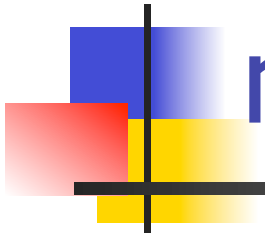


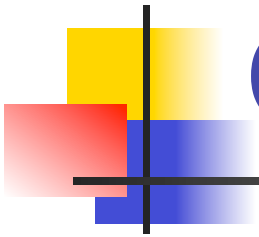
Il ruolo del Machine Learning nel Data Mining

Intelligenza Artificiale 2

Claudio Biancalana

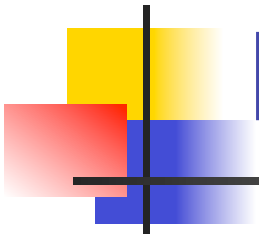
claudio.biancalana@dia.uniroma3.it

http://www.dia.uniroma3.it/~biancal



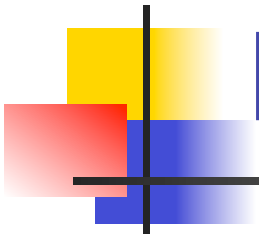
Obiettivi

- Comprendere le funzionalità e gli ambiti applicativi del Data Mining
- Comprendere il ruolo del Machine Learning nel Data Mining
- Regole di associazione
- Alberi di decisione



Dal warehousing al mining

- La maggior parte delle aziende dispone di enormi basi di dati contenenti dati di tipo operativo
- Queste basi di dati costituiscono una potenziale miniera di informazioni utili
- Nei sistemi DBMS e DW attuali le possibilità di estrarre conoscenza sono limitate



Data Mining

- Processo di estrazione di conoscenza da banche dati di grandi dimensioni tramite l'applicazione di algoritmi che individuano le associazioni "nascoste" tra le informazioni (pattern) e le rendono visibili.
- I pattern devono essere:
  - Validi
  - Precedentemente sconosciuti
  - Potenzialmente utili
  - Comprensibili

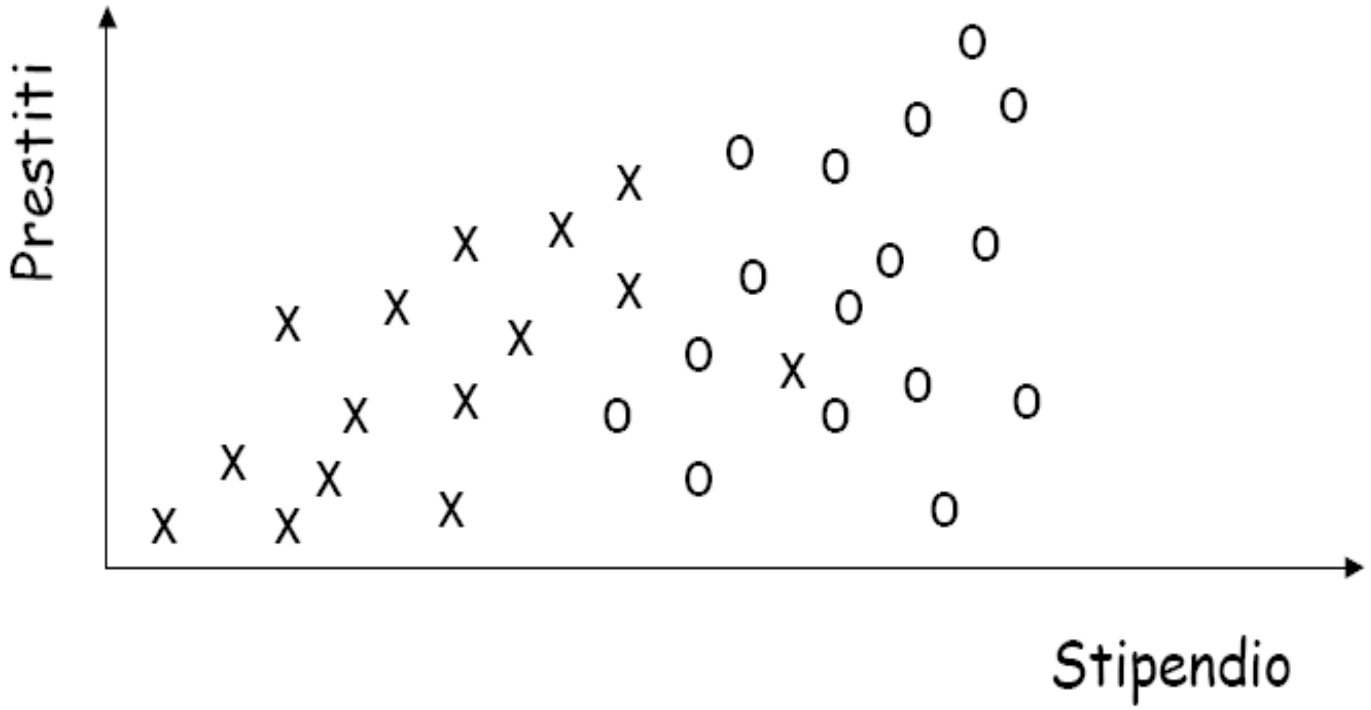
Data Mining Esempio

Record di vendite

id transazione	id cliente	prodotti comprati
T1	cust33	p2, p5, p8
T2	cust45	p5, p8, p11
T3	cust12	p1, p9
T4	cust40	p5, p8, p11
T5	cust12	p2, p9
T6	cust12	p9

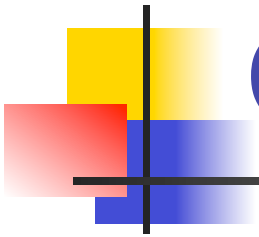
- Osservazioni:
  - I prodotti p5 e p8 vengono spesso comprati
  - Al Cliente 12 piace il prodotto p9

Data Mining: Esempio

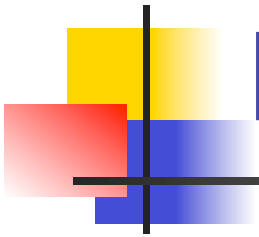


- Persone che hanno ricevuto un prestito in banca
  - X: persone che hanno mancato la restituzione di rate
  - O: persone che hanno rispettato le scadenze

Confronto tra tecniche



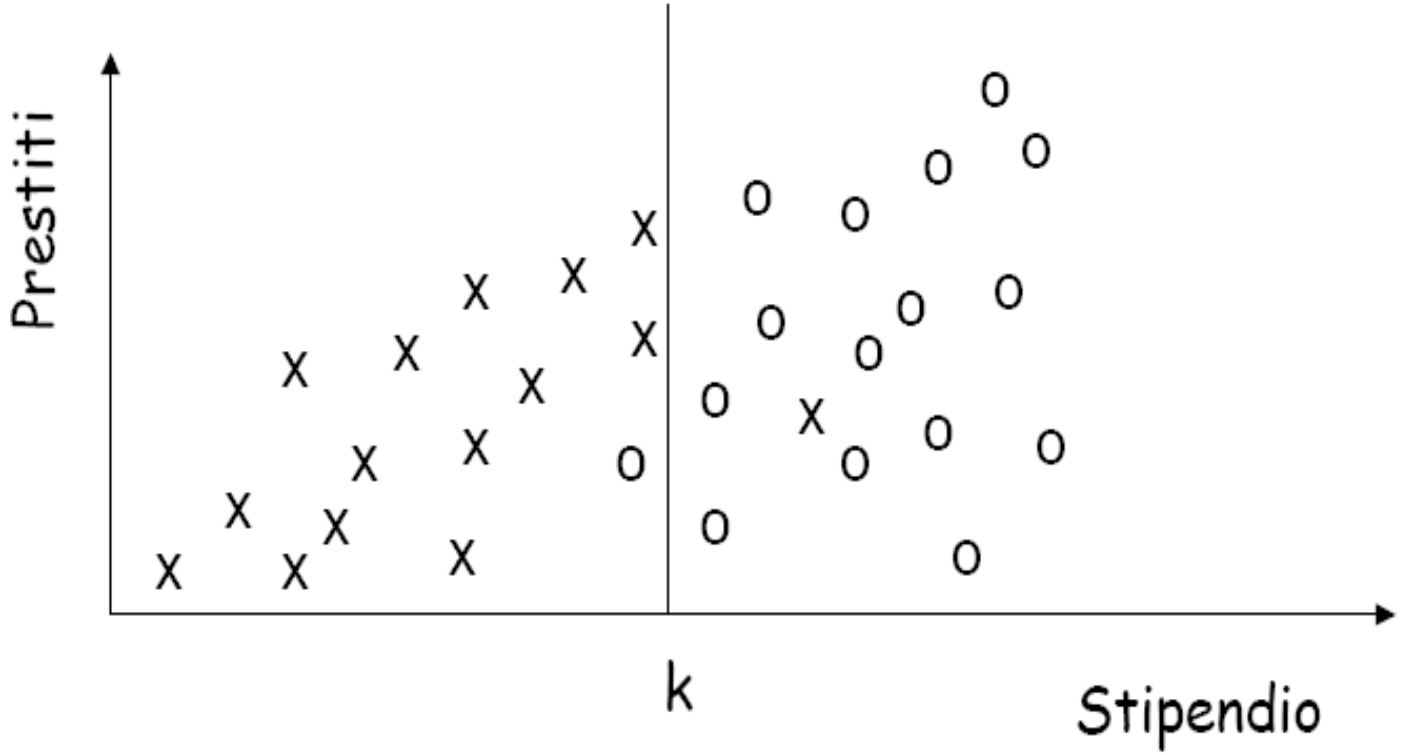
- Retrieval
  - Quanti clienti hanno età compresa tra 40 e 50 anni e comprano Diet Coke
  - Quali documenti contengono la parola "Sanità"
  - Quali brevetti ha depositato la società Colgate nel 1995
- Mining
  - Quali sono le caratteristiche dei miei clienti
  - Quali sono gli argomenti trattati da un insieme di documenti
  - Quali sono i miei concorrenti e come evolve la loro attività



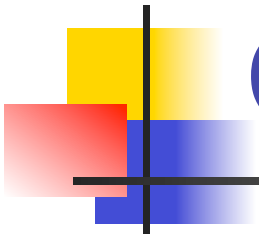
Kowledge Discovery

- Un processo KD si basa sui seguenti elementi:
  - DATI: insieme di descrizioni non interpretate di eventi contenute in una base di dati o in un data warehouse
  - PATTERN: espressione in un linguaggio opportuno che descrive in modo sintetico le informazioni estratte dai dati
    - Regolarità
    - Informazione di alto livello

ESEMPIO

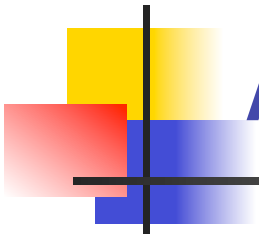


`IF` stipendio < k `THEN` mancati pagamenti



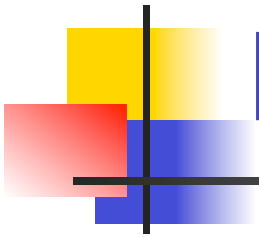
Caratteristiche dei pattern

- Validità:
  - I pattern scoperti devono essere validi su nuovi dati con un certo grado di certezza
    - Esempio: spostamento a destra del valore k porta la riduzione del grado di incertezza
- Novità:
  - Rispetto a variazioni dei dati o della conoscenza estratta
- Utilità:
  - Esempio: aumento di profitto atteso dalla banca associato alla regola estratta
- Comprensibilità:
  - Sintattica
  - Semantica



Applicazioni

- Text categorization
- Approvazione di prestiti e crediti
- Analisi degli acquisti
- Segmentazione di mercato
- Profilazione dei clienti
- Applicazioni finanziarie
- Commercio elettronico
- Rilevazione di fondi



Esempi di tecniche

- Regole di associazione
  - Il 99% dei clienti che acquistano pannolini di sabato acquista anche birra
- Classificazione
  - Le persone sotto i 25 anni che guadagnano meno di 20K€ sono cattivi debitori
- Sequenze simili
  - I DNA di A e B sono simili
- Rilevazione di outliers
  - Questa connessione è in realtà un attacco (Intrusion Detection Sysytem alias IDS)

