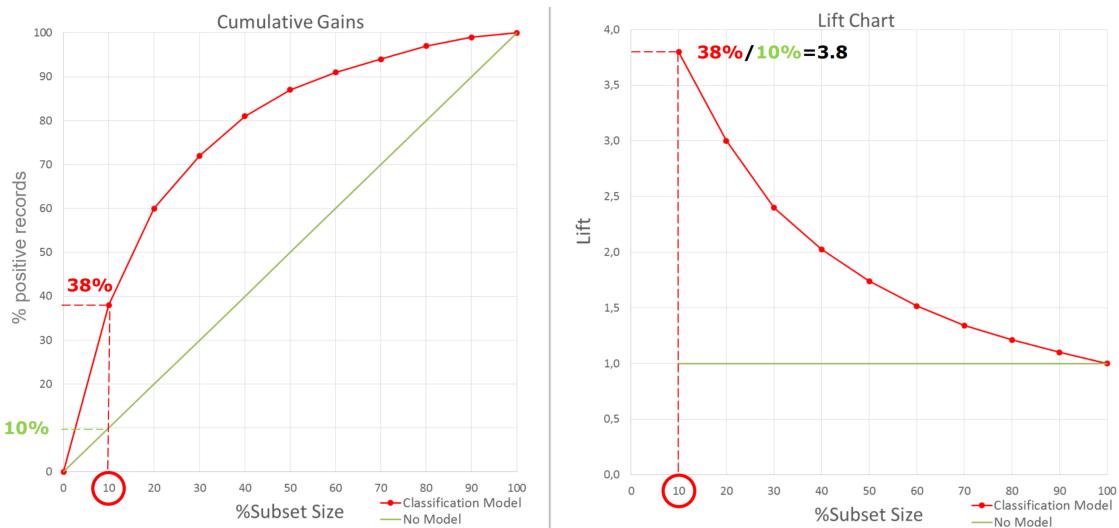
subset consisting of 5 records
2 records out of the 3 positive ones
are correctly identified
 $2/3 = 0.66$ of positive records
Lift = $0.66/0.5 = 1.33$

La seguente situazione può essere efficacemente riassunta nella **Cumulative Gains**, ovvero un grafico in cui vengono messe a confronto la percentuale di set considerata (*subset*) e il tasso di record positivi (*positive rate*):



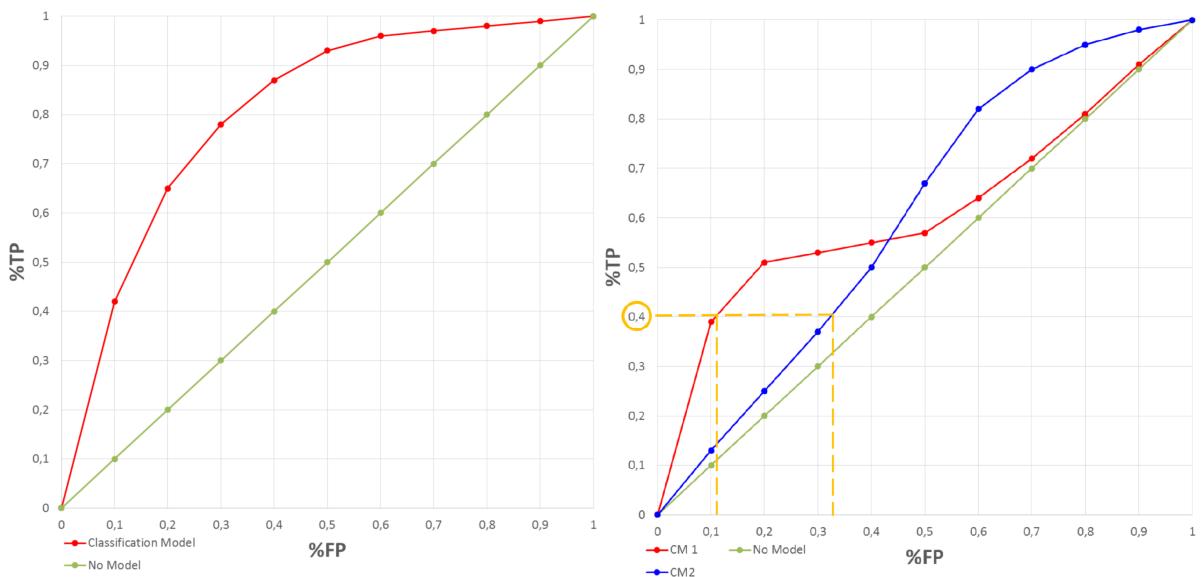
La curva rossa mostra la *cumulative gain*, mentre la bisettrice indica il risultato di un classificatore *random sampling*. La curva mostra come, con il modello in esame, passando da un *subset* del 10% ad uno del 30% si ha un aumento del *positive records rate* dal 38% al 72%.

Una visualizzazione alternativa strettamente collegata alla *cumulative gains* è il grafico denominato **Lift Chart**; in tale grafico viene infatti messa in risalto la relazione tra la percentuale di dati utilizzati come *subset* e la *lift*.



Si può notare come i due grafici abbiano lo stesso asse delle ascisse, poiché entrambi valutano le caratteristiche di un modello rispetto alla partizione di set considerato; un'altra caratteristica in comune è la curva verde, associata alla *ZeroR Rule* (no modello).

Un ultimo grafico correlato ai precedenti è la *Receiver Operating Characteristic Curve* (**ROC-Curve**); la ROC mette in relazione il *true positive rate* (TPR) con il *false negative rate* (FPR) come segue:



La curva ROC ideale costeggerebbe la zona sinistra del grafico (formando un angolo retto con l'asse delle ordinate); la curva verde rappresenta ancora una volta la condizione *ZeroR Rule* (niente modello). Nel secondo grafico viene messa in risalto

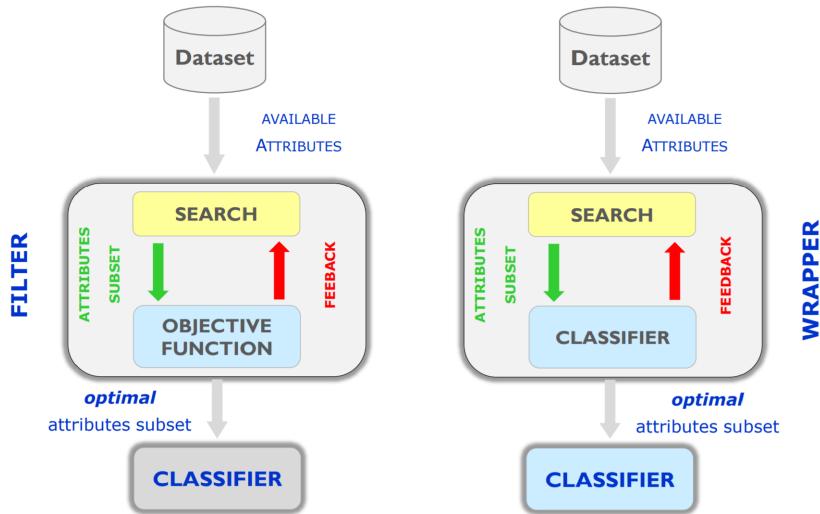
un confronto tra le ROC di due modelli, M_1 e M_2 : la ROC ci dice che il modello 1 risulta migliore per *subset* piccoli (inferiorio a 45% circa), mentre il modello 2 per set più ampi. L'area sottostante la curva ROC (**Area Under Curve**, AUC) rappresenta un indicatore sintetico di accuratezza del modello.

2.4 Feature Selection

Una fase fondamentale del processo di classificazione è la selezione delle variabili: lo scopo è quello di eliminare attributi ridondanti (con informazioni ricavabili da altri attributi) o irrilevanti (non utili alla risoluzione della *classification task*). Il vantaggio di una buona *feature selection* riguarda sia la precisione del modello, sia il miglioramento del processo in termini di onerosità computazionale. Esistono differenti metodi:

- **Brute-Force**: studio di tutti i possibili *subset* di attributi che possono essere impiegati; risulta quasi sempre troppo oneroso (20 *features* possono comporre $\sum_{n=1}^{20} \binom{20}{n} = 1,048,575$ possibili *subset*).
- **Embeeded**: la selezione delle variabili è un prodotto di un classificatore (*Decision Tree*, *Bayesian Network*, ...)
- **Filter**: la selezione avviene prima del processo di *learning* del classificatore.
- **Wrapper**: un classificatore viene utilizzato per trovare il subset ottimale dagli attributi a disposizione.

