I dati Netflix

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

Si ringraziano Giles J. Hooker e Saharon Rosset per aver condiviso i dati Netflix.

- AS § 2.1.3 per la regressione lineare multidimensionale
- An Introduction to Statistical Learning, 2nd edition § 12.3 e § 12.5.2 per valori mancanti e completamento della matrice

Il Netflix Prize

Il Netflix Prize è stata una competizione il cui scopo era quello di prevedere le valutazioni (*rating*) di diversi film fornite dagli utenti. Netflix ha fornito le valutazioni di 17.770 titoli di film da parte di 480.189 utenti, insieme alla data di ciascuna valutazione. Il compito era quello di prevedere le valutazioni per 282.000 combinazioni di utente-film-data che non erano presenti nel training set.

Netflix ha misurato la bontà delle previsioni con la radice quadrata dell'errore quadratico medio (*Root Mean Square Error*) e ha offerto un premio di $ 1.000.000 al primo classificato (che ha migliorato la bontà delle previsioni del loro approccio di oltre il 10%). Il premio è stato vinto nel 2009. I dettagli del Premio Netflix sono disponibili presso www.netflixprize.com

I dati

Poiché la competizione Netflix prevedeva un dataset molto grande e un problema non-standard, semplificheremo notevolmente il problema.

Il training set fornisce le valutazioni di $n = 10000$ utenti per 99 film, insieme alle date in cui sono state effettuate le valutazioni. Si noti che 14 di questi film sono stati valutati da tutti gli n utenti, mentre i restanti 85 film contengono valori mancanti.

L'obiettivo

L'obiettivo è prevedere la valutazione da parte dei $m = 2931$ utenti del test set per il film *Miss Detective* (titolo originale: *Miss Congeniality*); vi viene anche fornita la data in cui è stata effettuata ciascuna valutazione. Come per il training set, tutti gli m utenti del test set hanno valutato 14 film, mentre per i restanti 85 ci sono valori mancanti. Il test set fornisce le stesse informazioni del training set: le date e le valutazioni di questi 99 film insieme alla data delle valutazioni per *Miss Congeniality*.

Come per la competizione Netflix, la bontà delle previsioni verrà valutata con la radice quadrata dell'errore quadratico medio (RMSE) sul test set.

$$\text{RMSE}_{\text{Te}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i^* - \hat{y}_i^*)^2}$$

I dati sono disponibili nell'archivio del corso in formato file di test delimitato da tabulazioni, e comprendono:

- `Train_ratings_all.dat` : il training set contiene le valutazioni che gli utenti hanno assegnato a ciascuno dei 99 film
- `Test_ratings_all.dat` : come sopra per il test set
- `Train_dates_all.dat` : per il training set, le date in cui sono state effettuate le valutazioni di cui sopra.
- `Test_dates_all.dat` : come sopra per il test set
- `Train_y_rating.dat` : le valutazioni che gli utenti del training set hanno assegnato a Miss Congeniality
- `Train_y_date.dat` : per il training set, le date in cui gli utenti del hanno valutato Miss Congeniality
- `Test_y_date.dat` : Stesse informazioni per il set di test
- `Movie_titles.txt` : i titoli e le date di uscita per i 99 film, indicati nello stesso ordine delle colonne dei dati descritti sopra

Alcune osservazioni

Le valutazioni sono numeri interi da 1 a 5. Un valore di 0 indica un valore mancante

Per comodità, le date sono fornite come numero di giorni trascorsi a partire dal primo Gennaio 2017 fino ad una certa data. L'etichetta che identifica i valori mancanti per le date è '0000'

Alcune osservazioni

Le valutazioni sono numeri interi da 1 a 5. Un valore di 0 indica un valore mancante

Per comodità, le date sono fornite come numero di giorni trascorsi a partire dal primo Gennaio 2017 fino ad una certa data. L'etichetta che identifica i valori mancanti per le date è '0000'

Esercizio 1

Caricare solo i dati di training (senza considerare le date).

Visualizzare la percentuale di valori mancanti per film.

Come è stato valutato il film Miss Congeniality? Fornire una sintesi grafica e numerica.

Valutazione media rispetto ai film senza dati mancanti?

Quali sono i film (senza dati mancanti) maggiormente correlati con Miss Congeniality?

Esercizio 2

Divisione in training e test

```
m <- 2000
n <- nrow(X) - m
set.seed(123)
test.id <- sample(n+m,m)
test <- data.frame(y=y[test.id,], X[test.id,])
train <- data.frame(y=y[-test.id,], X[-test.id,])
```

Calcolare il RMSE sul test set per il modello nullo.