Stili di Data Mining

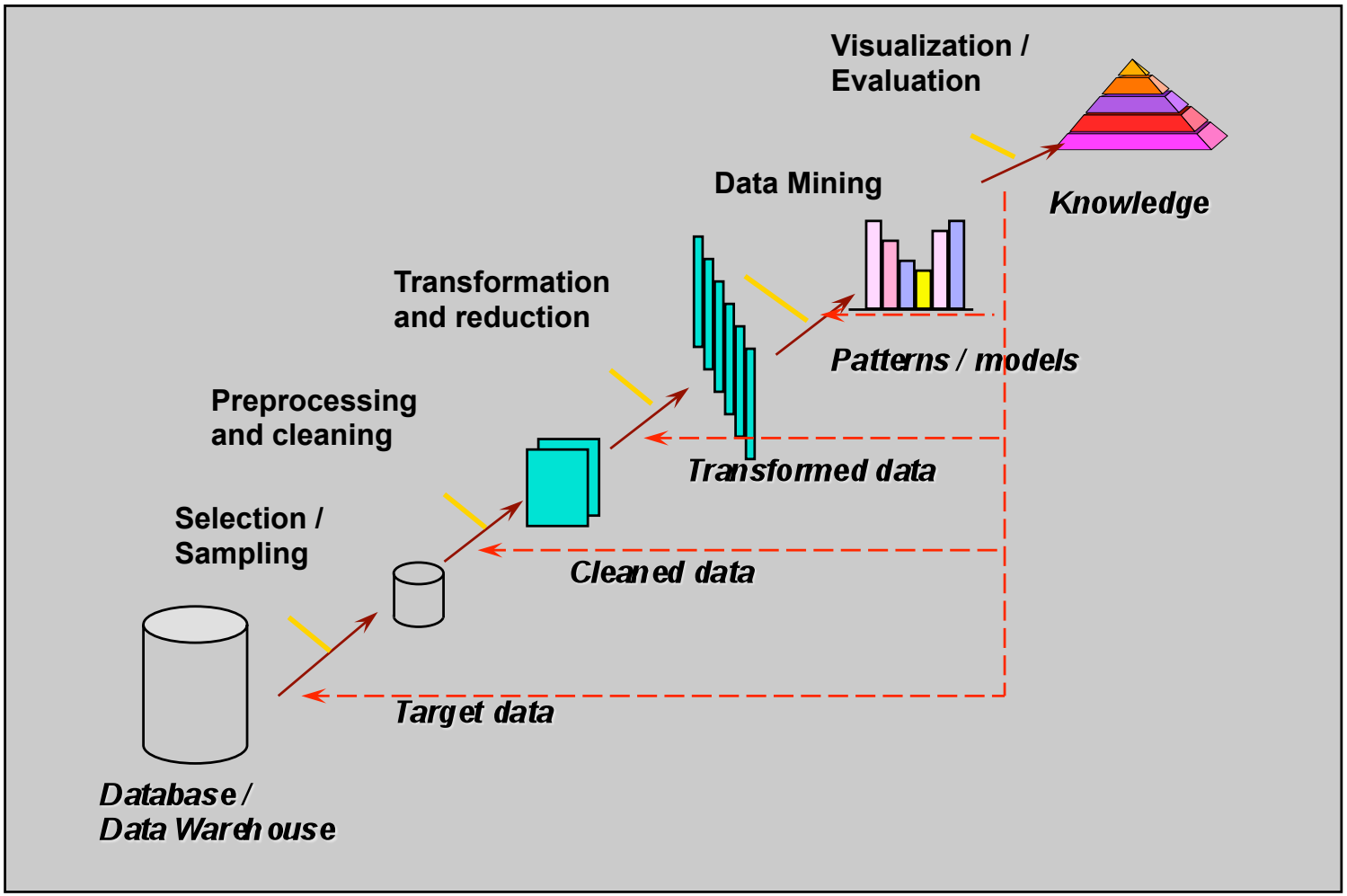
- **Top-Down**
  - Quanto sappiamo cosa stiamo cercando
- **Bottom-up**
  - Lasciare i dati liberi i parlare

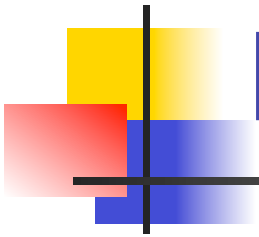
Trovare i pattern → L'utente decide se sono più o meno rilevanti

Non sono mutuamente esclusivi

Entrambi richiedono l'intervento umano!

Processo di estrazione (KDD)



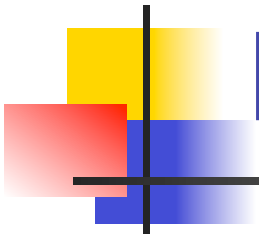


KDD

1. ***definizione e comprensione del dominio applicativo***: individuare le effettive problematiche di business e gli obiettivi da realizzare;

2. ***creazione di un target data set***: selezione di un sottoinsieme di variabili e di dati o di un campione dei dati;

3. ***data cleaning e pre-processing***: operazioni per attenuare il rumore nei dati, o degli outlier, selezione delle informazioni necessarie per generare il modello; decisioni sul trattamento dei campi mancanti o incompleti, dei dati rari (con un'eventuale sovra-campionatura) sulla definizione della storicità e sull'aggiornamento dei dati; aggiunta di variabili derivate e indicatori che hanno valori ricavabili da dati già esistenti.

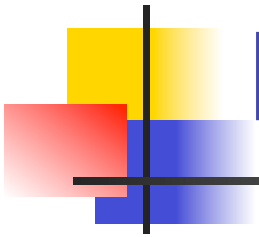


KDD

***4. data reduction e projection*:** definizione della modalità di rappresentazione dei dati secondo gli obiettivi posti, utilizzo di metodi per ridurre il numero delle variabili;

***5. scelta del ruolo dei sistemi di data mining per l'analisi*:** utilizzo dei sistemi di data mining per classificazione, regressione, clusterizzazione, etc.

***6. scelta del o degli algoritmi di data mining*:** selezione dei metodi per la ricerca di pattern, decidendo quali modelli o parametri possono essere più appropriati, integrazione dei metodi di data mining scelti con l'intero processo di scoperta della conoscenza;

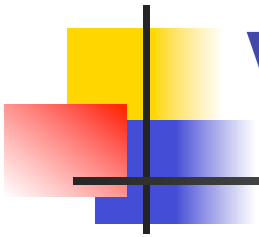


KDD

7. data mining**:** ricerca di modelli di interesse per l'utente, con raffinamenti successivi, presentati secondo definite modalità di rappresentazione (classificazione, alberi di decisione, regressione, cluster analysis…)

8. interpretazione dei modelli identificati**:** analisi e verifica dei risultati con possibile retroazione ai punti precedenti per ulteriori iterazioni al fine di migliorare l'efficacia dei modelli trovati;

9. consolidamento della conoscenza scoperta**:** integrazione della conoscenza e valutazione delle performance del sistema, mettendo a confronto i risultati con l'effettivo andamento nella realtà dei fatti e produzione della documentazione agli utenti finali o a terze parti interessate.

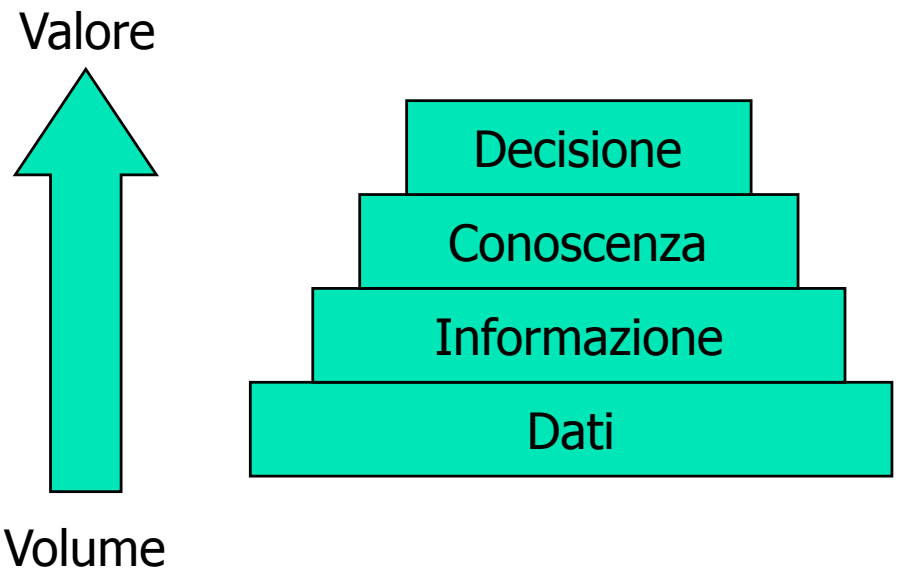


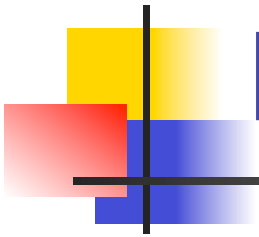
Vantaggi del DataMining

- Trattamento di dati quantitativi, qualitativi, testuali, immagini e suoni;
- Non richiede ipotesi a priori da parte del ricercatore;
- Non richiede ipotesi sulla forma distributiva delle variabili;
- Possibilità di elaborare un numero elevato di osservazioni;
- Possibilità di elaborare un numero elevato di variabili;
- Algoritmi ottimizzati per minimizzare il tempo di elaborazione;
- Semplicità di interpretazione del risultato;
- Visualizzazione dei risultati.

Perchè DataMining?

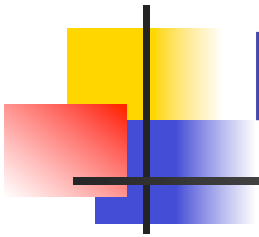
- Quantità dei dati
- Natura dei dati
- Rapida evoluzione del mercato
- Inadeguatezza degli strumenti tradizionali





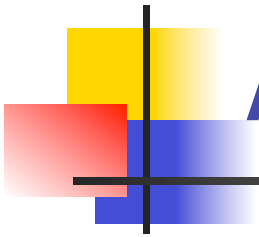
Machine Learning

- Definizione1:
  - Learning is constructing or modifying representations of what is being experienced [Michalski 1986]
- Definizione2:
  - Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time [Simon 1984]



Machine Learning

- Apprendere = migliorare la capacità di esecuzione di un certo compito, attraverso l'esperienza
  - Migliorare nel *task* T
  - Rispetto ad una misura di prestazione P
  - Basandosi sull'esperienza E
  - **E** = esempi di comportamenti "positivi" o "negativi" forniti da un istruttore, oppure un sistema di "ricompense".
- In A.I. : Tecniche che consentono ad agenti (o sistemi automatici) di migliorare il proprio comportamento attraverso l'analisi delle proprie esperienze.



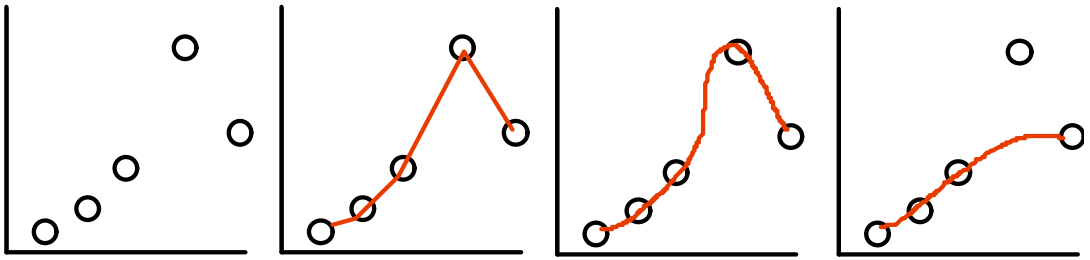
Apprendimento induttivo

- Il sistema parte dai fatti e dalle osservazioni derivanti da un insegnante o dall'ambiente circostante e le generalizza, ottenendo conoscenza che, auspicabilmente, sia valida anche per casi non ancora osservati (induzione). Due tipi di apprendimento induttivo:
  - **Apprendimento da esempi**: la conoscenza è acquisita a partire da un insieme di esempi positivi che sono istanze del concetto da imparare e di esempi negativi che sono non-istanze del concetto.
  - **Apprendimento di regolarità**: non c'e' un concetto da imparare, l'obiettivo e' quello di trovare regolarità (ovvero caratteristiche comuni) nelle istanze date.

