Introduzione al Machine Learnig

Prof. Fabio Divino
Tutorial 2: Binary classification

Considerare il dataset **caravan_data.RData** contenente 5822 record di clienti potenziali per l'acquisto di una polizza assicurativa per Caravan. Ogni record è costituito da 85 variabili esplicative contenenti dati socio-demografici (variabili 1-43) e dati sulla proprietà del prodotto (variabili 44-85). I dati socio-demografici derivano da codici postali. Tutti i clienti che vivono in aree con lo stesso codice postale hanno gli stessi attributi socio-demografici. La variabile 86 (Purchase) indica se il cliente ha acquistato una polizza assicurativa per Caravan ("Yes") oppure no ("No"). Ulteriori informazioni sulle singole variabili sono disponibili all'indirizzo http://www.liacs.nl/~putten/library/cc2000/data.html

Nel dataset, sono già definiti i due dataframe per i passi di **training** e **test** di una sessione di machine learning: caravana.training e caravan.test

Quesiti

A) Utilizzando le variabili che si ritiene rilevanti fra le 85 esplicative, confrontare i due seguenti classificatori binari:
   Regressione logistica (glm);
   K nearest neighbours (knn).
   Il numero di variabili che si possono considerare è a piacere, ma utilizzare sempre il criterio di parsimonia (max=10/15). I due classificatori devono utilizzare le stesse variabili.
B) Fare un'analisi semantica dei risultati.

Suggerimento

Per selezionare le variabili rilevanti si può utilizzare la significatività statistica applicata in prima fase al classificatore logistico, ad esempio attraverso una procedura **backward**.

Risultati

Dopo aver svolto il tutorial, salvare la sessione di lavoro come file

nome_cognome_tutorial2.RData (rigorosamente tutto in minuscolo!)

Inviare il file a

fabio.divino@unimol.it

indicando nel testo del messaggio le variabili che sono state considerate e quale metodo è risultato migliore, oltre alla relativa analisi semantica dei risultati.