Calcolare il RMSE sul test set per il modello lineare con tutte le variabili con i valori mancanti codificati come 0

Calcolare il RMSE sul test set per il modello lineare con solo i film senza dati mancanti

Regressione lineare multidimensionale

Se ci sono q variabili risposta $\underset{n\times q}{Y}$, e consideriamo q modelli di regressione lineare usando la stessa matrice del disegno $\underset{n\times p}{X}$, arriviamo alla formulazione

$$\underset{n\times q}{Y} = \underset{n\times p}{X}\,\underset{p\times q}{B} + \underset{n\times q}{E}$$

dove B è la matrice formata da q colonne di dimensione p, ciascuna delle quali rappresenta i parametri di regressione per la corrispondente colonna di Y, e la matrice E è costituita di termini di errore tali che

$$\mathbb{V}\mathrm{ar}(e_i) = \Sigma$$

dove e_i^T rappresenta la i-sima riga di E, per $i = 1, \ldots, n$, e Σ è la matrice di varianza/covarianza che esprime la correlazione delle variabili risposta.

Allora la soluzione del problema dei minimi quadrati multidimensionali è

$$\hat{B} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$$

che sono i q vettori stimati per ciascuna variabile risposta, mentre la stima di Σ è $\hat{\Sigma} = \frac{1}{n-p} Y^\mathsf{T} (I_n - H) Y$

Esercizio 3

Per i dati Neflix, la variabile risposta $Y \in \mathcal{Y} = \{1, 2, 3, 4, 5\}$. Sia

$$Y_j = \begin{cases} 1 & \text{se } Y = j \\ 0 & \text{altrimenti} \end{cases}$$

per $j \in \mathcal{Y}$. Otteniamo quindi una variabile risposta multidimensionale $\underset{n \times 5}{Y}$. (utilizzare i valori mancanti codificati con o).

Si stimi

$$\mathbb{E}(Y_j | X = x) = \mathbb{P}(Y_j = 1 | X = x), \qquad j = 1, \ldots, 5$$

utilizzando la regressione lineare multidimensionale. Si preveda y^* dato x^* con il valore j che massimizza $\mathbb{P}(Y_j = 1 | X = x^*)$. Calcolare il RMSE sul test set.

Stimare il valore atteso condizionato

$$\mathbb{E}(Y | X = x) = \sum_{j \in \mathcal{Y}} j \cdot \mathbb{P}(Y = j | X = x)$$

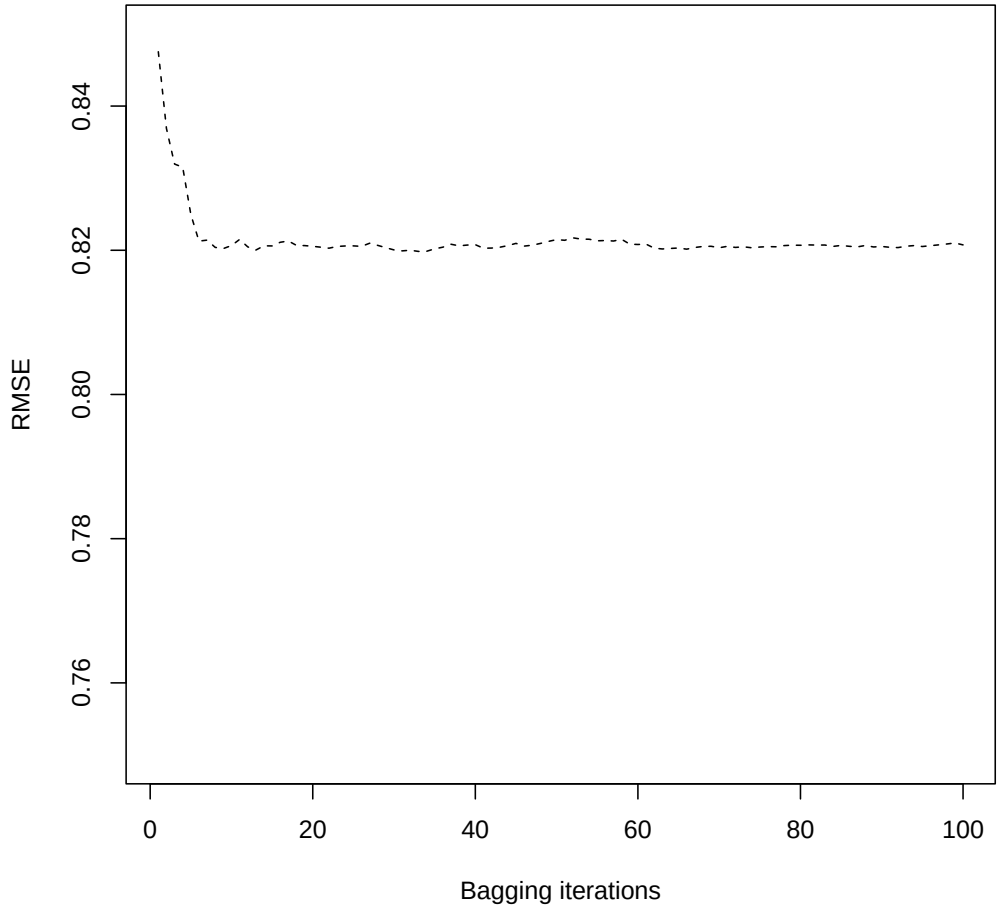
e calcolare il RMSE sul test set.