Apprendimento induttivo

Problema:

In generale possono esserci molte ipotesi per la funzione obiettivo, alle quali è possibile assegnare, al di là del semplice criterio di consistenza con gli esempi, un grado di preferenza detto ***inclinazione***



Approccio Incrementale: Si cerca di modificare l'ipotesi corrente ad ogni nuovo esempio fornito

Definizione del problema

Dato:

- Un insieme di **osservazioni** (esempi, istanze) $x \in X$
  (es: x *feature vector):* x: $(a_1, a_2 .. a_n)$, a_i assume valori su $\{0,1\}$)
- Una **funzione obiettivo** o *concetto* da apprendere (indicata con ***f***)
  (es. valutazione dell'interesse di una pagina web per un utente):
  Interessante: X➔$\{0,1\}$ dove X è l'insieme delle rappresentazioni feature-vector
- Uno **spazio di ipotesi** H (che dipende dalla modalità di rappresentazione di f)
- Un **insieme di apprendimento** D:

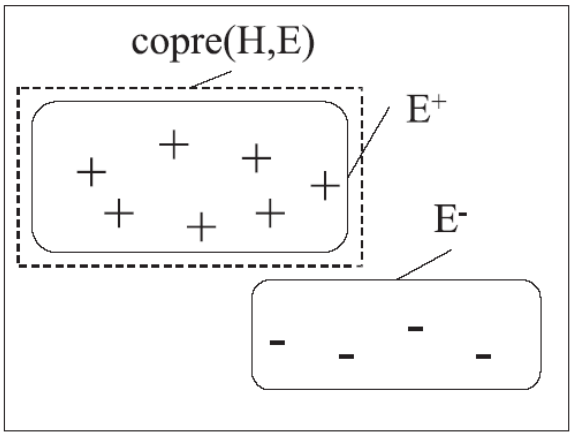
  D: $<(x1, f(x_1)) \ldots (x, f(x_m))>$

  Dove $f(x_i)=1$ se x_i è un esempio positivo di f, 0 altrimenti

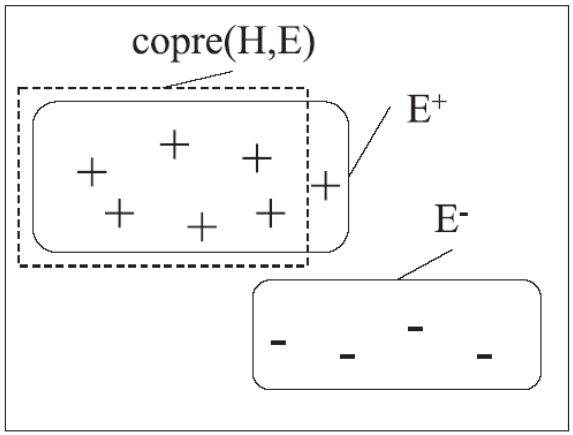
Determinare: un'ipotesi *h* in H tale che $h(x) = f(x)$ $\forall x$ in D

Completezza e constistenza

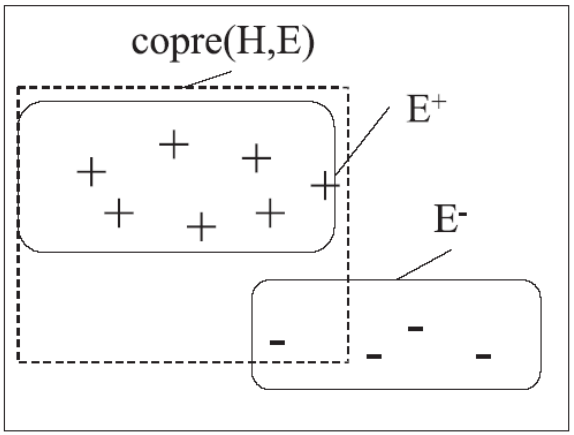
H: completa, consistente



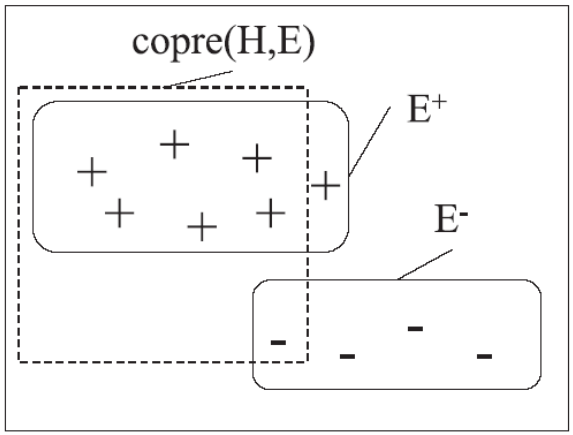
H: incompleta, consistente



H: completa, inconsistente



H: incompleta, inconsistente



Esempi di applicazioni

Esempio	*Tipo di problema*	***Tecnica adottabile***
Quali sono i tre principali motivi che hanno indotto il mio cliente a passare alla concorrenza?	Classificazione	Reti Neurali Decision Tree
Quali sono le fasce di clienti a cui posso offrire nuovi prodotti?	Clustering	Reti Neurali *kohon. Cluster Analysis
Quali sono le probabilità che un cliente ha aperto un c/c acquisterà anche il prodotto x in breve tempo?	Sequencing	Tecniche statistiche Rule induction
Quali sono le probabilità che un cliente acquisti due prodotti completamente differenti?	Associazione	Tecniche statistiche Rule induction
Quale sarà il prezzo del titolo tra un giorno/mese ecc?	Previsione	Reti neurali Tecniche statistiche

Tipi di dati

- Matrice dei dati
  - n oggetti con p attributi

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Il **Data Mining** è la risposta tecnologica all'esigenza di saper analizzare e ricavare conoscenze utili, dalle enormi quantità di dati grezzi che si raccolgono ormai in tutti i contesti operativi della nostra società.

ESTRAZIONE DELLE INFORMAZIONI IMPLICITE

Tipi di dati

- Matrice di dissimilarità
  - d(i,j) misura di dissimilarità tra oggetti i e j
  - d(i,j) = 0 oggetti molto simili

$$\begin{bmatrix} 0 & \cdots & \cdots \\ d(2,1) & 0 & \cdots \\ d(3,1) & d(3,2) & 0 \end{bmatrix}$$

Esempio giocattolo
con tre oggetti

Tipi di dati

- VARIABILI
  - Variabili numeriche
  - Variabili binarie
  - Variabili categoriche nominali
  - Variabili di tipo misto

Variabili numeriche

- Standardizzazione dei dati
  - Per evitare la dipendenza sull'unità di misura scelta per ogni variabile f
- Deviazione assoluta media:

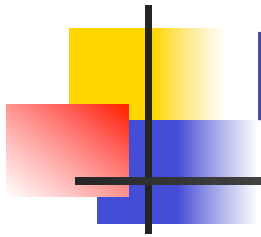
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$$

Misura standardizzata (z-score)

- Standardizzazione dei dati
  - Per evitare la dipendenza sull'unità di misura scelta per ogni variabile f
- Z-score:

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$



Distanza (definizione)

- Proprietà:
  - d(a,b) >= 0
  - d(a,b) = d(b,a)
  - d(a,a) = 0
  - d(a,b) <= d(a,c) + d(c,b)

Distanza e similarità

- Per misurare similarità tra coppie di oggetti spesso si utilizza la distanza di MINKOWSKI:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

  - (q intero positivo)
    - q = 1 distanza di Manhattan
    - q = 2 distanza euclidea

Possiamo pesare le variabili, ottenendo una misura di distanza pesata

Variabili binarie

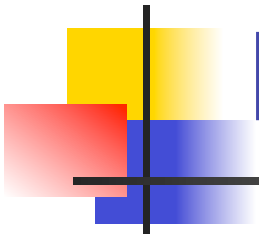
		OGGETTO j		
		1	**0**	**SUM**
OGGETTO i	**1**	q	r	**q+r**
	0	s	t	**s+t**
	SUM	**q+s**	**r+t**	**p**

Coefficiente di matching semplice (variabili simmetriche):

$$d(i,j) = \frac{r+s}{p}$$

Coefficiente di Jaccard (variabili asimmetriche):

$$d(i,j) = \frac{r+s}{q+r+s}$$



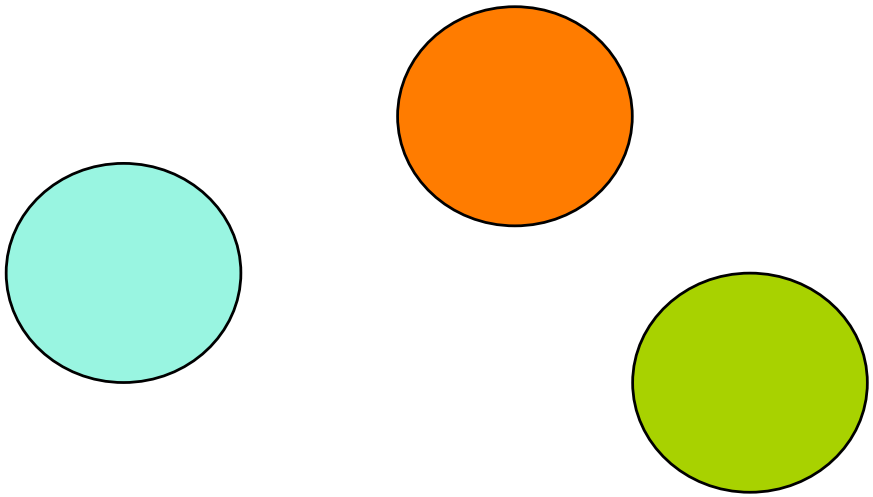
ESERCIZIO

Esempio	Grand.	Colore	Forma	Categoria
1	Piccolo [1]	Rosso [1]	Cerchio [1]	Positivo [1]
2	Grande [0]	Rosso [1]	Cerchio [1]	Positivo [1]
3	Piccolo [1]	Rosso [1]	Triangolo [0]	Negativo [0]
4	Grande [0]	Blu [0]	Cerchio [1]	Negativo [0]

ESERCIZIO: calcolare d(2,3) con JACCARD

TIPI DI CLUSTER

- BEN SEPARATI

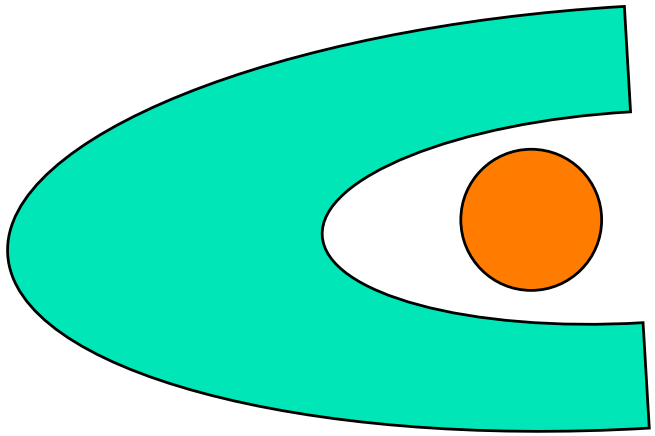


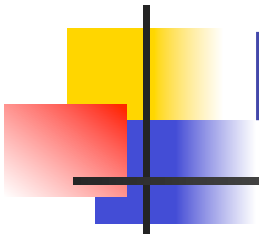
TIPI DI CLUSTER

- CENTER-BASED (mediod vs centroid)

TIPI DI CLUSTER

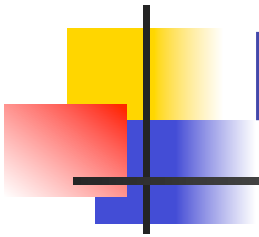
- CONTIGUOUS CLUSTER (NN)





Misure puntuali

- Criterio del legame singolo o del "vicino più vicino": la misura di distanza tra due gruppi è la minima distanza tra tutta le coppie di punti di cui il primo elemento è nel primo gruppo e il secondo nel secondo



Misure puntuali

- Criterio del legame completo o del "vicino più lontano": la misura di distanza tra due gruppi è la massima distanza tra tutte le coppie di punti di cui il primo elemento è nel primo gruppo e il secondo nel secondo.