2.4.1 Filter e Wrapper



Come mostra la figura, in cui vengono confrontati il funzionamento della tecnica **Filter** e **Wrapper**, la principale differenza è identificabile nello strumento utilizzato per la selezione delle variabili:

- Il *Filter* utilizza una Funzione Obiettivo (**Objective Function**).
- il *Wrapper* utilizza invece un **Classificatore**

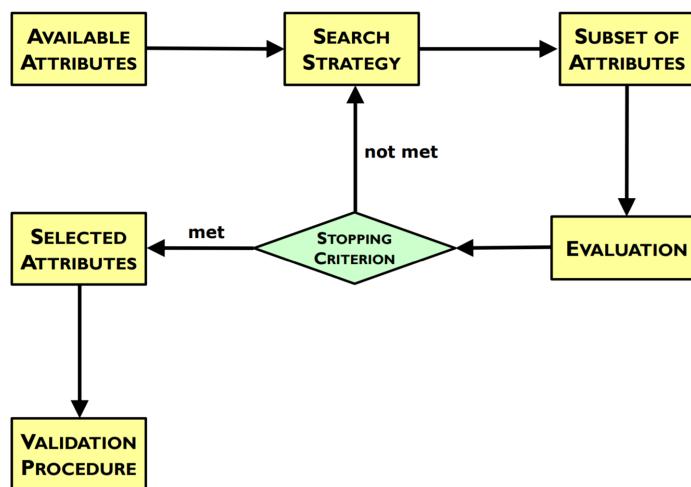
Risulta chiaro che entrambi i metodi, e di conseguenza i *subset* di variabili risultanti, sono strettamente legati alla scelta della funzione o del classificatore: cambiando tali elementi si ha un cambio nel set di attributi selezionati.

Il *filter* può essere di due tipologie:

Uni-Variate	Multi-Variate
Parametric	Correlation Feature Selection (CFS)
t-test ANOVA Mutual Information	Relief Blanket
Non Parametric	
Mann-Whitney Kruskal-Wallis Permutation test	

	Advantages	Disadvantages
Uni-Variate	speed scalability independent on the classifier	ignore that attributes can be dependent ignore interactions with the classifier
Multi-Variate	model dependency between attributes independent on the classifier computational cost compares favorably to wrapper	slower than Uni-Variate techniques less scalable than Uni-Variate Techniques ignore interactions with the classifier

- **Univariato** → viene definita una misura di associazione tra le *features* candidate, successivamente si riordinano le variabili in accordo con le misure di associazione e infine vengono selezionate gli *R* attributi da utilizzare come input per la classificazione. Tale metodologia consente l'eliminazione degli attributi irrilevanti, ma non di quelli ridondanti.
- **Multivariato** → a differenza del precedente, un *Multivariate Filter* consente l'identificazione sia degli attributi irrilevanti, sia degli attributi ridondanti; la regola di selezione segue il ragionamento secondo il quale un buon *subset* deve contenere attributi fortemente correlati con la variabile *target*, ma incorrelati tra loro. Vengono impiegate misure di simmetria o di correlazione.



Spesso può essere conveniente, al fine di ridurre la dimensionalità degli attributi, applicare un processo di *Feature Creation*, ovvero la creazione di nuove variabili capaci di catturare, e riassumere, l'informazione contenuta in altre. Esistono principalmente 3 tecniche:

- Feature Extraction
- Mapping data to a New Space
- Feature Construction

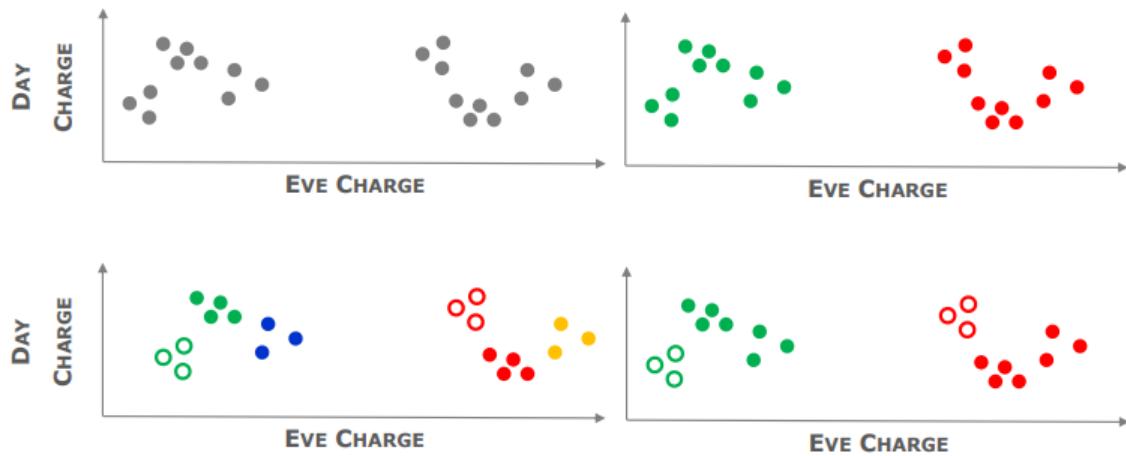
Clustering

3.1 Introduzione

Il **Clustering** comprende una serie di tecniche volte all'identificazione di gruppi (*cluster*) di osservazioni (all'interno di una determinata popolazione di riferimento) che abbiano alcune caratteristiche comuni/simili. Tra i maggiori vantaggi dell'impiego di metodologie di *clustering* ne possiamo individuare 2:

- **Understanding** → il *clustering* permette di sviluppare alcune conoscenze circa le caratteristiche più significative che legano (o differenziano) i diversi *cluster*
- **Utility** → il *clustering* consente di mettere in atto azioni (per esempio di *marketing*) che possano essere applicate a individui considerati simili.

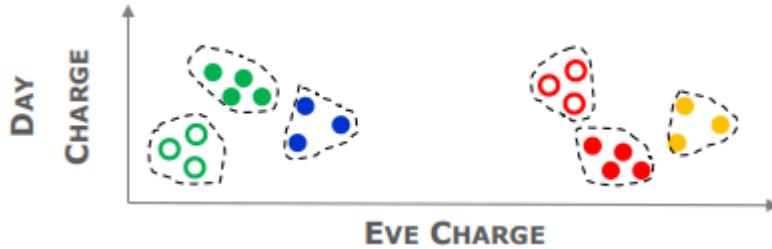
Nell'immagine sottostante viene mostrato il possibile risultato di diverse metodologie di *clustering* (a partire dalla popolazione in alto a sinistra).



Esistono un gran numero di tecniche (algoritmi) di *clustering*, tuttavia possiamo distinguerli per alcune caratteristiche principali:

1. Famiglia:

- *Partitional Clustering* → vengono definiti anche *un-nested*; consiste in una divisione delle osservazioni in *cluster* esclusivi (*Non-Overlapping*), ovvero in modo che ciascuna osservazione venga associata ad un unico cluster.



- *Hierarchical Clustering* → vengono definiti anche *nested*; a differenza del precedente i metodi gerarchici consentono la formazione di sotto-*cluster* (i quali possono essere aggregati o disaggregati), i quali vengono organizzati in una struttura ad albero (*Dendrogram*) come mostrato nell'immagine sottostante.



2. Inclusività delle osservazioni

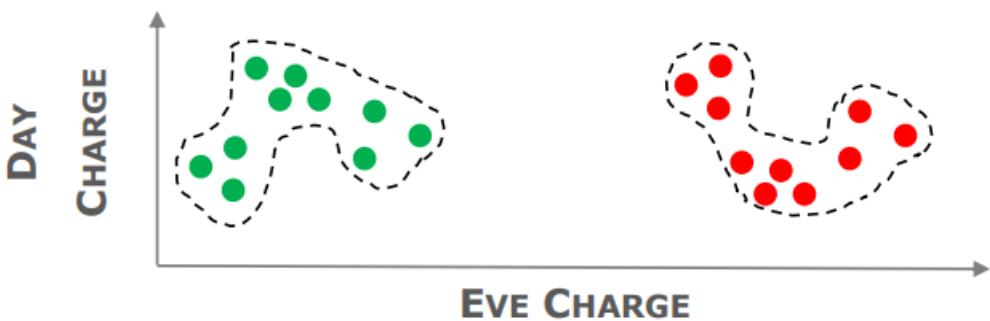
- *Exclusive* → ne fanno parte tutte quelle tecniche che prevedono che ciascuna osservazione possa essere assegnata ad un unico cluster.
- *Overlapping* → situazione opposta rispetto alla precedente; ne fanno parte tutte quelle tecniche di *clustering* che prevedono la possibilità che due differenti *cluster* condividano un'osservazione (*overlap*)
- *Fuzzy* → permettono a ciascuna osservazione di appartenere ad ogni *cluster*, tuttavia assegnando un peso ($w \in [0, 1]$) all'appartenenza.