Esercizio 4

Calcolare il RMSE sul test set per il modello multinomiale (libreria `nnet`), utilizzando come previsione lo score più probabile e quello atteso.

Calcolare il RMSE sul test set per con l'analisi discriminante lineare e quadratica (libreria `MASS`), utilizzando come previsione lo score più probabile e quello atteso.

Codificare i valori mancanti come `NA`. Calcolare il RMSE sul test set con un albero di regressione (libreria `rpart`) e con un bagging di alberi di regressione (senza utilizzare alcuna libreria aggiuntiva).


RMSE
0.76
0.78
0.80
0.82
0.84
0
20
40
60
80
100
Bagging iterations

Esercizio 5

Implementare l'algoritmo L2-boosting per alberi di regressione

1. Initialize $\hat{f}(x) = \bar{y}$ and $r_i = y_i - \bar{y}$ for $i = 1, \ldots, n$
2. For $b = 1, 2, \ldots, B$, repeat:
   - Fit a tree $\hat{f}^b$ with d splits to the data $(x_1, r_1), \ldots, (x_n, r_n)$
   - Update $\hat{f}$ by adding in a shrunken version of the new tree:

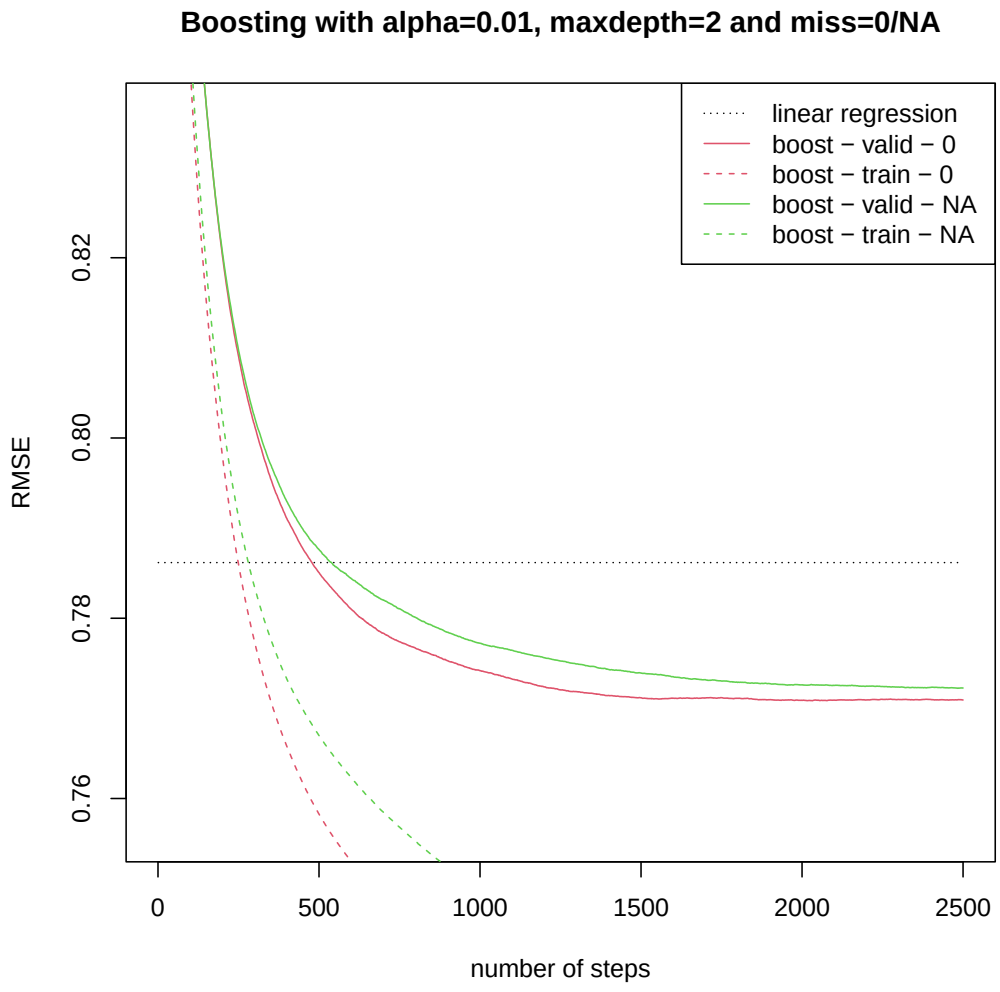
$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

