Statistica a Bocconi

September 9, 2015

L'analisi dei dati a disposizione è stata effettuata principalmente grazie all'applicazione del Least Absolute Selection and Shrinkage Operator, comunemente noto come LASSO. I risultati finali sono tuttavia stati raggiunti dopo numerosi tentativi per quanto riguarda la gestione dei dati iniziali.

Proponiamo un elenco di operazioni svolte, relative all'ultima submission:

1. Abbiamo inserito alcune variabli ottenute dall'elaborazione di quelle esistenti, ad esempio anno-settimana, giorno della settimana, data di test vicina o antecedente a quella di imbottigliatura, numero di bottiglie ignoto, basso o alto, formato ignoto, standard (75 cl) o alternativo...

2. I valori mancanti delle variabili numeriche sono stati sostituiti con la media calcolata considerando anche il campione esterno; per quanto riguarda le variabili categoriche, i valori mancanti sono stati trasformati in un livello;

3. La variabile "articolo" è stata sostituita con 47 variabili dummy, ognuna corrispondente a una delle parole possibili ad esclusione di "vino", sempre presente all'interno della variabile articolo.

Per l'esecuzione del LASSO, si sono considerate tutte le variabili, con le interazioni di secondo grado tra le dummy relative alla vecchia variabile "articolo", per un totale di 1282 variabili.

Il LASSO utilizza in questo caso una funzione di link logit, il modello ottimale è ricercato attraverso 10-fold cross-validation, utilizzando la funzione obiettivo fornitaci, per passare dalle probabilità stimate ad una variabile binaria finale.

Il modello scelto è quello che è in grado di minimizzare la media out-of-sample dei punteggi calcolata tra i 10 raggruppamenti dei dati forniti. In senso conservatore, abbiamo scelto un coefficiente di penalizzazione più basso per cercare di trattenere più variabili all'interno del modello.

In alternativa avevamo provato anche con interazioni tra altre variabili, trasformazioni delle variabili (es. logaritmi), selezione di variabili bayesiana (usando una prior "spike and slab"), ma non siamo riusciti ad ottenere dei risultati significativamente diversi, pertanto abbiamo deciso di mantenere il modello il più semplice possibile.

Considerato che il LASSO induce sparsità nelle variabili esplicative, i risultati parziali che abbiamo ricevuto come feedback sono interpretabili in diversi modi. Da un lato, sono lusinghieri, in quanto non compaiono errori di falsi idonei; tuttavia il modo in cui ci arriviamo non suggerisce particolari meccanismi fisici per l'identificazione anticipata dei vini da rivedere. Infatti sopravvive alla selezione brutale del LASSO una sola variabile, ovvero la dummy "ATTO": essa assume valore 1 se la variabile "articolo" include la parola "ATTO", 0 altrimenti. Pur non essendo questo molto interessante, questa strategia non si è rivelata peggiore di altre da noi provate.