3. Assegnazione delle osservazioni

- *Complete* → il metodo di *clustering* è vincolato ad assegnare ogni osservazione ad uno specifico *cluster*
- *Partial* → non vi sono obblighi di assegnazione; la motivazione è che alcune osservazioni potrebbero non appartenere ad un *cluster* definito (*outliers*).

3.1.1 Well-Separated Cluster

Il risultato ideale, il quale può giustificare l'impiego di una tecnica di segmentazione, è l'ottenimento di una suddivisione della popolazione in *cluster* che contengano osservazioni molto più simili tra loro, rispetto alle osservazioni appartenenti ad altri gruppi; per spiegare meglio lo scopo risiede nella divisione della popolazione in gruppi che siano omogenei all'interno, ma eterogenei tra loro (*Well-Separated Cluster*).

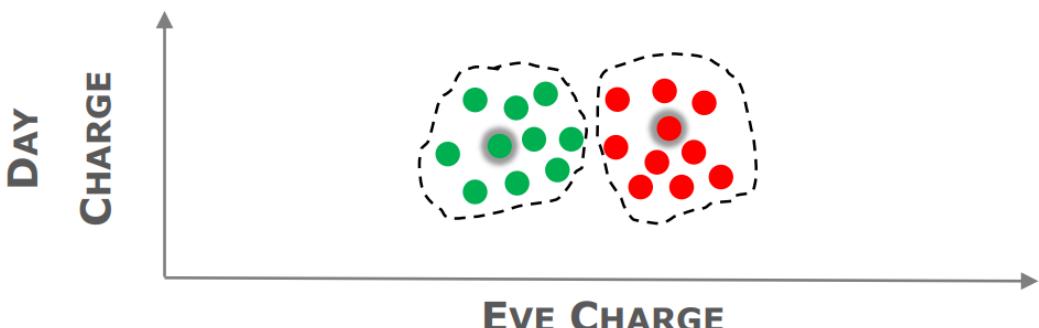


3.1.2 Prototype-Based Clustering

Ovviamente questa è una condizione che si può verificare solo in caso di presenza di *cluster* naturali, per questo risulta spesso conveniente utilizzare metodi **Prototype-Based**: tale tecnica prevede l'assegnazione di ciascuna osservazione ad uno specifico gruppo sulla base della sua vicinanza ad un oggetto (naturale o artificiale) rappresentativo del *cluster*.

Nel caso di attributi continui, viene spesso utilizzato un **centroide** come *Prototype-Object*, ovvero il punto che rappresenta la media di tutti gli oggetti del cluster; si dirà ora che i cluster saranno ben separati se ciascuna osservazione appartenente ad un gruppo risulta più vicina al proprio centroide, rispetto che a tutti gli altri. Nel caso di attributi nominali, i quali non hanno una rappresentazione media, vengono utilizzati i **medoidi**; i metodi *prototype-based* tendono ad assumere forme globulari.

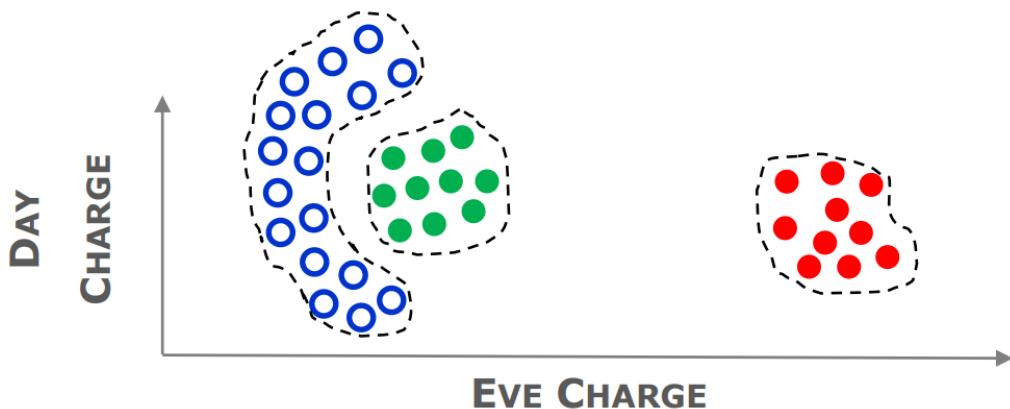
Nel grafico in esempio vengono mostrati due *Prototype-Based clusters*, evidenziando il loro centroidi e la forma globulare. Si può notare che, rispetto alla situazione *well-separated*, vi sono alcune osservazioni che risultano più vicine ad alcuni oggetti di un altro *cluster*, rispetto ad altri appartenenti allo stesso gruppo



3.1.3 Density-based Cluster

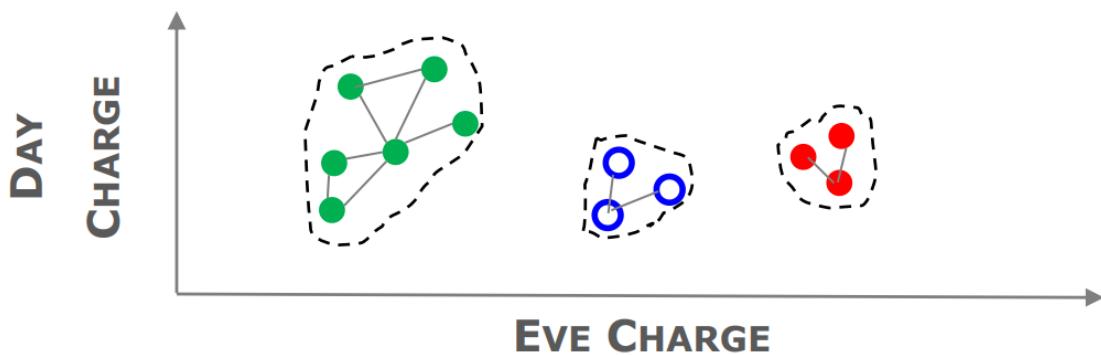
Una situazione differente rispetto al *prototype-based* è evidenziata dalle tecniche **Density-Based Cluster**; per questa particolare tipologia di *clustering* si fa uso del concetto di densità degli oggetti, o meglio di determinate regioni nello spazio delle osservazioni. Basandosi su tale strumento, tali tecniche individuano i *cluster* come regioni dello spazio in cui vi è un'alta densità di osservazioni, e li separano quando vengono identificate zone a bassa densità di osservazioni (come nell'immagine sottostante). Lo spazio verrà quindi suddiviso in segmenti di popolazione definite da una particolare forma (*shape*) che rappresenta la concentrazione di osservazioni.

Risultano estremamente utili nel caso di cluster irregolari (non globulari), intrecciati o in presenza di rumori e *outliers*.