   - Update the residuals:

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

3. Output the boosted model:

$$\hat{f}(x) = \bar{y} + \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$


Boosting with alpha=0.01, maxdepth=2 and miss=0/NA
linear regression
boost − valid − 0
boost − train − 0
boost − valid − NA
boost − train − NA
RMSE
0.82
0.80
0.78
0.76
0
500
1000
1500
2000
2500
number of steps

Esercizio 6

Svolgere l'analisi delle componenti principali con solo i film senza dati mancanti. Interpretare il risultato.

Calcolare il RMSE sul test set per con la *principal component regression* (PCR) per un numero di componenti da 1 a 14.

Table of Contents

Teorema di Eckart-Young

Data una matrice Z di dimensione $n \times p$, la soluzione di

$$\min_{A \in \mathbb{R}^{n \times q}, B \in \mathbb{R}^{p \times q}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{p} (z_{ij} - \sum_{k=1}^{q} a_{ik} b_{jk})^2 \right\}$$

è data da $A = Y_q$ (la matrice dei punteggi delle prime q componenti principali di Z) e $B = V_q$ (la matrice dei primi q autovettori della matrice di varianze/covarianze di Z), quindi

$$z_{ij} \approx \sum_{k=1}^{q} y_{ik} v_{jk} = \{ Y_q V_q^t \}_{ij}$$

Se sono presenti dati mancanti, possiamo risolvere

$$\min_{A \in \mathbb{R}^{n \times q}, B \in \mathbb{R}^{p \times q}} \left\{ \sum_{(i,j) \in \mathcal{O}} (z_{ij} - \sum_{k=1}^{q} a_{ik} b_{jk})^2 \right\}$$

dove $\mathcal{O}$ è l'insieme delle coppie di indici (i, j) osservati.

Una volta trovate le soluzioni A^* e B^*, possiamo

- sostituire le osservazioni mancanti z_{ij} con $z_{ij}^* = \sum_{k=1}^{q} a_{ik}^* b_{jk}^*$
- calcolare le q componenti principali sui dati completi

La soluzione del problema di minimo è più complicata rispetto al caso di dati completi, ma è possibile utilizzare il seguente algoritmo iterativo (denominato *hard impute*)

Algoritmo iterativo

1. Per una matrice incompleta di dati X, costruire la matrice $\hat{X}$ con elementi
$$\hat{x}_{ij} = \begin{cases} x_{ij} & (i,j) \in \mathcal{O} \\ \bar{x}_j & (i,j) \notin \mathcal{O} \end{cases}$$
dove $\bar{x}_j$ è il valore medio della j-sima variabile.
2. Ripetere i passi a.-c. fino a convergenza:
   a. risolvere
$$\min_{A \in \mathbb{R}^{n \times q}, B \in \mathbb{R}^{p \times q}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{x}_{ij} - \sum_{k=1}^{q} a_{ik} b_{jk})^2 \right\}$$
   b. Per ogni elemento $(i,j) \notin \mathcal{O}$, $\hat{x}_{ij} \leftarrow \sum_{k=1}^{q} a^*_{ik} b^*_{jk}$
   c. Calcolare
$$e = \sum_{(i,j) \in \mathcal{O}} (\hat{x}_{ij} - \sum_{k=1}^{q} a^*_{ik} b^*_{jk})^2$$

Esercizio 7

Si consideri un sottoinsieme dei dati X: $n = 1000$ utenti e $p = 14$ film (quelli senza dati mancanti).

Creare un 15% di dati mancanti (a caso).

Imputare i dati mancanti considerando l'algoritmo iterativo. Confrontare l'errore di approssimazione e confrontarlo con quello ottenuto dal modello

$$y_{ij} = \mu + \mu_i + \mu_j + \varepsilon_{ij}$$

dove μ_i è l'effetto dell'i-simo utente e μ_j è l'effetto dell'i-simo film