I dati Orange

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- Guyon, Lemaire, Boullé, Dror, Vogel (2009) Analysis of the KDD Cup 2009: Fast Scoring on a Large Orange Customer Database
- Chapter 6 of Zumel and Mount (2014) Practical Data Science with R , Manning Publications
- The support site of Zumel and Mount (code and data) on GitHub

Table of Contents

KDD cup

- La Conference on Knowledge Discovery and Data Mining (KDD) è la principale conferenza sui metodi di machine learning
- Ogni anno KDD ospita una competizione di data mining, in cui i team analizzano un dataset
- La KDD Cup è stata l'ispirazione per il famoso Premio Netflix e per le competizioni Kaggle
- La KDD Cup 2009 ha riguardato i dati Orange, un dataset sui clienti della società di telecomunicazioni francese Orange

Orange data

L'obiettivo è prevedere la propensione dei clienti a cancellare il proprio account, un evento chiamato *churn*. Altri obiettivi erano prevedere la tendenza dei clienti all'uso nuovi prodotti e servizi (un evento chiamato *appetency*) e disponibilità a rispondere favorevolmente alle proposte di marketing (un evento chiamato *upselling*)

Il dataset riguarda $p = 230$ predittori su $n = 50000$ conti di carte di credito. Per motivi di privacy, i predittori sono resi anonimi: non si conosce il significato di nessuno dei predittori. Questo dataset è di interesse per

- dati rumorosi eterogenei (predittori numerici e categorici con valori mancanti)
- *Class imbalance*: frequenza relativa di *churn* 7.3% (3672/50000)

- Training set con $n = 22253$ osservazioni
- Test set con $m = 27747$ osservazioni
- Variabile risposta : `churn` = -1 (no churn), +1 (churn)
- Class imbalance: 7% positive class nel train set (1633/22253)
- $p = 230$ predittori: `Var1`, `Var2`, .. , `Var230`. Non conosciamo il significato di nessun predittore
- Il metrica per la valutazione delle previsioni è l' Area Under the Curve (AUC). Il team vincitore ha ottenuto AUC = 0.76

Table of Contents

Gestire i dati mancanti

I dati mancanti non sono rari nei set di dati reali. La prima e più importante domanda quando si incontrano dati mancanti è: perché mancano questi valori?

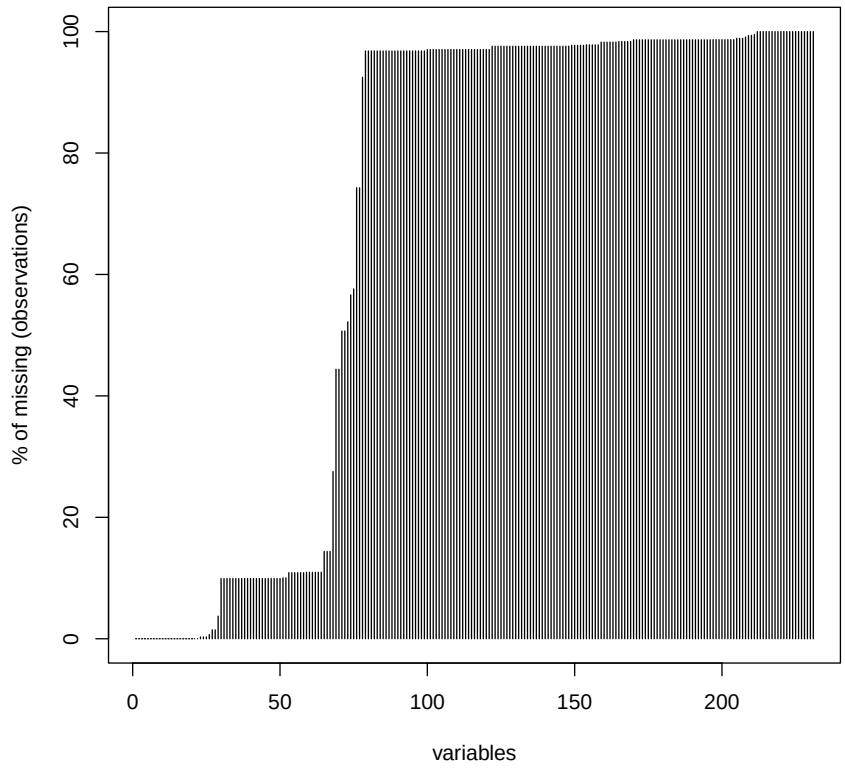
I valori mancanti sono generalmente causati da tre meccanismi:

- Una carenza strutturale nei dati
- Un evento casuale, o
- Una causa specifica

Vedere il capitolo 8 di FES

Esercizio 1

Si fornisca una sintesi sulla presenza di dati mancanti, commentando il risultato.