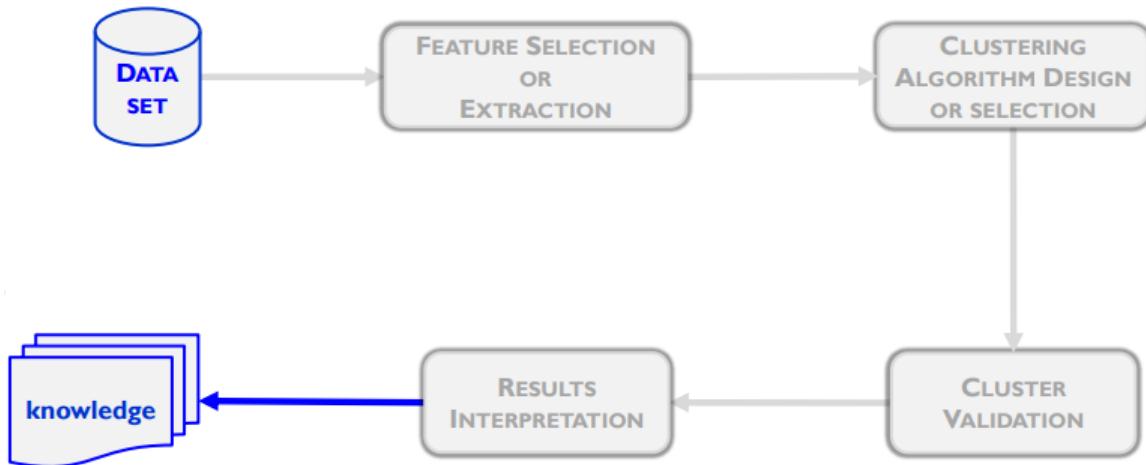
3.1.4 Graph-based Cluster

Un'ultima tipologia di tecniche di *clustering* è la **Graph-Based cluster**: l'assunzione di base è che lo spazio delle osservazioni possa essere rappresentato come una struttura a grafo, dove le osservazioni rappresentano i nodi, mentre i collegamenti (*links*) tra esse sono le connessioni. Partendo da tale struttura, un *cluster* viene identificato come un gruppo di osservazioni collegate tra loro, ma che non hanno connessioni con gli altri segmenti; i *Graph-based* tendono spesso ad assumere forme globulari (sferiche, ellittiche, ...). L'immagine sottostante mostra un esempio di una possibile struttura di grafi.



3.1.5 The clusetring "Cicle"

Come nella classificazione, anche il *clustering* prevede diverse fasi, tutte necessarie al fine di estrarre informazione di qualità dai dati originali; l'immagine sottostante mostra i principali passaggi, necessari a sviluppare un buon metodo di segmentazione.



3.2 Proximity

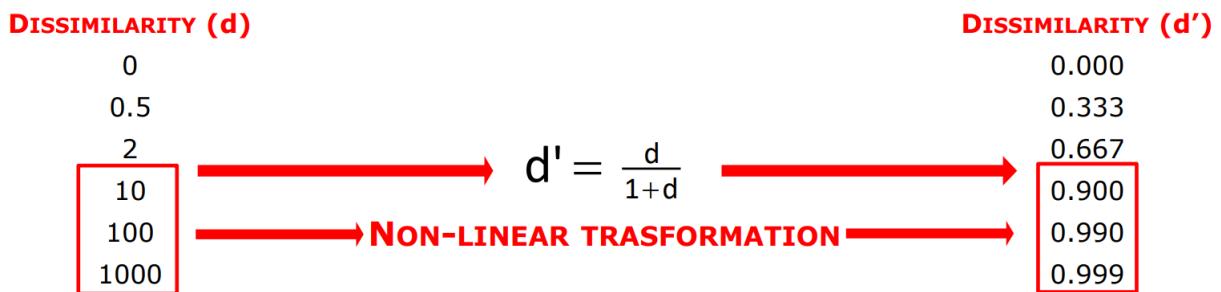
L'aggregazione di osservazioni, ovvero lo scopo del *clustering*, si fonda sul "mettere insieme" oggetti simili; di conseguenza risulta necessario chiarire cosa s'intenda per similarità/non similarità (*Similarity/Dissimilarity*) tra osservazioni. Iniziando a distinguere i due concetti:

- **Similarity** → la similarità tra due oggetti si definisce come una misura del grado di somiglianza/vicinanza di due oggetti. La *similarity* $\in [0, 1]$ (non negativa), con valore 0 quando 2 oggetti non sono simili, e valore 1 quando si è in condizione di *complete similarity*.
- **Dissimilarity** → ha un'interpretazione opposta, ovvero misura il grado in cui due oggetti sono non somiglianti; maggiore sarà la somiglianza, minore sarà la *dissimilarity*. Solitamente $\in [0, 1]$, ma non è inusuale che vari in un range $[0, +\infty)$

Solitamente si utilizza tuttavia un'altra misura per lo sviluppo di modelli di *clustering*: la **Proximity**. Tale misura rappresenta una trasformata utilizzata per passare da una misura di similarità ad una di dissimilarità (e viceversa); viene ricondotta ad un intervallo $[0, 1]$, e misura la frazione di similarità/dissimilarità tra due osservazioni. Partendo da una misura di *similarity* $\in [1, 10]$, una possibile misura di prossimità è data da:

$$S' = \frac{S - S_{min}}{S_{max} - S_{min}} = \frac{S - 1}{9} \in [0, 1]$$

la quale vale anche per una misura di dissimilarità ($D \rightarrow D'$) La principale criticità nell'applicazione di una trasformata che porti ad una misura di *proximity* $\in [0, 1]$ si riscontra in caso S o D fossero definite in range molto ampi (per esempio $[0, +\infty]$); in tale situazione andrebbe infatti applicata una trasformazione non lineare, e si otterebbe una compressione per i valori più elevati, modificando drasticamente la scala dei valori originali (come mostrato in figura).



Il caso più semplice per passare da *Similarity* a *Dissimilarity* (e viceversa) risulta quello in cui entrambe le misure esistono nell'intervallo $[0, 1]$, e in particolare si avrebbe :

$$D = 1 - S ; S = 1 - D \quad (S, D \in [0, 1])$$

Un'altra possibilità è quella di definire l'una come il negativo dell'altra:

$$S = -D ; D = -S$$

Dal momento che la prossimità di due osservazioni è definita dalla prossimità degli attributi per tali osservazioni, possiamo definire alcuni esempi di *Proximity* per casi diversi (come mostra l'immagine sottostante in riferimento ad un unico attributo).

ATTRIBUTE TYPE	DISSIMILARITY	SIMILARITY
CATEGORICAL (QUALITATIVE)	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$
	$d = \frac{ x - y }{n - 1}$	$s = 1 - d$
NUMERIC (QUANTITATIVE)	$d = x - y $	$s = -d \quad s = \exp(-d)$ $s = \frac{1}{1 + d}$
		$s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

con x, y valori che assume un determinato attributo per due osservazioni; vengono definite le trasformate per attributi qualitativi (nominali/ordinali) e quantitativi.

3.2.1 Proximity Measures

Il caso precedente è ovviamente vincolato alla presenza di un unico attributo, utilizzato per la valutazione della similarità/dissimilarità; nella realtà tuttavia si ha molto spesso a che fare con dati multidimensionali, e di conseguenza bisogna ridefinire le tecniche che calcolano la di *Proximity*. Per fare ciò, solitamente, si fa riferimento alle **Distanze**, una particolare tipologia di misure con alcune proprietà in comune con la *dissimilarity*; le principali appartengono alla famiglia delle distanze **Minkowsky**, le quali seguono la seguente distribuzione

$$d_M(\vec{x}, \vec{y}) = \sqrt[r]{\sum_{k=1}^n |x_k - y_k|^r}$$

A partire da tale formula, variando il parametro r , si ottengono diverse misure di distanza:

- **Manhattan Distance** ($r = 1$) $\rightarrow d_m(\vec{x}, \vec{y}) = \sum_{k=1}^n |x_k - y_k|$
- **Euclidean Distance** ($r = 2$) $\rightarrow d^2(\vec{x}, \vec{y}) = \sqrt[2]{\sum_{k=1}^n (x_k - y_k)^2}$
- **Supremum Distance** ($r = \infty$) $\rightarrow d_\infty(\vec{x}, \vec{y}) = \lim_{r \rightarrow \infty} \sqrt[r]{\sum_{k=1}^n |x_k - y_k|}$

Tali misure devono rispettare alcune proprietà :

Proprietà	Definizione
Non negatività	$d(\vec{x}, \vec{y}) \geq 0$ e $d(\vec{x}, \vec{y}) = 0 \leftrightarrow \vec{x} = \vec{y}$ ($\forall \vec{x}, \vec{y}$)
Simmetria	$d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$ ($\forall \vec{x}, \vec{y}$)
Disuguaglianza triangolare	$d(\vec{x}, \vec{z}) \leq d(\vec{x}, \vec{y}) + d(\vec{y}, \vec{z})$ ($\forall \vec{x}, \vec{y}, \vec{z}$)

Una misura di *dissimilarity* che rispetti tutte e 3 tali proprietà viene definita **Metrica**; e misure di *Similarity* non rispettano la terza proprietà, mentre le prime due sono solitamente applicabili.

