Introduzione al Data Mining

e applicazioni al dominio del Contact Management

Parte I*: Il Processo di Data Mining, Principali tipologie di Learning*

Andrea Brunello

Università degli Studi di Udine

In collaborazione con dott. Enrico Marzano, CIO Gap srl
progetto Active Contact System

Contenuti

Parte I:

- cos'è e come si svolge il processo di Data Mining;
- classificazione delle principali tipologie di Learning.

Cos'è il Data Mining

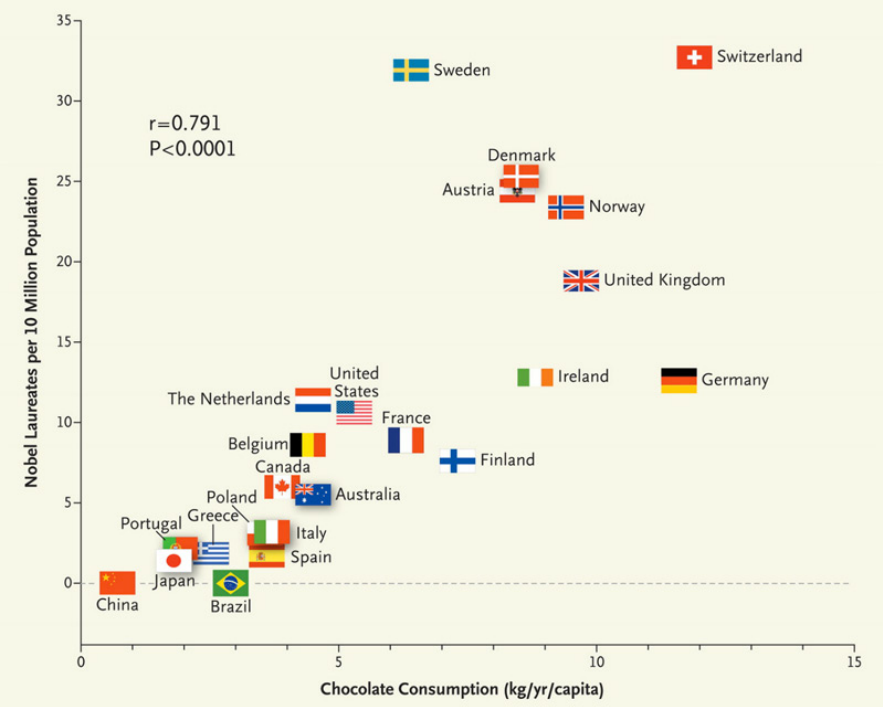
- **Dati** $\approx$ **fatti** memorizzati, registrati;
- L'**informazione** è costituita dall'insieme dei **concetti**, delle regolarità, degli schemi che si trovano "nascosti" fra i dati;
- Il **Data Mining** si occupa dell'**estrazione** e della presentazione di **informazione** utile, precedentemente sconosciuta, ed implicitamente contenuta in una (grande) mole di dati.
  - E' un processo di astrazione (generazione di un modello).
- Il **Machine Learning** costituisce la "base tecnica" del Data Mining.

Cos'è il Data Mining (2)

- Utilizzare i modelli/pattern appresi per:
  - **conoscere**: comprendere che determinate fasce di popolazione sono più propense ad acquistare un determinato bene;
  - **inferire**: probabile guasto ad un macchinario, da un insieme di sintomi;
  - **predire**: stabilire se e quale variazione nelle vendite risulterà da un aumento del budget pubblicitario.
- tali fini possono mescolarsi, si pensi alla ricerca di un modello che fornisca la valutazione di un'abitazione sulla base di diversi valori in input.

