Tecniche statistiche per il data mining e i big data

antonio.lucadamo@unisannio.it

Informazioni generali

Lezioni:
Mercoledí ore 14-16
Giovedí ore 9-11
Venerdí ore 9-11

Ricevimento: Su prenotazione!!

Testi di riferimento:

- Analisi dei dati e data mining (Azzalini - Scarpa)
- An introduction to Statistical Learning with Applications in R (James - Witten - Hastie - Tibshirani)

Programma

- Introduzione al data mining e ai big data
- Richiami di modelli lineari
- Richiami di regressione logistica (binaria e multinomiale)
- Tecniche per la selezione dei modelli
- Previsione di una variabile quantitativa: Stima non parametrica, regressione locale, spline, modelli additivi, alberi di regressione e reti neurali.
- Metodi di classificazione: analisi discriminante, alberi di classificazione, reti neurali, combinazioni di classificatori.
- Metodi di analisi interna: metodi di raggruppamento, associazioni tra variabili.

Applicazioni attraverso l'uso del software R

Introduzione

Situazioni tipiche dell'utilizzo dei big data:

- Ogni mese una catena di supermercati emette milioni di scontrini, uno per ogni carrello della spesa che si presenta ad una cassa. Il contenuto di ciascuno di questi carrelli riflette le esigenze di spesa, le propensioni e il comportamento economico del consumatore. L'insieme di tutte queste liste lista della spesa costituisce un'importante base informativa per orientare le politiche sia di vendita che di acquisto di prodotti da parte del supermercato. Questa operazione diventa ancora piú interessante se si mettono in corrispondenza le singole liste della spesa alle "carte fedeltá" dei clienti, dato che possiamo cosí seguire il loro comportamento attraverso una sequenza di acquisti.

- Situazione analoga si riscontra con l'utilizzo delle carte di credito, con la particolaritá che qui i clienti sono tutti identificati perfettamente. Un altro elemento di diversitá é che l'azienda che emette la carta non vende direttamente nulla al cliente, ma puó offrire ad altre ditte l'opportunitá di fare offerte speciali a clienti mirati.
- Le societá telefoniche generano ogni giorno dati relativi a milioni di chiamate e di altri servizi forniti. Le aziende sono interessate all'analisi del comportamento del consumatore sia per cercare di individuare le opportunitá di ampliamento dei servizi di cui il cliente usufruisce, sia per individuare la possibilitá che lo stesso disdica l'utenza passando ad un altro operatore.

- Internet costituisce un enorme deposito di informazioni, contenute in tanti documenti, una piccolissima frazione dei quali rilevante rispetto ad una specifica interrogazione sottoposta ad un motore di ricerca. L'operazione di selezione dei documenti rilevanti che deve essere compiuta dal motore di ricerca é complicata da vari elementi: la numerositá dell'insieme complessivo dei documenti; l'insieme non costituisce una forma strutturata come in un data-base organizzato; all'interno del singolo documento gli ingredienti che determinano la sua pertinenza o meno rispetto all'interrogazione non sono collocati in una posizione predeterminata.

- Anche nella ricerca scientifica sono moltissimi gli ambiti in cui i moderni metodi di rilevazione danno luogo a raccolte di dati imponenti. Uno dei campi di ricerca piú attivi riguarda la microbiologia con particolare riferimento alla struttura del DNA. L'analisi della sequenza di porzioni di DNA porta alla costruzione di gigantesche tabelle dette DNA microarray in cui ciascuna colonna costituisce una sequenza di alcune migliaia di valori numerici associata al DNA di un individuo. L'obiettivo é quello di mettere in relazione la configurazione di queste sequenze con la presenza di determinate malattie.
- L'insieme delle rilevazioni di natura fisica, chimica e di altro genere volte ad esaminare l'evoluzione del clima terrestre ha ormai raggiunto dimensioni gigantesche. La semplice organizzazione in modo strutturato di tali dati pone problemi significativi e anche in questo ambito il data mining o le tecniche per i big data sono strumenti indispensabili.

Alcuni concetti importanti:

- Numerositá
- dimensionalitá

Il data mining rappresenta l'attivitá di elaborazione in forma grafica o numerica di grandi raccolte o di flussi continui di dati con lo scopo di estrarre informazione utile a chi detiene i dati stessi. Il concetto di utilitá é molto vario e dipende dal contesto in cui si opera e dagli obiettivi che ci si prefigge.

I modelli

All models are wrong but some are useful (George Box)

Un modello é una rappresentazione semplificata del fenomeno di interesse, funzionale ad un obiettivo specifico.

- É essenziale che si tratti di una rappresentazione semplificata, perché se fosse molto simile non servirebbe, conservando la complessitá.
- Il modello deve essere funzionale ad un obiettivo specifico, ma questo vuol dire che ci potrebbero essere modelli diversi per lo stesso fenomeno, a seconda degli obiettivi.
- Anche una volta definito l'aspetto del fenomeno che si vuole descrivere, resta comunque margine di scelta nella modalitá con cui si esprimono le relazioni tra le componenti in gioco.
- Ci sono quindi diverse "dimensioni": il grado di semplificazione della realtá, la scelta degli elementi da riprodurre, la natura delle relazioni delle componenti in gioco.

Quindi il modello sará anche "sbagliato", ma l'importante é che lo sia in maniera utile.