Matrice di dissmilarità

- Supponiamo di avere un campione di 5 osservazioni su un certo numero di variabili, e di aver scelto una delle distanze descritte per misurare la distanza fra punti.
- Segue la matrice di dissimiglianza:

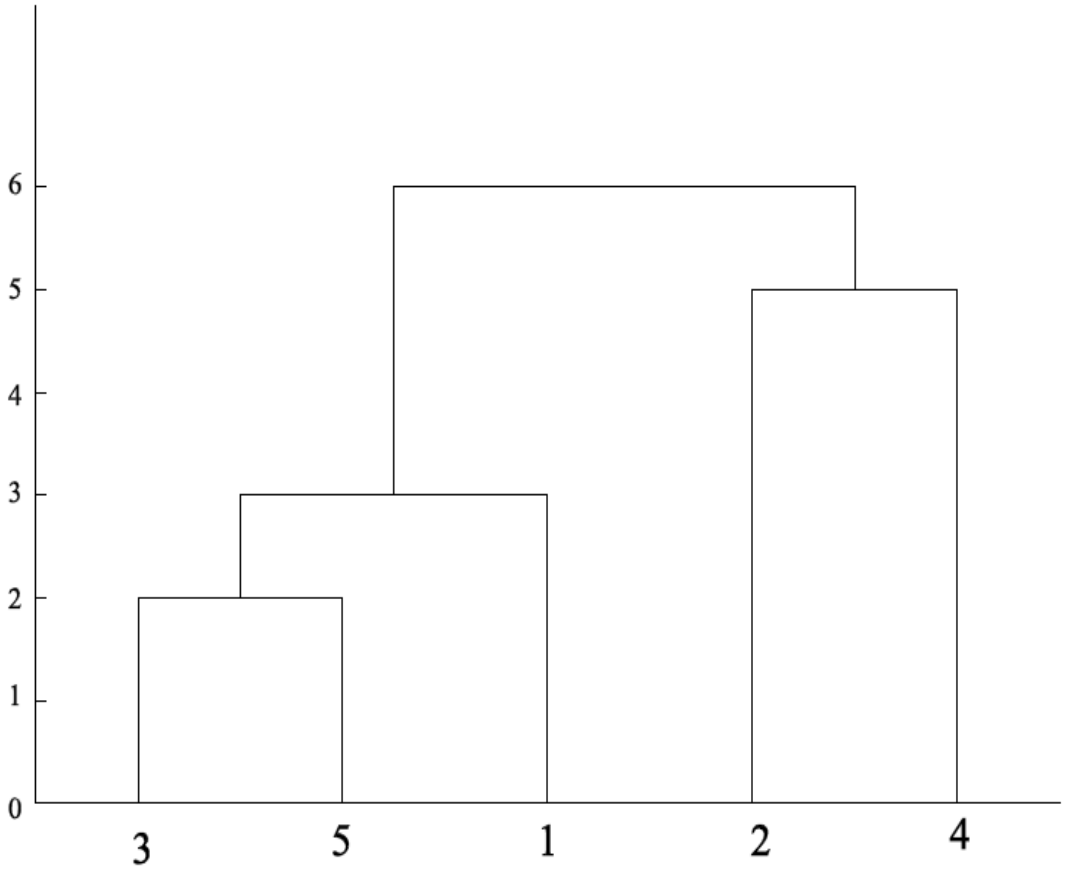
$$\Delta = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

Legame singolo

$$d((3,5);1) = \min\{d(3;1); d(5;1)\} = \min\{3;11\} = 3$$
$$d((3,5);2) = \min\{d(3;2); d(5;2)\} = \min\{7;10\} = 7$$
$$d((3,5);4) = \min\{d(3;4); d(5;4)\} = \min\{9;8\} = 8$$

$$\Delta_1 = \begin{matrix} (3,5) \\ 1 \\ 2 \\ 4 \end{matrix} \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \qquad \Delta_2 = \begin{matrix} (1,3,5) \\ 2 \\ 4 \end{matrix} \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix}$$

Dendrogramma

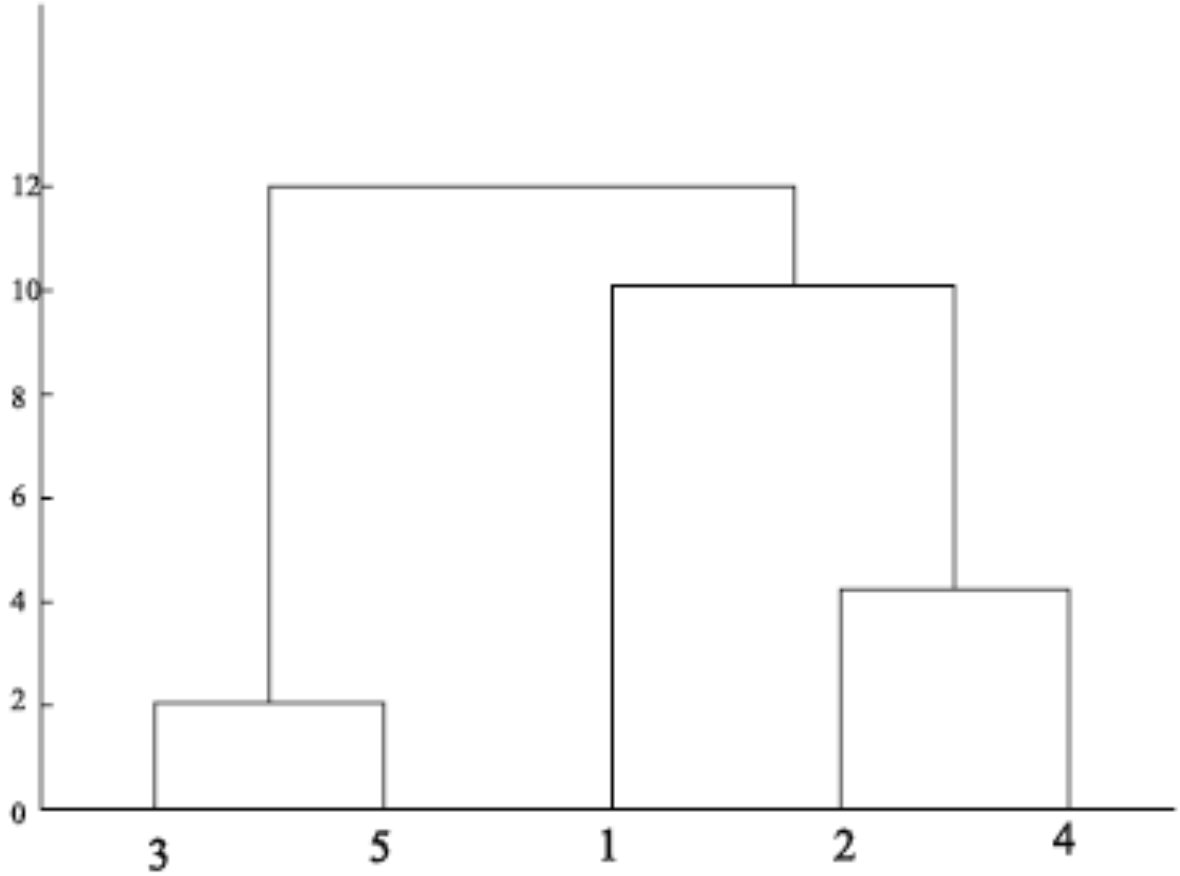


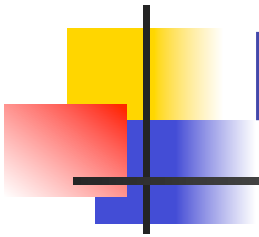
Legame completo

$$d((3,5);1) = \max\{d(3;1); d(5;1)\} = max\{3;11\} = 11$$
$$d((3,5);2) = max\{d(3;2); d(5;2)\} = max\{7;10\} = 10$$
$$d((3,5);4) = max\{d(3;4); d(5;4)\} = max\{9;8\} = 9$$

$$\Delta_1 = \begin{matrix} (3,5) \\ 1 \\ 2 \\ 4 \end{matrix} \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix} \qquad \Delta_2 = \begin{matrix} (3,5) \\ (2,4) \\ 1 \end{matrix} \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & 9 & 0 \end{bmatrix}$$

Dendrogramma



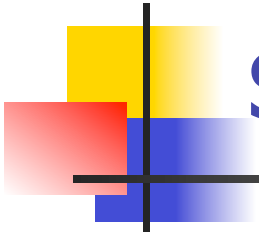


Partitioning Method

Scopo: partizionare il database D di n oggetti in un insieme di k cluster.

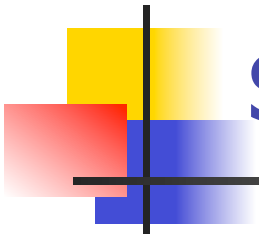
- Apprendimento Non Supervisionato
  - K-MEANS vs K-MEDIOIDS (PAM)
- Apprendimento Supervisionato
  - Reti Neurali (MLP)

Apprendimento non supervisionato

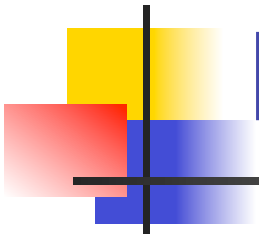


- Etichettare un elevato numero di pattern può essere particolarmente costoso.
- Può essere più conveniente procedere in senso inverso: scoprire le "classi naturali" di appartenenza senza usare le etichette, e poi etichettare solo i gruppi trovati.
- Le caratteristiche dei pattern potrebbero cambiare nel tempo.
- "Scoprire" feature particolarmente significative

Apprendimento non supervisionato



- La classificazione non supervisionata ha lo scopo di trovare raggruppamenti (“cluster”) significativi nei dati
- E’ facile intuire che molta della difficoltà in questo tipo di analisi risiede nel fatto che il concetto di “cluster”, sebbene intuitivo, è molto difficile da definire rigorosamente.

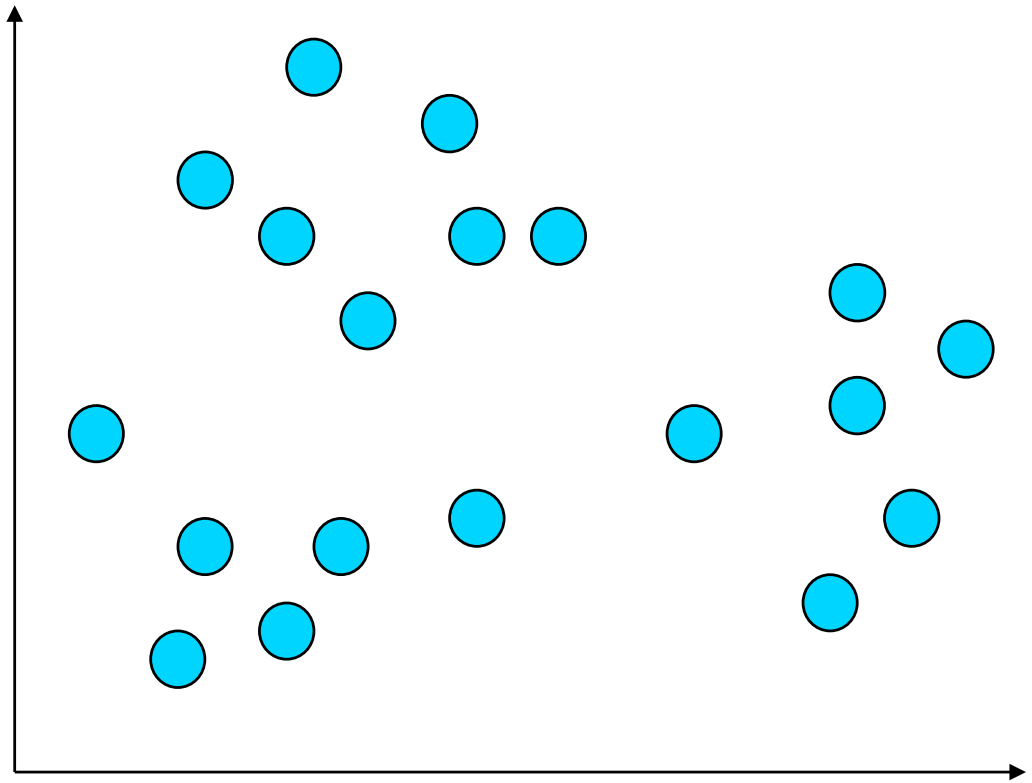


K-Means

1) Fissare il numero *c* di "cluster" da trovare ed una misura di similarità fra due pattern.
   - Inizializzare l'algoritmo definendo *c* centri di "cluster" (si può scegliere a casc *c* punti).
2) Assegna ciascun punto ad uno dei c cluster sulla base della misura di similarità
3) Calcola i nuovi centri del "cluster"
4) Ripeti i passi 2 e 3 fino a quando non si ha alcun "spostamento" dei centri di "cluster"