Se per le variabili continue l'utilizzo delle distanze come misura della *proximity* è molto utilizzata, per le variabili con attributo binario si utilizzano i **Coefficienti di similarità**, i quali assumono valore 1 in caso di perfetta similarità, e valore 0 in caso di assenza totale. Tra i principali troviamo:

- **Simple Matching Coefficient** → Rispetta le proprietà di simmetria e non negatività; poichè considera ugualmente la presenza (1) e l'assenza (0) di una determinata variabile binaria, è utilizzabile solo quando gli attributi sono binari e simmetrici.

$$SMC(\vec{x}, \vec{y}) = \frac{\text{matching attributes}}{\text{attributes}} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{10} + f_{01}} \in [0, 1]$$

- **Jaccard Coefficient** → A differenza del precedente viene utilizzato quando gli attributi sono binari, ma asimmetrici; ciò implica che la comune assenza di un attributo non può essere considerata indice di somiglianza tra le osservazioni (il fatto che due persone abbiano comprato lo stesso prodotto le rende simili, il contrario no).

$$J(\vec{x}, \vec{y}) = \frac{\text{matching presences}}{\text{attributes except 00matches}} = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} \in [0, 1]$$

- **Tanimoto Coefficient** → Viene definito anche *Extended Jaccard Coef.*; solitamente utilizzato in caso di attributi binari per i dati documentali

$$EJ(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|^2 + \|\vec{y}\|^2 - \vec{x} \cdot \vec{y}}$$

- **Cosine Similarity** → Viene utilizzata quando tutti gli attributi sono numerici; come Jaccard ignora gli 00matches, ma può essere impiegata anche per trattare attributi non binari. Solitamente viene utilizzata in caso di dati sparsi, e in particolare nell'*Information Retrieval* (con i documenti rappresentati come vettori).

$$COS(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} \in [0, 1]$$

- **Pearson Correlation Coefficients** → Indica la correlazione tra le due variabili, e assume valore massimo quando è uguale a |1|, mentre indica assenza di relazione quando assume valore 0. Si utilizza solo in caso in cui tutti gli attributi risultino numerici

$$Cor(\vec{x}, \vec{y}) = \frac{cov(\vec{x}, \vec{y})}{stdv(\vec{x}) \cdot stdv(\vec{y})} \in [-1, 1]$$

Ciascuna misura può essere ottimale in diverse situazioni; esistono tuttavia numerose **Criticità** che si possono presentare durante la selezione dell'indicatore di prossimità.

1. La prima riguarda senza dubbio le **differenze di scala**: nel caso vi sia una variabile con un *range* di variazione molto più ampio rispetto alle altre, durante il calcolo di una misura di prossimità (per esempio una distanza euclidea), si avrebbe che tale variabile dominerebbe rispetto alle altre, distorcendo così i risultati ottenuti. Per ovviare a tale situazione si applicano **standardizzazioni/normalizzazioni**, al fine di riportare ciascun attributo ad un *range* di variazione comune.
2. La seconda si ha quando gli **attributi** utilizzati per definire la prossimità tra attributi siano **correlati**. Solitamente se vi è correlazione, un *range* di

variazione differente e la possibilità di assumere una distribuzione Gaussiana, può essere utile l'impiego della **Distanza di Mahalanobis**:

$$D_{ML}(\vec{x}, \vec{y}) = (\vec{x} - \vec{y})\Sigma^{-1}(\vec{x} - \vec{y})^T \in [0, +\infty]$$

con Σ^{-1} inversa della matrice di varianze/covarianze. Assume valore 0 in caso \vec{x} e \vec{y} siano uguali ed è applicabile solo ad attributi numerici.

3. La terza riguarda i casi in cui si riscontra la presenza di **attributi di tipologie diverse** (binari, continui, nominali...). Una possibile soluzione consiste nel valutare la similarità di ciascun attributo singolarmente; dati K attributi, si calcola la similarità tra \vec{x} e \vec{y} per il singolo attributo k ($S_k(\vec{x}, \vec{y}) \in [0, 1]$) e si assume che ciascuno dei K attributi contribuisca ugualmente.

$$\delta_k = \begin{cases} 0, & \text{se il k-esimo è asimmetrico e assume valore 00, o se è missing} \\ 1, & \text{altrimenti.} \end{cases}$$

Successivamente si definisce un indicatore δ_k (per il k -esimo attributo), che assume la seguente forma: Infine si calcola un indicatore aggregato dato da:

$$similarity(\vec{x}, \vec{y}) = \frac{\sum_{k=1}^K \delta_k \cdot s_k(\vec{x}, \vec{y})}{\sum_{k=1}^K \delta_k} \quad \text{con } k = 1, \dots, K \in N$$

4. L'ultima problematica riguarda le situazioni in cui gli **attributi contribuiscono con pesi differenti** alla definizione della prossimità. Nel caso in cui si assuma che i pesi w_k (con $k = 1, \dots, K$) sommino a 1 ($\sum_{k=1}^K w_k = 1$) si può aggiustare la precedente formula per tenere conto del peso di ciascun attributo:

$$similarity(\vec{x}, \vec{y}) = \frac{\sum_{k=1}^K \delta_k \cdot w_k \cdot s_k(\vec{x}, \vec{y})}{\sum_{k=1}^K \delta_k} \quad \text{con } k = 1, \dots, K \in N$$

oppure utilizzando la distanza di Minkowsky:

$$d_{Mw}(\vec{x}, \vec{y}) = \sqrt[r]{\sum_{k=1}^n w_k \cdot |x_k - y_k|^r} \quad \text{con } k = 1, \dots, K \in N$$

In generale è consigliabile seguire il seguente schema:

✓ DENSE AND CONTINUOUS DATA	metric distance measures (Euclidean) are often used
✓ SPARSE DATA, ASYMMETRIC BINARY	similarity measures which ignore 00 matches cosine, Jaccard and Extended Jaccard

3.3 Clustering Evaluation

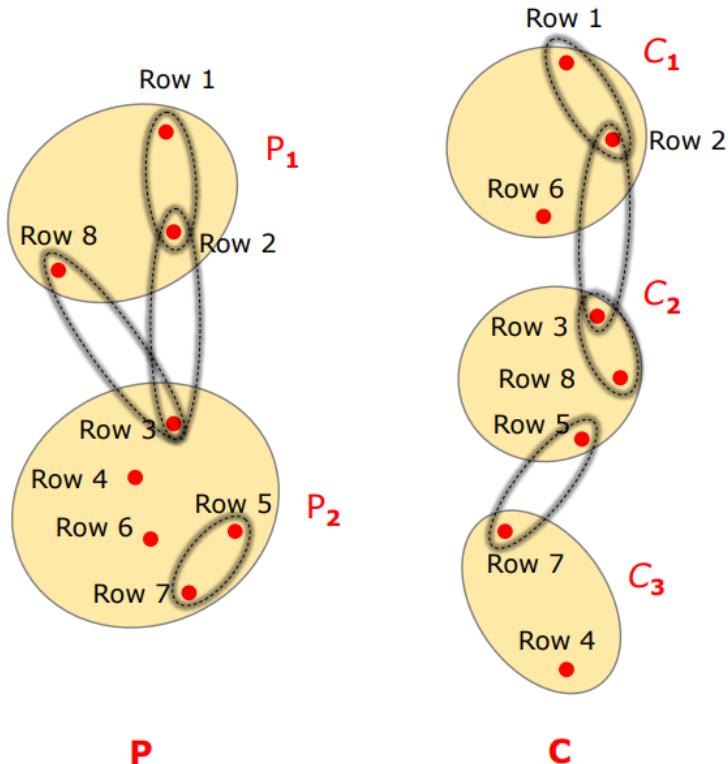
Dopo aver definito la misura di similarità, il metodo di **clustering** e il numero di gruppi ideale, è necessario valutare la qualità dei risultati ottenuti. A differenza della classificazione (dove esistono apposite misure e metodi), nella segmentazione è spesso difficile valutare il risultato, soprattutto quando non esistono *cluster* naturali. Esistono comunque alcune misure che permettono di applicare la **Cluster Validation**

3.3.1 External/Supervised measures

Sono metodi basate sull'utilizzo di strutture esterne al *clustering*. Supponendo che esista una variabile di classe che assuma R valori (con $P = \{P_1, \dots, P_r\}$ R partizioni del set di dati), e che il metodo di *clustering* abbia identificato K gruppi (con $C = \{C_1, \dots, C_K\}$ partizioni del set di dati), possiamo distinguere alcuni casi:

- x e y appartengono allo stesso cluster C e alla stessa classe P .
- x e y appartengono allo stesso cluster C , ma non alla stessa classe P .
- x e y non appartengono allo stesso cluster C , ma appartengono alla stessa classe P .
- x e y non appartengono né allo stesso cluster, né C e alla stessa classe P .