Cos'è il Data Mining (3)

- spesso i pattern scoperti risulteranno banali, frutto di correlazioni casuali, o non completamente corretti.
- ciò può essere dovuto a:
  - errori nei dati;
  - caratteristiche del dominio (es. esistenza di dipendenze funzionali)



Cos'è il Data Mining (4)

Riassumendo:

- Il Data Mining sfrutta tecniche di Machine Learning
- per estrarre semi-automaticamente
- da (grandi) quantità di dati
- informazioni, pattern utili

Input del processo:

- istanze, esempi dei concetti che si vogliono apprendere

Output del processo:

- predizioni/classificazioni
- modelli

Il processo di Data Mining

In genere, il processo di Data Mining si articola come segue:

- il tutto inizia con il porsi una **domanda**, ben chiara e specifica;
- in seguito, si passa alla raccolta dei **dati** da utilizzare come input;
- viene selezionato un insieme di **caratteristiche** (features) ritenute importanti su tali dati, e per il fine che si vuole ottenere;
- si applica un **algoritmo** di machine learning sul dataset così definito, in modo da "addestrare" un modello;
- dopo un'eventuale fase di tuning, il modello prodotto dall'algoritmo viene **valutato**, ed è infine pronto per essere utilizzato.

Un primo esempio: SPAM vs HAM

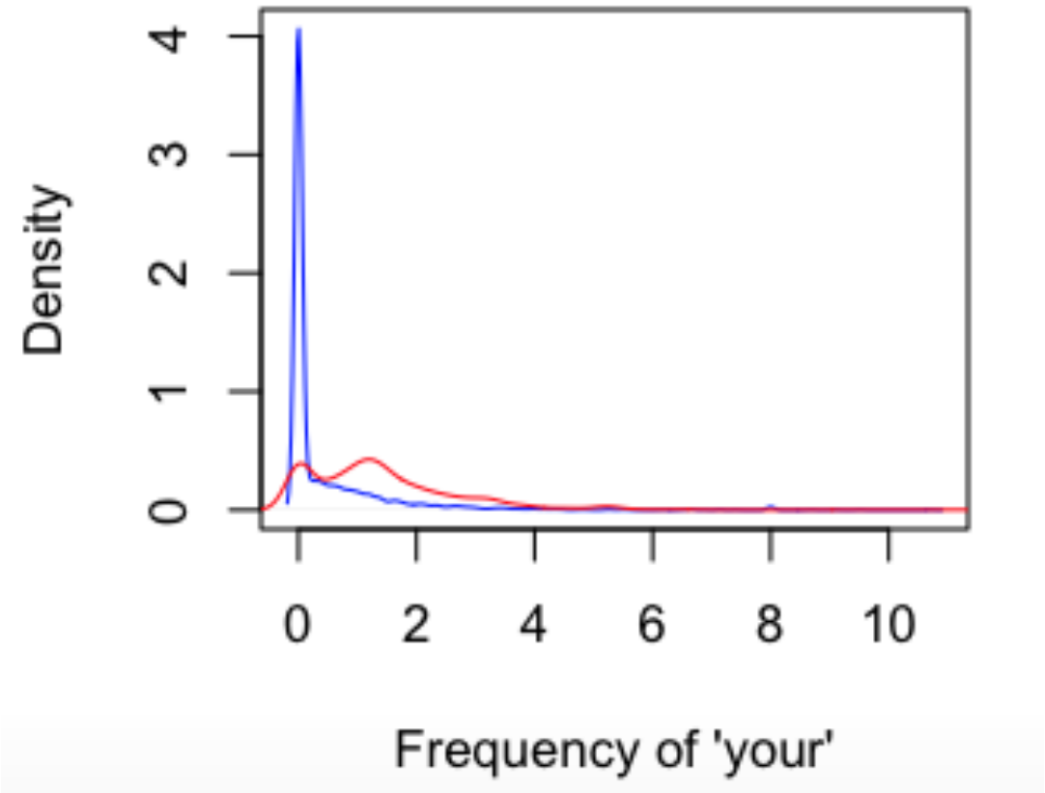
Ripercorriamo ora il processo di Data Mining con un esempio focalizzato sulla predizione.