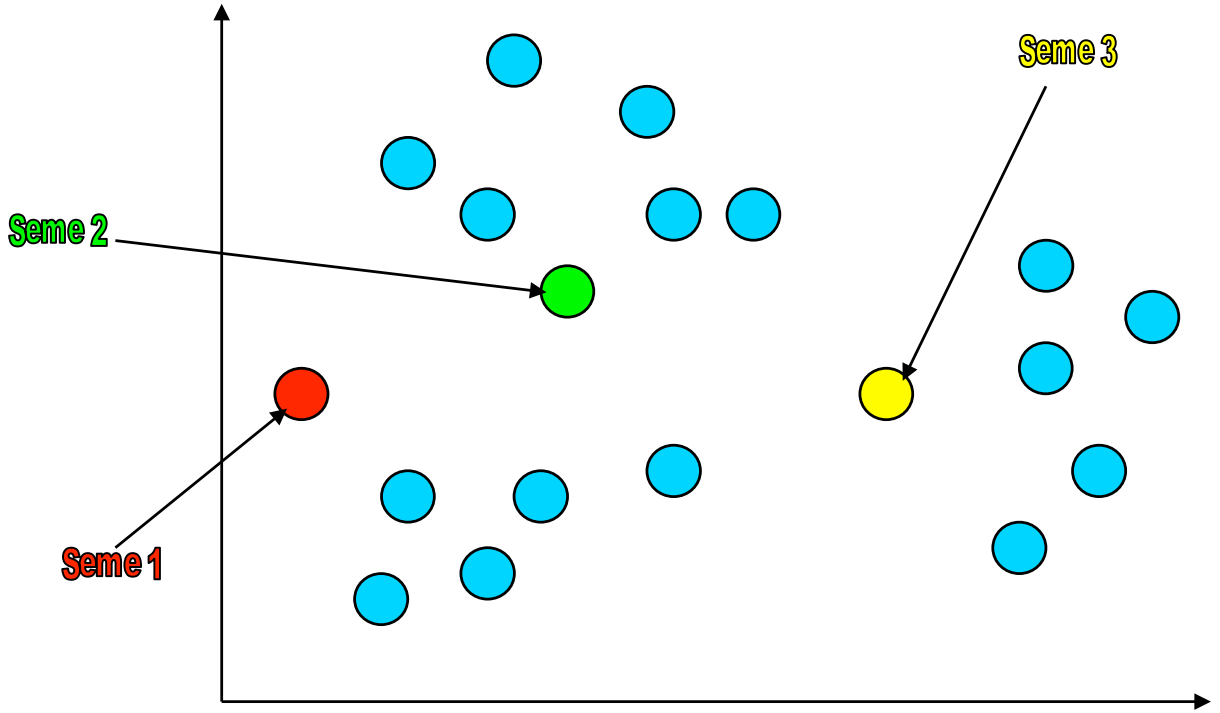
K-Means

Spazio di rappresentazione



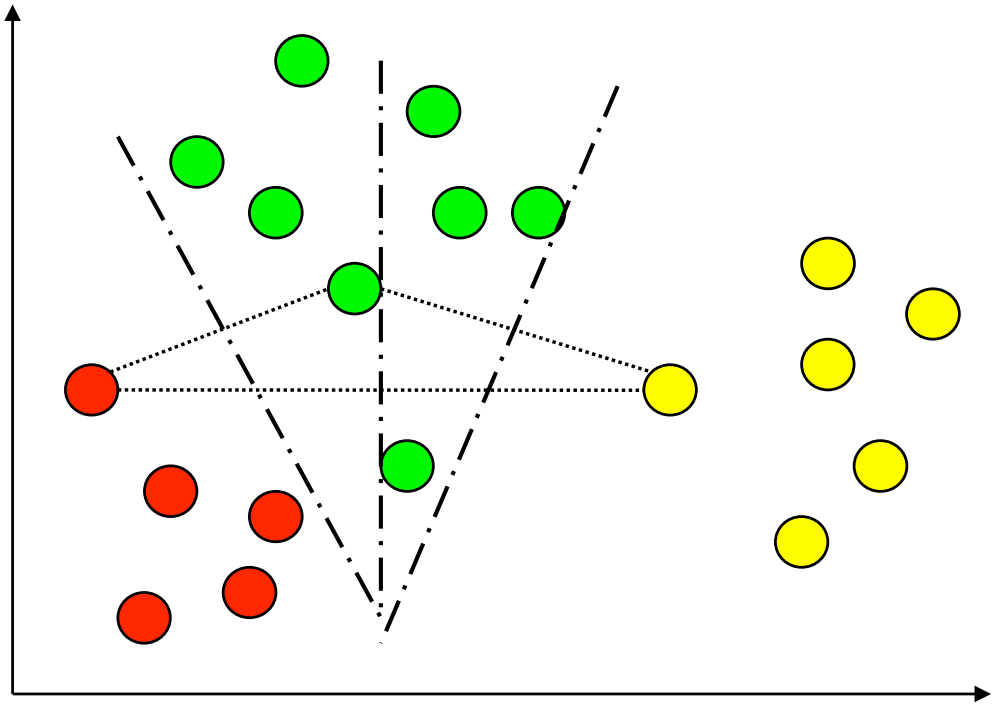
K-Means

Step 1: scegliamo k=3 e i seguenti semi iniziali:



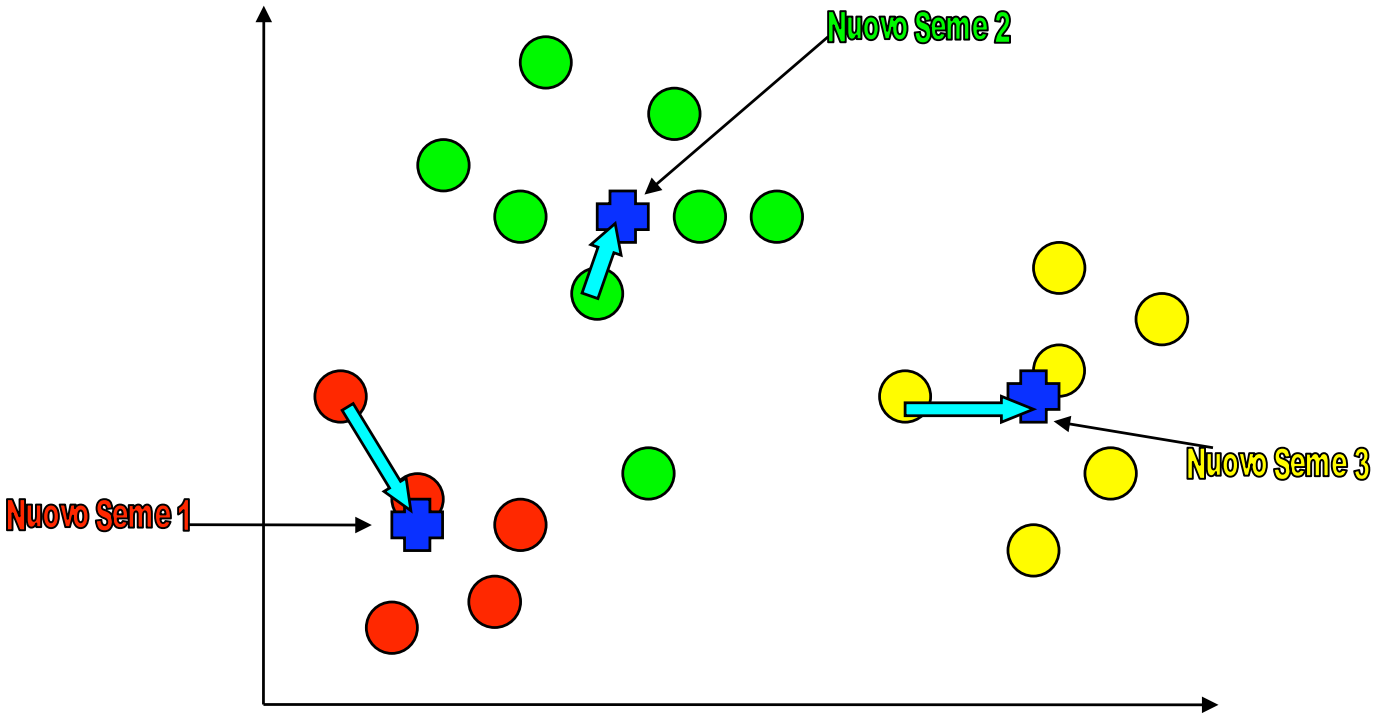
K-Means

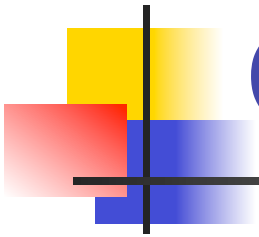
Step 2: assegniamo ogni record al cluster con il centroide (o seme) più vicino



K-Means

Step 3: Il passo due ha individuato nuovi cluster. Ci calcoliamo i centroidi (o semi) di questi





Complessità K-Means

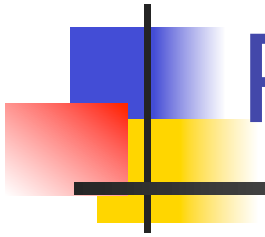
$$O(tknd)$$

n = numero di oggetti

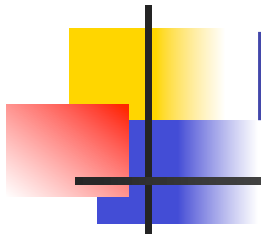
k = numero di cluster

t = numero di iterazioni

d = numero di attributi

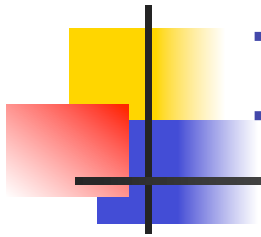


Regole di associazione



Regole di associazione

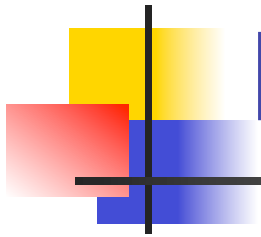
- Caratteristiche
  - Consentono di determinare le regole di implicazione logica presenti nella base di dati, quindi di individuare i gruppo di affinità tra oggetti
  - Tipicamente utilizzate nel marketing per lo studio delle abitudini di acquisto per la pubblicità mirata e l'organizzazione della merce sugli scaffali, e per lo studio della variabilità delle vendite in assenza di un certo prodotto.



I dati del problema

- Insieme di item I
- Transazione T: sottoinsieme di item
- Base di dati D: insieme di transazioni

- Problema:
  - Trovare regole che correlano la presenza di un insieme di prodotti X con un altro insieme di Y



Esempio

- Market-basket analysis:
  - Gli item sono tutti i prodotti venduti da un supermercato
  - Una transazione è l'insieme di oggetti acquistati nella stessa transazione di cassa
  - Si vogliono trovare regole del tipo:
    - Scarpe → calze con confidenza 98%

Supporto e confidenza

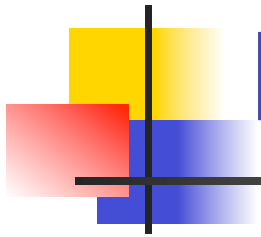
- Regola X➔Y (X,Y contenuti in I)
  - Supporto S, indica la rilevanza statistica:

$$\frac{\#\text{trans. contenenti } X \cup Y}{\#\text{trans. totali}}$$

  - Confidenza C, indica la significatività dell'implicazione:

$$\frac{\#\text{trans. contenenti } X \cup Y}{\#\text{trans. contenenti } X}$$

- Obiettivo: determinare tutte le regole con supporto e confidenza superiori ad una certa soglia



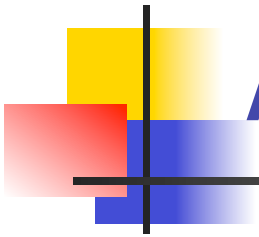
Esempio

- Latte ➔ Uova
  - S = 30%
  - C = 30%
- Supporto
  - Il 30% degli acquisti include entrambi gli elementi
- Confidenza
  - Il 2% degli acquisti di latte includono anche uova

Esempio

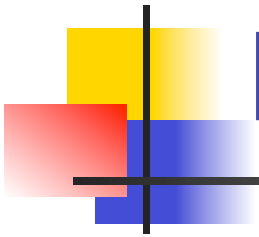
ID Transazione	Prodotti acquistati
2000	A, B
1000	A, C
4000	A, C
5000	B, E, F

- Assumiamo
  - Supporto minimo 50%
  - Confidenza minima 50%
- Regole
  - A ➔ C supporto 50% confidenza 66.6%
  - C ➔ A supporto 50% confidenza 100%



Applicazioni

- Analisi Market-Basket
  - * ➔ uova
    - Cosa si deve promuovere per aumentare il livello delle vendite di uova?
  - latte ➔ *
    - La vendita di quali altri prodotti viene favorita dalla vendita di latte?
  - Dimensioni del problema
    - Oggetti: 10^4, 10^5
    - Transazioni: > 10^6
    - Base di dati: 10-100 GB

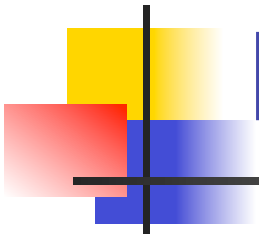


Decomposizione del problema

1. Trovare tutti gli insiemi di item che hanno un supporto minimo

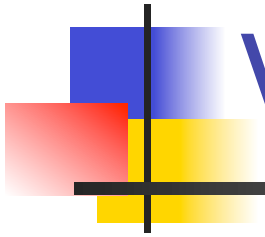
   frequent itemset (FI)

3. Generazione delle regole a partire da FI
4. Algoritmo fondamentale: APRIORI



Regole interessanti

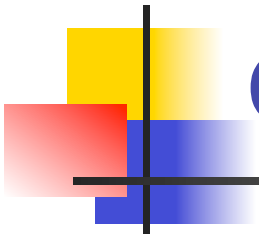
- Non tutte le regole sono interessanti
  - Esempio:
    - Scuola con 5000 studenti
    - Il 60% (3000) gioca a pallacanestro
    - Il 75% (3750) mangia fiocchi a colazione
    - Il 40% (2000) gioca a pallacanestro e mangia fiocchi a colazione
    - Con supporto min 40% e confidenza min. 60%:
      - Gioca a pallacanestro ➔ mangia fiocchi
      - Supporto = 0.4
      - Confidenza = 0.66
    - Regola fuorviante perché il 75% degli studenti mangia fiocchi!