% of missing (observations)
0
20
40
60
80
100
0
50
100
150
200
variables

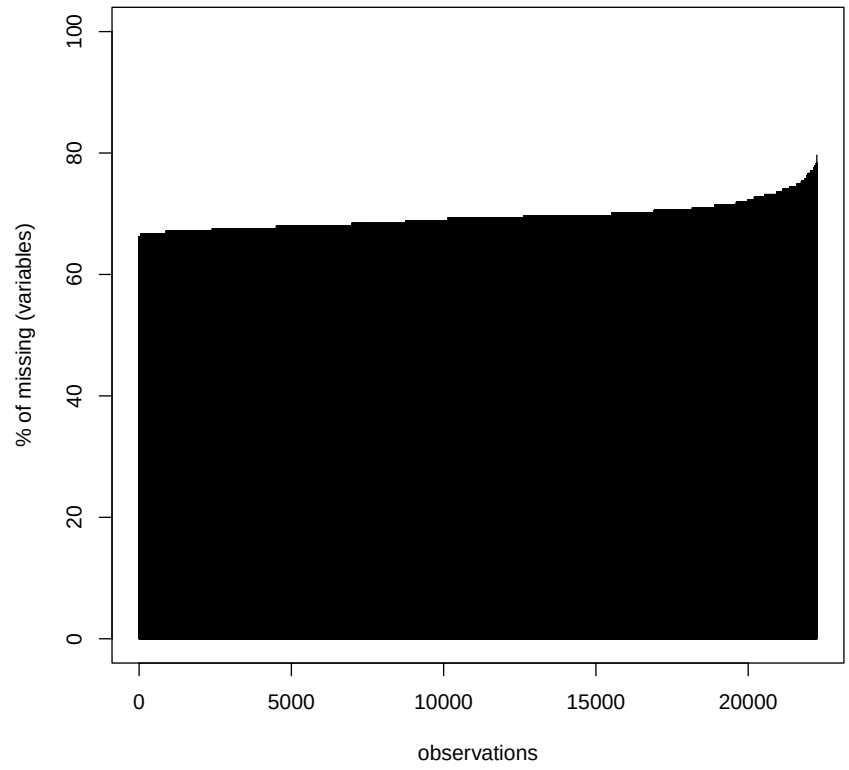

% of missing (variables)
0
20
40
60
80
100
0
5000
10000
15000
20000
observations

Table of Contents

Predittori a varianza zero e quasi zero

Per molti modelli (esclusi ad esempio i modelli basati su alberi), la presenza di predittori a varianza zero / quasi zero potrebbe causare l'impossibilità di stimare il modello o l'instabilità della stima Per identificare questi tipi di predittori, è possibile calcolare le seguenti due metriche:

- il rapporto di frequenza: la frequenza del valore più prevalente rispetto al secondo valore più frequente; ci si aspetta un valore vicino a uno per predittori bilanciati e molto grande per predittori altamente sbilanciati
- la percentuale di valori univoci: è il numero di valori univoci diviso per il numero totale di osservazioni (moltiplicato per 100); questo numero si avvicina allo zero all'aumentare della granularità dei dati

Esercizio 2

Se il rapporto di frequenza è maggiore di una soglia pre-specificata freqCut = 95/5 e la percentuale dei valori univoci è inferiore a una soglia uniqueCut = 10, potremmo considerare un predittore a varianza quasi zero.

Identificare i predittori a varianza 0 / quasi zero.

Escludere questi predittori nelle analisi successive.

Creare una variabile che identifica gli indici dei predittori `factor` e una variabile che identifica gli indici dei predittori `numeric`.