Nell'immagine sottostante viene mostrato un esempio con $K = 2$ e $R = 3$.



Definendo a, b, c, d rispettivamente il numero di osservazioni appartenenti ai casi 1,2,3,4, e M come il numero di coppie possibili ($M = \frac{m \cdot (m-1)}{2} = a + b + c + d$), possiamo definire alcuni indici:

Indice	Formula	Intervallo
Rand	$R = \frac{a + d}{M}$	[0,1]
Jaccard	$J = \frac{a}{a + b + c}$	[0,1]
Fowlkes-Mallows	$FM = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}}$	[0,1]
Γ Statistics	$\Gamma = \frac{M \cdot a - (a + b) \cdot (a + c)}{\sqrt{(a + b) \cdot (a + c) \cdot (M - a - b) \cdot (M - a - c)}}$	[-1,1]

3.3.2 Internal/Unsupervised measures

Sono misure che si basano sui concetti di **Cohesion** e **Separation**; dati K segmenti della popolazione (C_1, \dots, C_K clusters), possiamo esprimere una misura di **Validity** come segue:

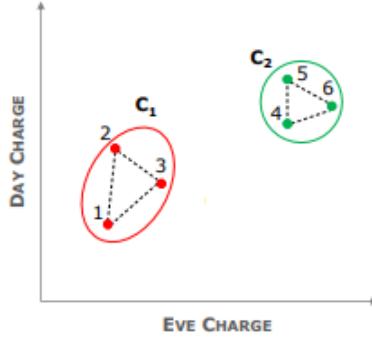
$$Val_{tot} = \sum_{k=1}^K w_k \cdot val(C_i)$$

dove la misura $val(C_i)$ può essere sia di coesione, sia di separazione, e i pesi w_k possono assumere diversi valori (1, cardinalità di C_k, \dots). In generale se si seleziona una misura di coesione, maggiore sarà il valore di Val_{tot} , migliore sarà il risultato.

Nel caso particolare in cui si abbia a che fare con una tipologia *Graph-Based*, una misura di coesione è data dalla somma dei pesi dei collegamenti tra i nodi (osservazioni) del grafo (all'interno del cluster k -esimo).

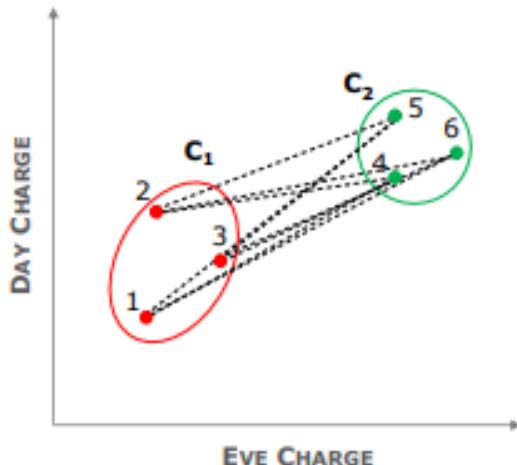
- La *Cohesion* sarà quindi data da:

$$cohesion(C_k) = \sum_{x,y \in C_k} prox(x,y) = \sum_{x,y \in C_k} simil(x,y)$$



- La *Separation* invece:

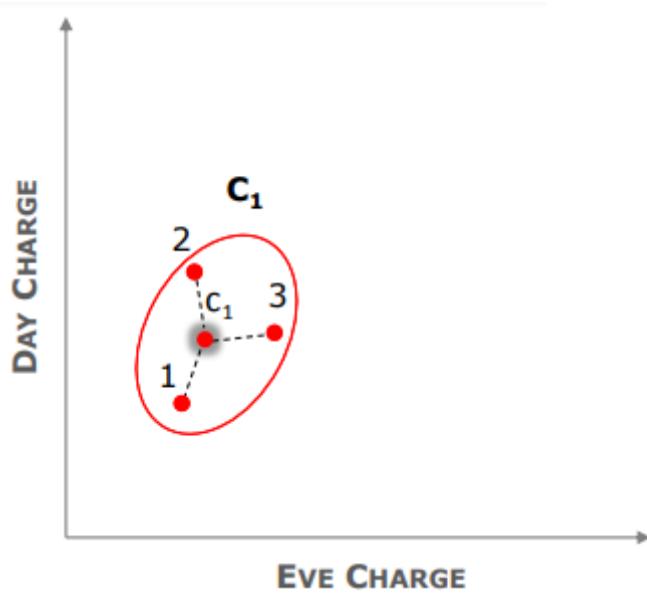
$$separation(C_k, C_j) = \sum_{x \in C_k, y \in C_j} prox(x, y) = \sum_{x \in C_k, y \in C_j} simil(x, y)$$



Nel caso si abbia invece a che fare con un *Prototype-Based clustering*, la *Cohesion* è data dalla somma delle prossimità tra le osservazioni e il *prototype-object* del medesimo *cluster* (centroidi/medoidi)

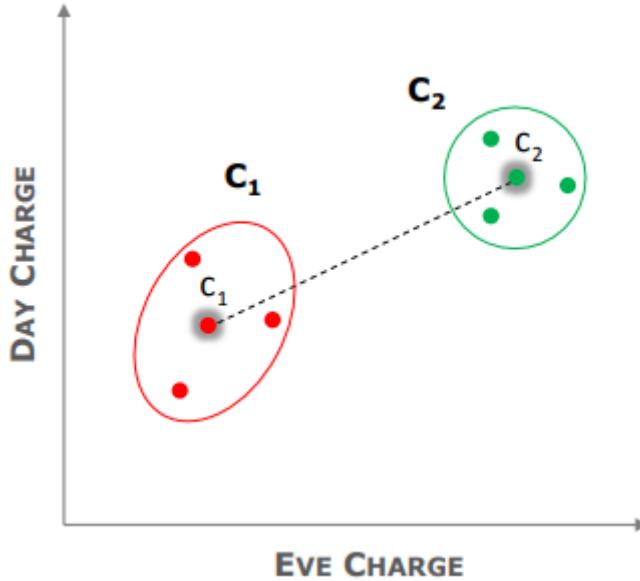
- La *Cohesion* sarà quindi data da:

$$cohesion(C_k) = \sum_{x \in C_k} prox(x, c_k) = \sum_{x \in C_k} simil(x, c_k) \quad \text{con } c_k \text{ centroide di } C_k$$



- La *Separation* invece:

$$separation(C_k, C_j) = prox(c_k, c_j) = simil(c_k, c_j) \text{ con } c_k, c_j \text{ centroidi di } C_k, C_j$$



La scelta dei pesi w_k solitamente avviene come segue:

Tipologia	Misura	Peso
Graph-based	Cohesion	$\frac{1}{m_k}$
Prototype-Based	Cohesion	1
	Separation	m_i

con m_k numerosità del cluster k -esimo.

Una misura aggregata che consente di riassumere i coefficienti di coesione e separazione è il coefficiente di **Silhouette**, il quale viene definito come segue:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, 1]$$

con $i \in C_k$, a_i pari alla distanza media tra la i -esima osservazione e tutte le altre del cluster, mentre b_i la distanza media minima con le osservazioni non appartenenti al cluster. Un valore negativo indica che $a_i > b_i$, il che indica che i non è nel cluster migliore; un valore vicino a 1 ($a_i < b_i$) indica invece una buona situazione. Una buona misura aggregata per la validazione del *clustering* è data dalla **Average Silhouette of all points**.