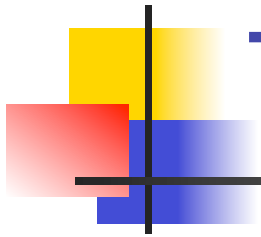
Valutare i classificatori

Best by the Test

Misurare le performance dei classificatori



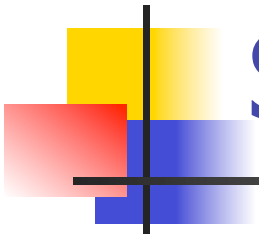
- Una volta ottenuto un classificatore, bisogna stimarne l'accuratezza.
- Ci serve una misura della prestazione del classificatore.
- per i problemi di classificazione, una misura naturale è il tasso di errore.



Tasso di errore vero

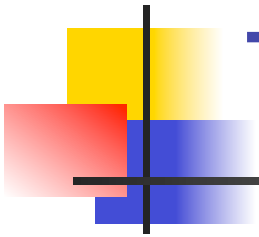
- Supponiamo che
  - I sia l'insieme di tutte le istanze possibili
  - Pr è una distribuzione di probabilità su I
  - c(I) è la classe "vera" di I
  - h(I) è la classe "predetta" di I
  - Si definisce allora il tasso di errore vero (true error rate) come

$$\Pr \{ i \in I \mid h(I) \neq c(I) \}$$



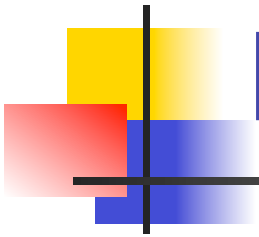
Stima del tasso di errore

- Il modo più semplice di stimare il tasso di errore vero è calcolare il tasso di errore sulle istanze di addestramento
  - si parla di errore di sostituzione
- La stima è troppo ottimistica!!
  - occorre utilizzare due insiemi distinti, uno per l'addestramento e l'altro per il test



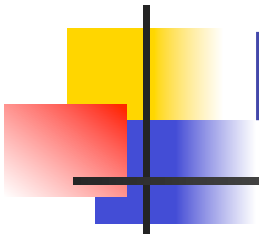
Training, Test, Validation

- E` importante che l'insieme di test non sia utilizzato in alcun modo nella fase di costruzione del modello
  - ad esempio, nella fase di reduced error pruning non va usato l'insieme di test, ma un terzo insieme di dati, indipendente sia da quello di addestramento che da quello di test.



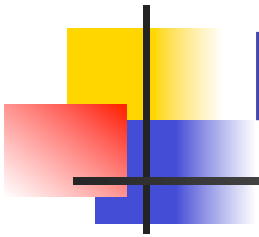
Holdout

- Holdout: metodo con cui si divide l'insieme di dati una parte usata per l'addestramento e una per il testing.
  - tipicamente 1/3 è usato per il test e 2/3 per l'addestramento
- Problema: il campione potrebbe non essere rappresentativo
  - ad esempio, alcune classi poco numerose potrebbero mancare nell'insieme di addestramento



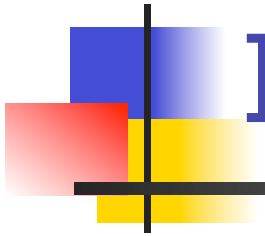
Leave one out estimator

- Vantaggi:
  - fa il massimo uso dei dati a disposizione
  - non ci sono campionamenti casuali
- Svantaggi:
  - computazionalmente oneroso (tranne che per il metodo k-NN)



Boosting e Bagging

- Sono due metodi standard per combinare dei classificatori C 1, ... CT per produrre un classificatore C* più accurato.
  - Analogia con i medici
    - Supponiamo di voler diagnosticare una malattia. Possiamo rivolgerci a vari medici invece che ad uno solo.
      - Bagging: prendo le risposte di tutti i medici e considero come diagnosi valida quella prodotta in maggioranza.
      - Boosting: peso la diagnosi di ogni medico in base agli errori fatti in precedenza.



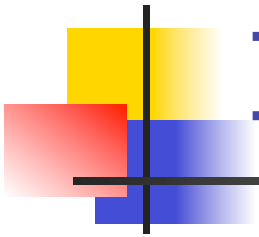
Introduzione a Weka

Analisi dei dati ed estrazione della conoscenza con WEKA

In theory, there is no difference between theory and practice. But, in practice, there is.
Jan L.A. van de Snepscheut (1953-1994), computer scientist and educator

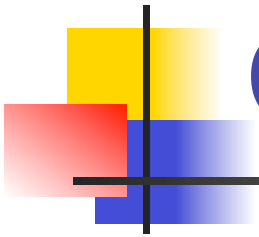
WEKA

- Weka è l'acronimo di Waikato Environment for Knowledge Analysis, ma è anche il nome di un uccello della nuova zelanda…



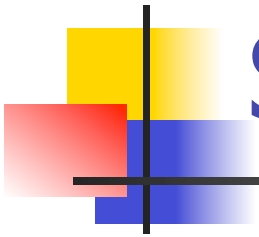
Installazione

- Weka è scritto in Java, quindi si può utilizzare su qualunque sistema operativo dotato di un ambiente di esecuzione Java.
  - Windows
    - Se il JRE (Java Runtime Environment) è già installato, basta scaricare il solo programma di installazione Weka ed eseguirlo.
    - Altrimenti, la cosa più conveniente è scaricare il programma di installazione Weka + JRE che installa entrambi gli ambienti in una volta sola.



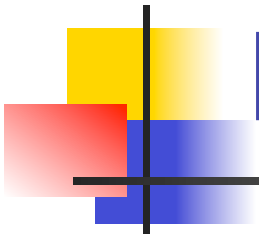
Gli ambienti operativi di Weka

- Una volta lanciato Weka possiamo scegliere tra 4 diversi ambienti operativi:
  - SimpleCLI
  - Explorer
  - Experimenter
  - KnowledgeFlow



SimpleCLI

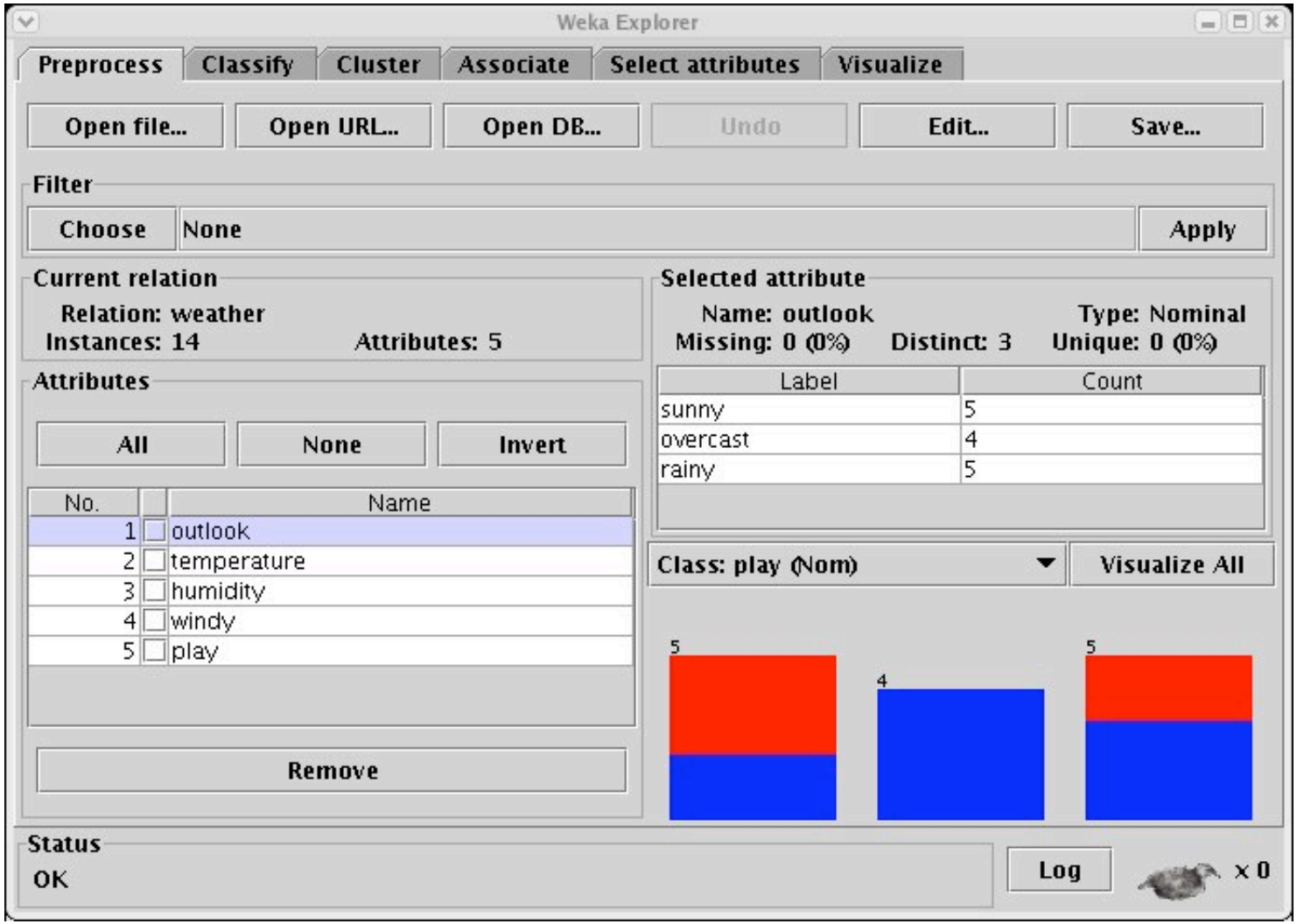
- E’ un ambiente a linea di comando, da usare per invocare direttamente le varie classi Java di cui Weka è composto.
- Tutto quello che si può fare dalla SimpleCLI è possibile farlo anche da un ambiente a linea di comando come il “prompt di DOS” di Windows o la shell di Unix.

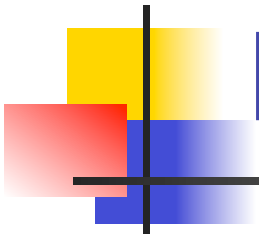


Explorer

- E’ l’ambiente che utilizzeremo più spesso. Con esso si possono caricare degli insiemi di dati, visualizzare in modo grafico la disposizione degli attributi, effettuare una serie di operazioni preliminari di preparazione, ed eseguire algoritmi di classificazione, clustering, selezione di attributi e determinazione di regole associative.

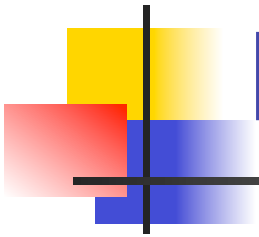
Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather
Instances: 14
Attributes: 5
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
sunny
5
overcast
4
rainy
5
Attributes
All
None
Invert
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0



Experimenter

- E’ una versione batch dell’Explorer. Consente di impostare una serie di analisi, su vari insiemi di dati e con vari algoritmi, ed eseguirle alla fine tutte insieme. E’ possibile in questo modo confrontare vari tipi di algoritmi, e determinare qual è il più adatto a uno specifico insieme di dati.