Contare il numero di modalità osservate / potenziali dei predittori `factor`.

```
V2
    0     5  <NA>
  536     1 21716

frequency ratio = 536
percent unique values = 8.987552e-05

V16
frequency ratio = 1.0625
percent unique values = 0.01599784

V210
 3av_  7A3j  DM_V  g5HH  oT7d  uKAI
   37   207    65   710    76 21158
frequency ratio = 29.8
percent unique values = 0.0002696266
```

Table of Contents

Supervised Encoding Methods

- Esistono diversi metodi per codificare i predittori categorici in colonne numeriche utilizzando la variabile risposta come guida (in modo che siano metodi supervisionati)
- Queste tecniche sono adatte ai casi in cui il predittore ha molte modalità o quando compaiono nuovi modalità non osservate nei dati di training
- Un metodo semplice è chiamato codifica dell'effetto o della verosimiglianza (*effect / likelihood encoding*): viene misurato l'effetto del livello del fattore sulla variabile risposta e questo effetto viene utilizzato come codifica numerica

- Per i problemi di regressione, potremmo calcolare il valore di risposta medio o mediano per ciascun livello del predittore categoriale dai dati di training e utilizzare questo valore per rappresentare numericamente il livello del fattore nel modello
- Per problemi di classificazione binaria, potremmo calcolare i rapporto di probabilità dell'evento (*odds / log-odds*) e utilizzarle come codifica
- Tuttavia, un problema con la codifica degli effetti è che aumenta la possibilità di overfitting

Esercizio 3

Creare un insieme di calibrazione di 5604 osservazioni, riducendo quindi il training set da 22253 osservazioni a 16698 osservazioni.

Calcolare la codifica delll'effetto per la variabile `Var218`. Quando `Var218` assume la modalità `cJvF`, il 6% dei clienti abbandona, `UYBR`, l'8% dei clienti abbandona, `NA`, il 28% dei clienti abbandona.

```
      <NA>       UYBR       cJvF
0.28033473 0.08175548 0.06011316
```

Costruire una funzione che calcola la codifica dell'effetto per variabili categoriali sui dati di training / calibrazione / test, convertendo `NA` in una modalità e trattando nuove modalità (se presenti) come non informative.

Calcolare il valore AUC per ciascun predittore codificato nel training set e nel calibration set.

Table of Contents

Binning di predittori numerici

- Il *binning*, noto anche come categorizzazione o discretizzazione, è il processo di trasformazione di una variabile quantitativa in un insieme di due o più categorie. Ad esempio, una variabile potrebbe essere trasformata in categorie corrispondenti ai quantili
- Il *binning* permette di evitare il problema di dover specificare la relazione tra il predittore e la variabile risposta

Esercizio 4

Convertire i predittori numerici in predittori categoriali con 10 categorie determinate dai decili.

Calcolare la codifica dell'effetto per queste nuove variabili categoriali sui dati di training / calibrazione / test.

Calcolare il valore AUC per ciascun predittore codificato nel training set e nel calibration set.

Table of Contents

Selezione dei predittori

- Per evitare il sovra-adattamento, selezioneremo le codifiche degli effetti che funzionano bene nei dati di calibrazione, misurandoli tramite la log-verosimiglianza
- Per un'osservazione con `churn` = 1 e una probabilità stimata di 0.9 di essere `churn`, la logverosimiglianza è $log(0.9)$; per un'osservazione con `churn`=-1, lo stesso punteggio di 0.9, logverosimiglianza è $log(1 - 0.9)$

$$\log \ell = \sum_{i=1}^{m} (I\{y_i^* = 1\} \log(x_i^*) + I\{y_i^* = -1\} \log(1 - x_i^*))$$

dove y_i^* è la $i$ma risposta nei dati di calibrazione e x_i^* è l'$i$mo punteggio del predittore

Esercizio 5

- Il modello nullo ha probabilità di of `churn` = (numero dei clienti `churn`)/(totale dei clienti) e logverosimiglianza $\log \ell_0$
- Selezionare i predittori con un miglioramento della devianza

$$2(\log \ell - \log \ell_0)$$

  superiore ad una certa soglia (e.g. 5)
- Stimare un modello logistico con i predittori selezionati, e calcolare il valore AUC