- **Domanda**: è possibile distinguere automaticamente fra i messaggi email che sono indesiderati (SPAM) e quelli legittimi (HAM)?
- **Dati**: insieme di 4601 istanze di email già classificate, ed ognuna avente 57 caratteristiche indicanti la frequenza di determinate parole e caratteri nel corpo del messaggio;
  - *www.inside-r.org/packages/cran/kernlab/docs/spam*

Un primo esempio: SPAM vs HAM (2)

Selezione delle **caratteristiche**:

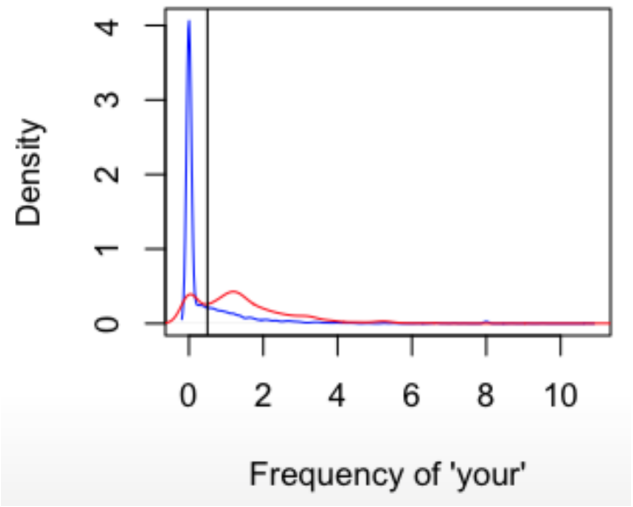
- processi di *attribute selection*;
- esplorazione e selezione manuale.



Un primo esempio: SPAM vs HAM (3)

Modello: utilizzo di un valore di **cutoff**

- *se frequenza della parola "your" nel testo > 0.5, allora l'email è SPAM*



- Otteniamo la **tabella di contingenza** (valori su insieme di training):

Predizione \ Classe	**nonspam**	**spam**
nonspam	0.4590	0.10017
spam	0.1469	0.2923

L'input del processo

Entrando nel dettaglio del processo di Data Mining, abbiamo che il suo input è costituito da:

- **concetti**
- **istanze**
- **attributi**:
  - *numerici* VS *nominali*
  - *feature* VS *target* (supervised learning)

Rivestono grande importanza l'integrazione, pulizia, e **trasformazione** dei dati.

$\rightsquigarrow$ raramente le istanze in input copriranno tutti i casi possibili per il dominio!

L'input del processo (2)

Condizioni	Temp.	Umidità	Vento	Si_gioca?
soleggiato	calda	alta	falso	no
soleggiato	calda	alta	vero	no
nuvoloso	calda	alta	falso	si
pioggia	mite	alta	falso	si
pioggia	fredda	normale	falso	si
pioggia	fredda	normale	vero	no
nuvoloso	fredda	normale	vero	si
soleggiato	mite	alta	falso	no
soleggiato	fredda	normale	falso	si
pioggia	mite	normale	falso	si
soleggiato	mite	normale	vero	si
nuvoloso	mite	alta	vero	si
nuvoloso	calda	normale	falso	si
pioggia	mite	alta	vero	no

Supervised vs Unsupervised Learning

Distinzione fondamentale, a seconda dell'obiettivo del processo:

- **Supervised Learning**: all'algoritmo di learning viene fornito un risultato noto per ciascuna istanza di training e si punta a determinare il valore per nuove istanze (attributo obiettivo).
  - *Classificazione con alberi, regressione lineare, . . .*
- **Unsupervised Learning**: non si cerca di prevedere il valore di uno specifico attributo, ma viene ricercata ogni possibile associazione/correlazione fra gli attributi.
  - *Clustering, regole di associazione, . . .*

Problemi di regressione e di classificazione

Nel *Supervised Learning*, a seconda della tipologia dell'attributo obiettivo, dinstinguiamo problemi di

- **classificazione**: (o predizione). Si vuole assegnare a ciascuna istanza uno di un insieme finito di valori (*cl. discreti*; in altri casi viene restituita la probabilità di appartenere a ciascuna classe, o un ranking delle istanze: *cl. continui*);
- **regressione**: si vuole assegnare a ciascuna isanza un valore numerico.

Alcune famiglie di algoritmi di learning si adattano ad entrambi i problemi (alberi di classificazione e di regressione, regressione lineare e logistica, . . .).

Riferimenti

I. H. Witten, E. Frank, M. A. Hall: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition, 2011.

G. James, D. Witten, T. Hastie, R. Tibshirani: *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.