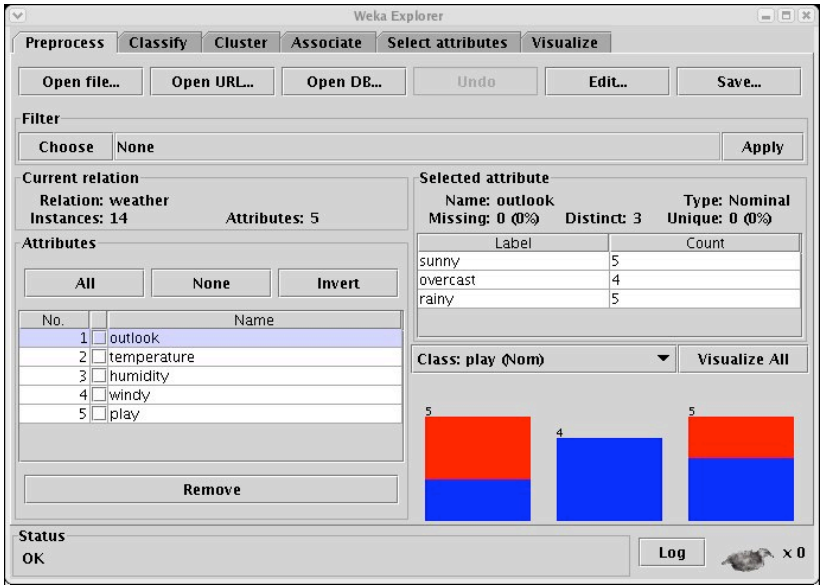


Knowledge Flow

- Una variante dell'explorer, in cui le operazioni da eseguire esprimono in un ambiente grafico, disegnando un diagramma che esprime il "flusso della conoscenza". E' possibile selezionare da una tavolozza varie componenti come sorgenti di dati, filtri, algoritmi di classificazione e collegarli tra loro in un diagramma tipicamente detto "data-flow".

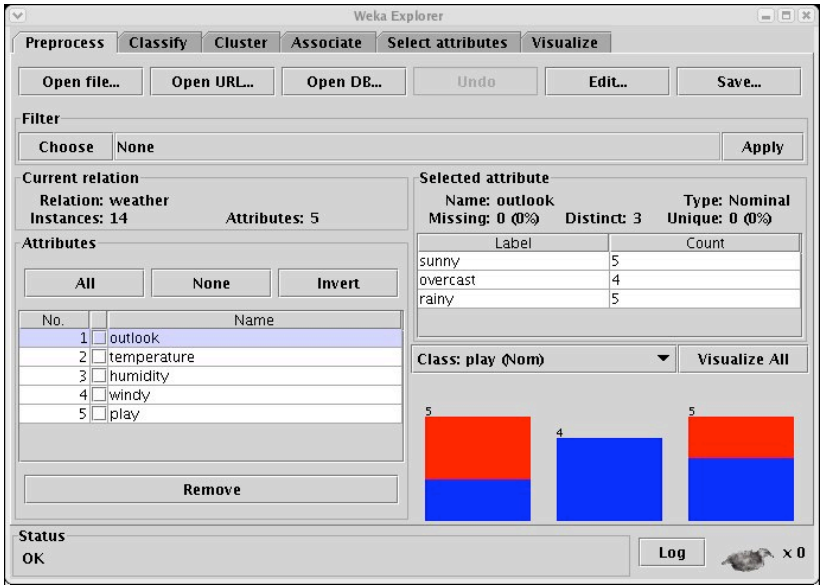
Explorer (1/2)

- Dall'Explorer, per gli attributi nominali abbiamo l'elenco dei possibili valori e, per ognuno di essi, il numero di istanze con quel valore. Interessante anche il conteggio del numero di istanze in cui l'attributo manca e del numero di valori che appaiono una sola volta.

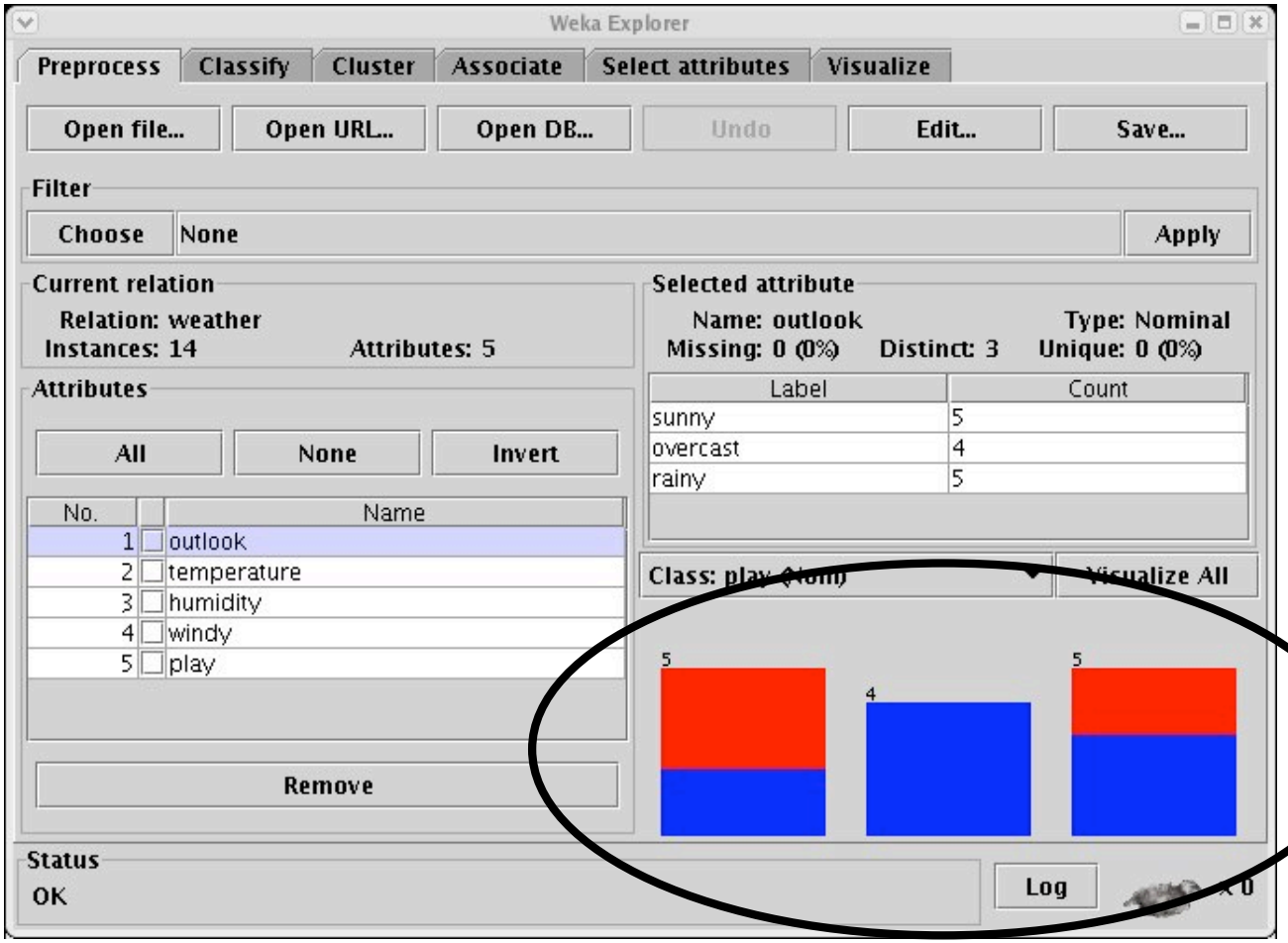


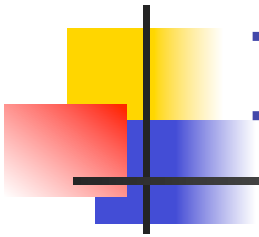
Explorer (2/2)

- Dall'Explorer, per gli attributi numerici, abbiamo le informazioni sul valore massimo, minimo, media e deviazioni standard, oltre alle solite informazioni su numero di valori diverse, numero di valori unici e numero di istanze col valore mancante.



Istogramma riassuntivo





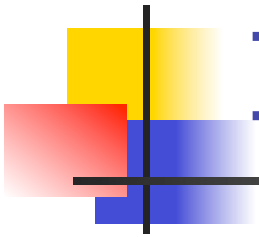
Il formato dati ARFF

- Weka può prelevare i dati da
  - Un file di testo sul computer locale, in formato ARFF
  - Un file su Web in formato ARFF
  - Un database, tramite il driver JDBC

Il formato dati ARFF

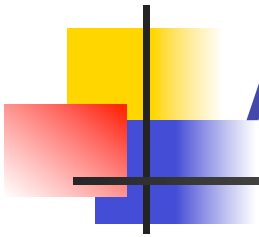
- Un file ARFF è composto da una intestazione e il corpo dei dati vero e proprio.

```
@relation weather

% Relazione weather-data

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
```



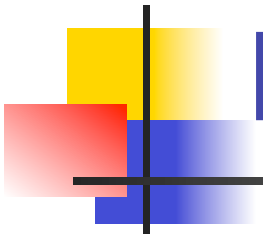
Il formato dati ARFF

- La riga `@relaion weather` specifica un nome per la relazione.
- La riga `"% Relazione weather-data"` è un commento.
- La riga `@attribute outlook {sunny, overcast, rainy}` specifica che il primo attributo è di tipo categoriale e può assumere i valori sunny, overcast e rainy. Il nome dell'attributo è "outlook"
- La riga `@attribute temperature real` specifica che il secondo attributo è di tipo numerico ed ha nome "temperature"
- La riga `@data` indica l'inizio dei dati veri e propri
- La riga `sunny,85,85,FALSE,no` indica che la prima istanza ha valori outlook=sunny, temperature=85, humidity=85, windy=FALSE e play=no.



ARFF per classificare…

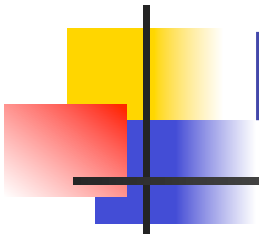
- In generale, ogni volta che serve individuare un attributo particolare come la "classe" dell'istanza (ad esempio per problemi di classificazione), l'ultimo attributo gioca questo ruolo.
- Si può utilizzare il valore ? come dato mancante.



ESERCIZIO

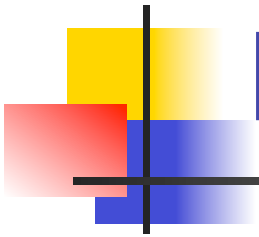
- Modificare i dati del seguente file, in modo da aggiungere un'istanza con attributo outlook mancante e un nuovo valore di outlook che occorre unicamente in una istanza.

```
@relation weather

% Relazione weather-data

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
```



Pre-elaborazione dei dati

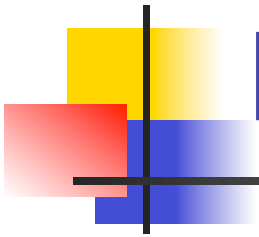
- Dall'Explorer di Weka, tutte le operazioni di pre-elaborazioni si possono eseguire dalla scheda "Pre-process".
- Filtri:
  - Supervisionati:
    - Esiste un attributo speciale, l'attributo classe, che viene usato per guidare le operazioni di filtraggio.
  - Non Supervisionati:
    - Tratta gli attributi allo stesso modo



Filtri supervisionati

- **AddCluster**: aggiunge un nuovo attributo che rappresenta la classe assegnata ad ogni istanza da un algoritmo di clutering (raggruppamento).
  - Esempio: set di dati iris.arff, sugli attributi petalwidth e petallength. Controllare la precisione ottenuta trmite il classificatore J48 dopo la rimozione degli attributi originari, rispetto al set di dati iniziale.
  - Esercizio: cosa succede se si esegue il raggruppamento solo su uno dei due attributi?
- **Discretize**: discretizza un attributo con il metodo dell'equi-widht o equi-depth binning
  - Esempio: set di dati iris.arff, provare i risultati delle due varianti sull'attributo sepallength con 4 intervalli.
- **Normalize**: normalizza col metodo min-max, restringendo tutti gli attributi numerici all'intervallo 0-1.
  - Esempio: set di dati iris.arff.
- **Numeric Transform**: applica una generica funziona matematica a determinati attributi.
- **Replace Missing Values**: rimpiazza tutti i valori mancanti con la moda dell'attributo (se si tratta di attributi nominali) o la media (per attributi numerici)
- **Standardize**: normalizza col metodo z-score tutti gli attributi numerici
- **Resample**: campionamento semplice con rimpiazzamento dei dati

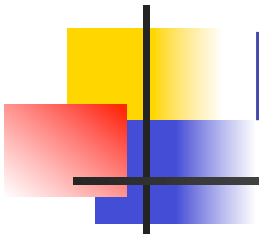
Filtri non supervisionati



- **Discretize**: discretizza gli attributi usando il metodo MDL di Fayyad & Irani's
  - esperimento: confrontare i risultati di precisione (col classificatore J4.8) dei dati originari, dei dati discretizzati con binning e con quest'ultimo metodo del set di dati "segmentation-challenge.arff"

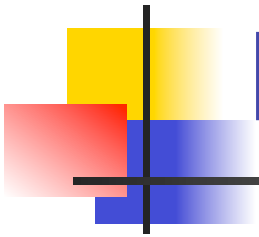
- **Resample**: campionamento con rimpiazzamento dei dati
  - può funzionare come il metodo non supervisionato oppure è in grado di effettuare il campionamento in modo che l'attributo classe abbia una distribuzione uniforme. Ciò avviene settando a 1.0 il valore dell'attributo biasToUniformClass.