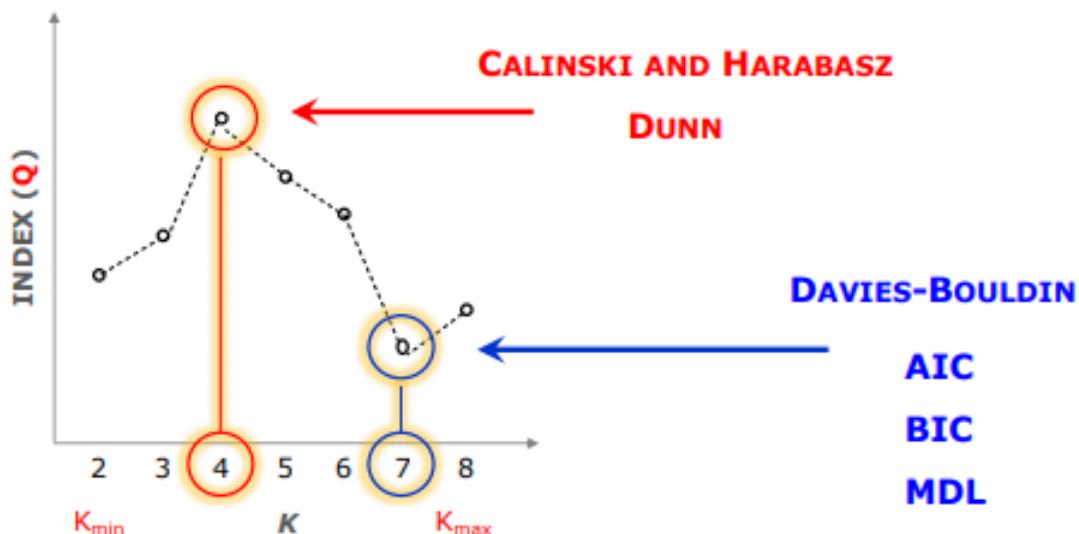
Per il *clustering* gerarchico si utilizza un'altro indicatore, detto **Cophenetic Correlation Coefficient**; misura il grado di similarità tra la matrice delle prosimità (\mathbf{P}) e la matrice cofenetica (\mathbf{Q}). Varia in un range $[-1, 1]$ dove 1 indica il massimo (tranne in caso di utilizzo di *average linkage* in cui un valore alto non garantisce la similarità).

3.3.3 *Relative measures* e il Problema Fondamentale del Clustering

Solitamente i metodi di validazione, siano supervisionati o non supervisionati, sono molto onerosi computazionalmente (inoltre richiedono test statistici). Gli **indici relativi** smorzano in parte tale problematica, e permettono inoltre di fornire possibili soluzioni al **Problema fondamentale del clustering**: scegliere il valore di K ! I principali sono:

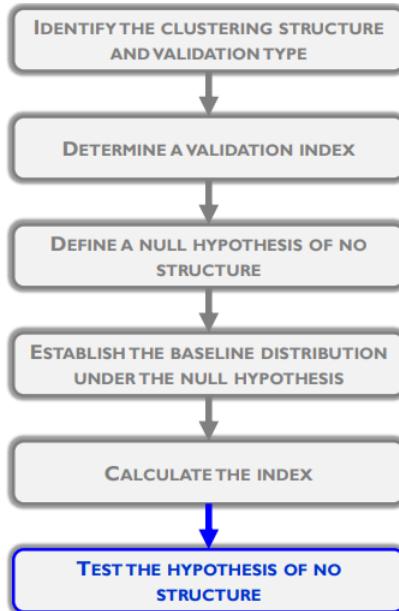
- **Calinski-Harabasz** → il massimo di tale indice ci restituisce valore ottimale di K .
- **Dunn** → stessa interpretazione; inoltre ci dice che in caso di valore abbastanza elevato siamo in presenza di *cluster* omogenei all'interno e eterogenei all'esterno.
- **Davies-Bouldin** → il minimo corrisponde al K ottimale.
- **AIC** → il minimo corrisponde al K ottimale. il minimo corrisponde al K ottimale.
- **Minimum Description Length (MDL)** → il minimo corrisponde al K ottimale. il minimo corrisponde al K ottimale.
- **Bic** il minimo corrisponde al K ottimale. il minimo corrisponde al K ottimale.

Per la selezione del K ottimale (per gli algoritmi che lo richiedono) si crea un ciclo nel quale si fa variare il valore $K \in [K_{min}, K_{max}]$; ad ogni iterazione si calcola un indice di *validity* e si sceglie infine il K associato al miglior risultato.



3.3.4 Validity Paradigm

Consiste nel procedimento mediante il quale si testa l'ipotesi nulla di "Assenza di struttura" all'interno del set di dati; in caso l'ipotesi nulla non venga rigettata, ciò implica che probabilmente applicare un metodo di *clustering* non sia conveniente. I principali passaggi sono:



La definizione di un ipotesi nulla dipende strettamente dal tipo di dati che si hanno a disposizione:

- **Random Position Hypothesis** → si utilizza per dati continui; afferma che tutte le posizioni delle m osservazioni in una specifica regione n -dimensionale sono ugualmente probabili
- **Random Graph hypothesis** → si utilizza per misure di prossimità tra coppie di osservazioni; afferma che tutte le $[m \times m]$ matrici di prossimità (\mathbf{P}) sono ugualmente probabili.
- **Random Label Hypothesis** → si utilizza per tutti i tipi di dati; tutte le permutazioni delle classi sulle m osservazioni sono ugualmente probabili.

Per stabilire una distribuzione sotto le ipotesi nulle appena descritte solitamente si fa uso di 2 metodi: **Bootstrap** e **Analisi Monte Carlo**.

Svolti tutti passaggi sarebbe desiderabile che H_0 (assenza di struttura) venisse rifiutata, altrimenti la definizione dei **cluster** potrebbe risultare insensata.

Analisi e Regole di Associazione

Con **analisi delle associazioni** si intende l'insieme di tecniche volte all'individuazione di **Regole Associative**, ovvero relazioni che intercorrono tra osservazioni: solitamente trova il suo principale impiego nella *Market Basket Analysis*, dove le osservazioni riguardano transazioni, mentre le regole di associazione individuano le relazioni negli acquisti dei prodotti. E' risultata particolarmente utile negli ultimi decenni per lo studio della posizione dei prodotti sugli scaffali dei negozi, e più recentemente, sotto forma di *click stream analysis*, per la strutturazione delle piattaforme di *e-commerce*. Un esempio di analisi delle transazioni, dove vi sono variabili binarie riferite a prodotti acquistabili (con 1 acquistato, 0 non acquistato, viene fornita dalla figura sottostante:

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

Possiamo identificare con $I = \{i_1, \dots, i_p, \dots, i_P\}$ l'insieme di tutti i prodotti da analizzare (attributi binari), mentre con $T = \{t_1, \dots, t_n, \dots, t_N\}$ l'insieme di tutte le transazioni di cui si hanno informazioni (osservazioni dell'analisi); ogni transazione si compone di un certo numero di elementi k , e si definirà quindi **k -itemset** ($k = 1, \dots, P$). Ciascun *itemset* contiene dei *subset* ($k_1, \dots, k_{k-1} < k$), ovvero insiemi di prodotti contenuti nella transazione, e di conseguenza k viene definito *transaction width* (il sottoinsieme più grande, ovvero formato da k elementi).

Un elemento fondamentale per lo studio delle associazioni è il **Support Count**, ovvero un indicatore che rappresenta il numero di volte in cui un determinato *itemset* (X) compare nelle diverse transazioni (T):

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

Si può dare ora una definizione generale di regola associativa:

1. Rappresenta una relazione del tipo $X \rightarrow Y$, dove X viene definito **antecedente** della relazione e Y il **conseguente**.
2. X e Y rappresentano due insiemi di prodotti (**itemset**) disgiunti, ovvero $X \cap Y = \emptyset$.
3. Il **Supporto** di una regola associativa indica quante volte si applica ad un set di dati:

$$S\{X \rightarrow Y\} = \frac{\sigma(X \cup Y)}{N}$$

con N numero di transazioni.