  - Esempio: provare sul set di dati soybean.arff



Regole associative

- Weka implementa l'algoritmo Apriori.
- Apriori ha bisogno di alcuni parametri in ingresso:
  - minMetric (soglia di confidenza)
  - numRules (numero di regole da determinare)
- L'algoritmo opera in modo da determinare numRules regole che superano la soglia di confidenza impostata. La soglia di supporto minimo invece non è fissata:
  - Il sistema parte da una certa soglia massima (upperBoundMinSupport) normalmente 1, e scende gradatamente verso una soglia minima (lowerBoundMinSupport) a passi impostati dal parametro delta.

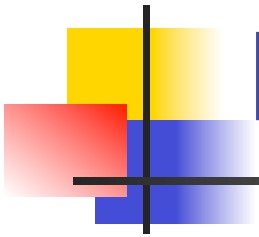


Regole associative

- Applicare Apriori sull'insieme di dati weather-nominal è banale. Con i valori di default (confidenza 90%, supporto da 100% a 10% a passi di 5%, 10 regole da determinare).

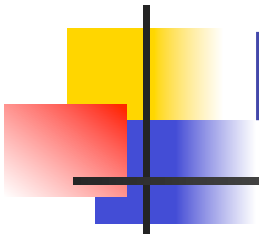
Output Apriori

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 39

Size of set of large itemsets L(4): 6

Best rules found:

 1. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
 3. outlook=overcast 4 ==> play=yes 4    conf:(1)
 4. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
 5. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 7. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
 8. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
 9. temperature=cool windy=FALSE 2 ==> humidity=normal play=yes 2    conf:(1)
10. temperature=cool humidity=normal windy=FALSE 2 ==> play=yes 2    conf:(1)
```



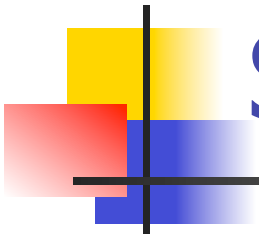
Risultati

- Otteniamo come risultato alcune informazioni come la soglia di supporto minimo finale (15%) e il numero di passi che sono stai necessari per raggiungerla (17). Poi segue il numero di elementi in L(k) (gli itemset più frequenti di k elementi). Infine, l'elenco delle regole, in ordine decrescente di confidenza.



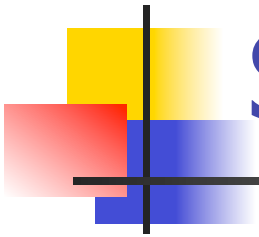
Metriche

- Esistono misure diverse per definire la bontà di una regola oltre la confidenza e il supporto…
  - lift: è il valore di confidenza diviso la percentuale delle istanze in cui è presente nella conseguenza della regola.
    - Ad esempio, nel caso della regola humidity=normal windy=FALSE ➔ play=yes la confidenza è 1, mentre la percentuale di istanze in cui windy=FALSE è 9/14. Dunque lift = 1/(9/14)=14/9=1.56



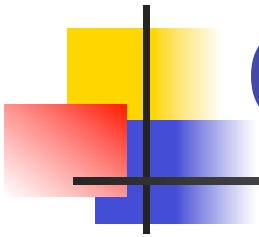
Selezione di attributi rilevanti

- Esistono vari modi per la selezione degli attributi. Alcuni considerano un attributo alla volta e determinano la misura della sua significatività in base alla capacità di discriminare una classe da un’altra. Altri considerano invece una collezione di attributi e ne valutano l’efficacia complessiva.



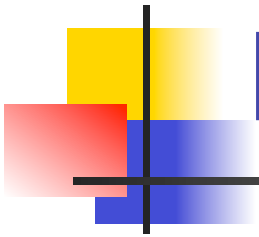
Selezione di attributi rilevanti

- Possiamo prendere come set di dati **zoo.arff** e scegliere come valutatore per gli attributi il meotodo InfoGainAttributeEval che calcola, per ogni attributo, il guadagno di informazione…
- Ricordiamo che
  - InfoGain(Attr)=H(Class) - H(Class|Attr)
  - GainRatio(Attr)=InfoGain(Attr)/H(Attr)



Output

```
=== Attribute Selection on all input data ===

Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 18 type):
    Information Gain Ranking Filter

Ranked attributes:
 2.3906   1 animal
 1.3108  14 legs
 0.9743   5 milk
 0.8657   9 toothed
 0.8301   4 eggs
 0.7907   2 hair
 0.7179   3 feathers
 0.6762  10 backbone
 0.6145  11 breathes
 0.5005  15 tail
 0.4697   6 airborne
 0.4666  13 fins
 0.3895   7 aquatic
 0.3085  17 catsize
 0.1331  12 venomous
 0.0934   8 predator
 0.0507  16 domestic

Selected attributes: 1,14,5,9,4,2,3,10,11,15,6,13,7,17,12,8,16 : 17
```

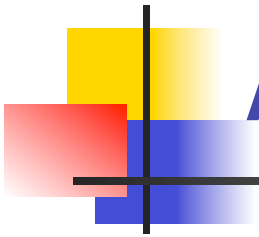


Feature Selection

- Per ogni attributo viene visualizzato un punteggio, in questo caso il gudadagno di informazione, a partire dall'attributo che ha il guadagno maggiore fino a quello che ha il guadagno minore. Normalmente tutti gli attributi vengono selezionati, indipendentemente dal valore del punteggio. Tuttavia, modificando i parametri del metodo di ricerca **`Ranker`**, è possibile cambiare il comportamento

- **`numToSelect`**: se viene impostato al valore *n* positivo, indica che si vogliono selezionare solo i primi *n* attributi in ordine di rilevanza. Se impostasto a un valore negativo, la selezione avviene in base al valore di soglia di cui si parla qui sotto.

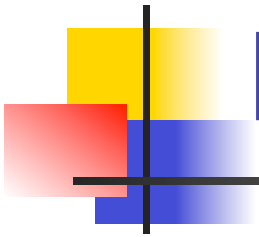
- **`threshould`**: se `numToSelect` è negativo, vengono selezionati gli attributi con ranking superiore a questa soglia. Il valore di default di -1.7976931348623157E308 causa la selezione di tutti gli attributi.



Applicare la riduzione...

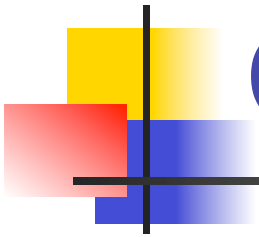
- Supponiamo di voler determinare un albero di classificazione usando l'algoritmo ID3. Dopo aver discretizzato i dati col filtro supervisionato **Discretize** (l'algoritmo ID3 si applica solo a dati discreti) si ottiene un grafo ad un solo livello in cui viene selezionato l'attributo animal. Questo perchè ID3 utilizza il guadagno di informazione per selezionare gli attributi sui vari nodi. L'accuratezza del metodo, valutata con la tecnica della "cross validation", è vicino allo 0!!

- L'algoritmo J48 (che è un clone di C4.5) si basa invece sul Gain Ratio, ed è immune a questo fenomeno, e genera un buon classificatore con accuratezza del 92% circa. Notare che se si elimina l'attributo animal manualmente, l'algoritmo ID3 produce un buon albero di classificazione, migliore di quello di C4.5 (accuratezza del 97%).



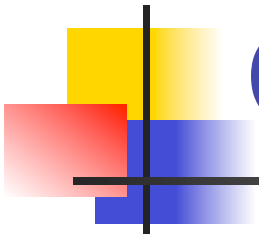
Metodi di classificazione

- Analizziamo un insieme di dati (zoo.arff) con i vari metodi di classificazione conosciuti.
- Prima di effettuare le varie analisi, discretizziamo il set di dati usando il filtro supervisionato Discretize. In questo modo possiamo utilizzare anche algoritmi che funzionano solo su dati discreti.



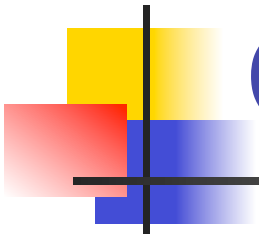
Classificare…

- **ZeroR**
  - Viene selezionata come predizione la classe più frequente nel set di dati. Equivale a un albero di classificazione di altezza 0.
    - Accuratezza = 41%
  - Notare che per ZeroR (e per tutti gli algoritmi di classificazione) è possibile ottenere uno scatter plot, ma con in più l'indicazione se l'istanza è classificata correttamente oppure no.
    - Istanza classificate correttamente sono marcate con '+', mentre le istanze classificate non correttamente sono classificate con un quadratino.



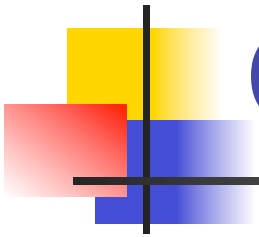
Classificare…

- J48 e ID3
  - Genera un albero di classificazione.
    - Accuratezza: 89%
- Naive Bayes
  - Utilizza il metodo bayesiano con incremento dei contatori di 1, in modo da evitare i problemi che sorgono con probabilità nulle.
    - Accuratezza: 93%



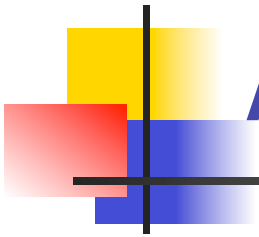
Classificare…

- BayesNet
  - È un metodo per apprendere reti bayesiane. La struttura della rete può essere fissata o può essere appresa dall'algoritmo, e si possono modificare sia l'euristica usata per la ricerca della struttura della rete, sia il metodo usato per stimare le probabilità condizionate.
    - Accuratezza: 95%



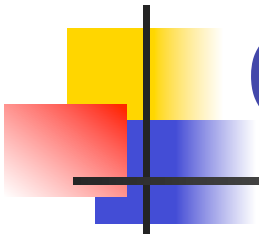
Classificare…

- IDk
  - Implementa il metodo “k-NN”. E’ possibile selzionare il valore di k e se si vuole che le istanze siano pesate a seconda della distanza.
    - Accuratezza:
      - K=1 ➔ 96%
      - K=5 ➔ 93%



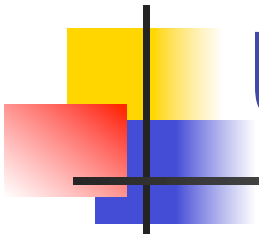
Analisi di raggruppamento

- Weka dispone di alcuni (pochi) tipi di analisi di raggruppamento.
- K-Means
  - Input:
    - numero di partizioni
    - “seme” per la generazione di numeri casuali.



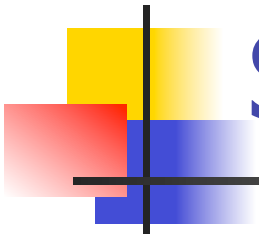
Cluster Mode

- Nel pannello di sinistra, cluster mode, è possibile selezionare come effettuare il raggruppamento. Si può scegliere tra:
  - Use Training Set.
  - Supplied Test Set.
  - Percentage Split.
  - Classes to cluster evaluation.



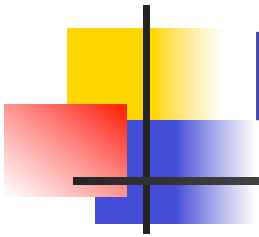
Use Training Set

- Viene applicato l'algoritmo di raggruppamento su tutti gli attributi. Come metodo per valutare la bontà del risultato, viene calcolata la somma delle distanze al quadrato dal centro del cluster.



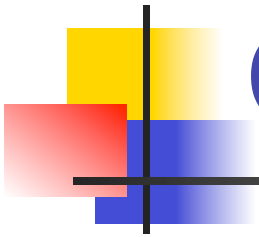
Supplied Test Set

- Simile al metodo precedente, ma la somma delle distanze al quadrato, viene calcolata su un insieme di test differente



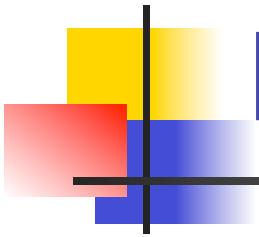
Percentage Split

- Il set di dati viene diviso in una parte di addestramento ed una di test secondo al percentuale indicata.



Classes to cluster Evaluation

- Si sfrutta l'esistenza di un attributo di classe. L'addestramento avviene su tutto l'insieme di dati, e successivamente viene mostrato, per ogni cluster, come si distribuiscono le varie classi al suo interno.



Bibliografia

- “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation”. [I. Witten, E. Frank].
- “Artificial intelligence modern approach” [Russel, Norvig]
- “Data Mining” [Paolo Giudici]