4. La **Confidence** di una regola associativa indica invece quante volte l'*itemset* Y compare nelle transazioni che contengono X :

$$C\{X \rightarrow Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Confidenza e Supporto sono i due indicatori che definiscono la forza di una regola associativa; in particolare:

- In caso *Support* assuma valore basso, si avrebbe una relazione che potrebbe rappresentare una casualità, e ciò renderebbe poco conveniente utilizzare la regola per promuovere determinati *itemset*. Solitamente il supporto si utilizza per eliminare le regole poco interessanti dal punto di vista del *business*, fissando una soglia minima (Sup_{min}) sotto il quale la regola non viene considerata.
- La *Confidence* viene interpretata invece come una misura di affidabilità della regola associativa, e un valore alto indica che il conseguente (Y) rientra spesso nelle transazioni contenenti il conseguente (X); in termini probabilistici misure la probabilità condizionata di Y rispetto a X . Come per il Supporto si fissano una soglia minima ($Conf_{min}$) sotto il quale la regola non viene considerata.

Bisogna precisare che nonostante la relazione antecedente → conseguente, una regola associativa non va interpretata in termini di causalità.

Esistono diversi metodi per valutare Supporto e Confidenza tra *itemset*. La prima è il metodo *Brute-Force*, che come dice il nome stesso consiste nel calcolo dei due indicatori per tutte le possibili coppie di *itemset*; il numero di regole da studiare sarebbe $R = 3^d - 2^{d-1} + 1$, ovvero un procedimento estremamente oneroso in caso di un elevato numero di prodotti. Solitamente risulta più conveniente studiare il problema eliminando gli *itemset* con basso supporto, ovvero sotto una certa soglia, conservando quindi solo i Frequent Itemset; in secondo luogo si studiano le confidenze delle regole tra *frequent itemset*, e un'altra volta si mantengono quelle oltre una certa soglia. Alla fine del procedimento si identificano così le **Strong Rules**.

4.1 Rules Evaluation

Svolto il processo di identificazione delle *Strong Rules*, risulta spesso necessario fornire un *ranking* di quelle che possano risultare le più interessanti dal punto di vista dell'analisi; per fare ciò esistono due criteri differenti, che possono essere basati su argomenti statistici oppure soggettivi.

4.1.1 Criteri soggettivi

Si basano su argomenti non basati su indici, ma su ragionamenti svolti dall'analista o dal committente dell'analisi. Sono spesso volti ad eliminare le relazioni che risultino infondate secondo logica, oppure troppo scontate (pane e burro). La difficoltà nell'applicazione di tali criteri risiede nell'enorme quantità di informazioni a priori di cui bisogna disporre.

4.1.2 Criteri statistici

Spesso è possibile che, durante la fase di analisi, vengano identificate delle regole associative basate su correlazioni spurie, le quali risultano spesso non interessanti. Il processo di valutazione ha quindi lo scopo di definire delle misure di interesse (**Objective Interestingness Measures**), le quali consentano l'individuazione e l'eliminazione dei *pattern* non significativi. Le più utilizzate sono il Supporto, la Confidenza e la Correlazione. Le caratteristiche di tali indicatori sono:

- Sono *data-driven*, ovvero forniscono informazioni ricavabili dai dati a disposizione.
- Non dipendono dal dominio, ovvero sono indipendenti dal settore cui si applica l'analisi associativa.
- Vengono calcolati utilizzando la tabelle delle contingenze.

Un esempio per chiarire tali punti è il seguente: prendendo in analisi il consumo di tè e il consumo di caffè si evidenzia la seguente **Contingency-Table**:

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

Assumendo di analizzare la regola associativa $Tè \rightarrow Caffè$, possiamo identificare alcuni indicatori:

- Supporto: $\frac{150}{1000} = 15\%$
- Confidenza: $\frac{150}{200} = 75\%$

Tali risultati sembrano affermare che i bevitori di tè solitamente consumano anche caffè. Per valutare tuttavia l'interesse della regola bisogna tenere conto che la *Confidence* di una regola può risultare ingannevole: infatti, nonostante l'elevato valore che assume nell'esempio, la confidenza non tiene conto del supporto del conseguente (caffè). Infatti, essendoci 1000 consumatori, la probabilità che uno di essi consumi caffè è del $\frac{1000-200}{1000} = 80\%$, ovvero la probabilità a priori di consumare caffè risulta maggiore della confidenza della regola associativa.

Possiamo ora introdurre alcune misure utilizzate nel processo di *Association Rule Evaluation*:

- La prima è l'**Interest Factor**, ovvero un indicatore capace di risolvere il problema della *confidence* (e del supporto del conseguente). Viene utilizzato per attributi binari e si compone nella forma:

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

$$= 1 \text{ se } A \text{ e } B \text{ sono indipendenti}$$

$$> 1 \text{ se } A \text{ e } B \text{ sono associati pos.}$$

$$< 1 \text{ se } A \text{ e } B \text{ sono associati neg.}$$

Tale indicatore confronta la frequenza del pattern (regola associativa) rispetto alla frequenza di base, sotto ipotesi di indipendenza statistica.

- La seconda è la **Lift**, ovvero il rapporto la la confidenza della regola ed il supporto del conseguente e viene impiegata per gli attributi numerici.

$$Lift = \frac{C(A \rightarrow B)}{s(B)}$$

- La terza riguarda l'**Analisi della correlazione**, in particolare attraverso l'impiego del coefficiente di Pearson per gli attributi continui, e attraverso il **Coefficiente- ϕ** per quanto riguarda i binari (simmetrici).

$$\phi = \frac{f_{11} \cdot f_{00} - f_{01} \cdot f_{10}}{\sqrt{f_{1+} \cdot f_{+1} \cdot f_{0+} \cdot f_{+0}}} \in [-1, 1]$$

ϕ assume valore 0 in caso di indipendenza statistica. Per ciò che riguarda i binari asimmetrici si utilizza la **IS Measure**, data da:

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

la quale, in caso di indipendenza statistica, risulta uguale a $\sqrt{s(A) * s(B)}$. Condivide molte delle problematiche della *confidence*, tra le quali la possibilità di assumere valori elevati anche in caso di incorrelazione.

Esistono altre due tipologie di indicatori per analizzare la qualità d una regola associativa, le quali vengono chiamate **Objective Measures**:

- **Simmetriche:** un indicatore M viene definito simmetrico se $M(A \rightarrow B) = M(B \rightarrow A)$. L'*interest factor* è una misura simmetrica ($I(A, B) = \frac{s(A, B)}{s(A) \cdot s(B)} = \frac{s(B, A)}{s(A) \cdot s(B)} = I(B, A)$)
- **Asimmetriche:** un indicatore M viene definito asimmetrico se $M(A \rightarrow B) \neq M(B \rightarrow A)$. La *Confidence* è una misura asimmetrica ($C(A \rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)} \neq \frac{\sigma(B \cup A)}{\sigma(B)} = C(B \rightarrow A)$)

Le principali misure, distinte in simmetriche ed asimmetriche, vengono riassunte di seguito:

SYMMETRIC

- Correlation (ϕ)
- Odds ratio
- Kappa
- Interest (I)
- Cosine (IS)
- Jaccard
- Collective strength

ASYMMETRIC

- Gini index
- Mutual information
- Certainty factor
- Added value
- J-measure
- Goodman-Kruskal

Alcune indicazioni sulla scelta della giusta misura obiettivo, al fine della valutazione delle regole associative, son:

- Coefficiente- ϕ , *Odds-Ratio*, Kappa e *Collective Strength* non sono solitamente impiegati per gli attributi binari asimmetrici; *Interest*, Coseno (IS) e Jaccard invece possono essere impiegati a tale scopo.
- Coseno e Jaccard sono spesso impiegati per la *Document Analysis* e per la **Market Basket Analysis** (MBA); le altre misure simmetriche no.

Nonostante siano stati analizzati solo i casi in cui si ha a che fare con misure riguardanti attributi binari, è possibile definire alcuni indicatori riferiti a tabelle di contingenza multidimensionali:

C	B	not B	
A	f_{111}	f_{101}	f_{1+1}
not A	f_{011}	f_{001}	f_{0+1}
	f_{+11}	f_{+01}	f_{++1}

not C	B	not B	
A	f_{110}	f_{100}	f_{1+0}
not A	f_{010}	f_{000}	f_{0+0}
	f_{+10}	f_{+00}	f_{++0}

Dove l'*Interest Factor* assumerebbe la forma:

$$I_3(A, B, C) = \frac{N^{3-1} \cdot f_{111}}{f_{1++} \cdot f_{+1+} \cdot f_{++1}}$$