I dati del Titanic

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

Si consiglia la lettura di Varian (2014) Big Data: New Tricks for Econometrics. In particolare

- l'esempio Titanic (sezione Classification and Regression Trees)
- il codice R utilizzato (potete scaricare il dataset nella sezione Additional Materials)

La competizione Kaggle Titanic: Machine Learning from Disaster. In particolare

- Exploring Survival on the Titanic : è un buon tutorial da cui partire
- Tidy TitaRnic : fornisce un buon esempio di EDA
- Titanic using Name only : fornisce un buon esempio di feature engineering

Table of Contents

Il contesto della classificazione

Siano $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ variabili casuali con distribuzione congiunta (ignota), dove

$Y \in \{0, 1\}$ è una variabile risposta binaria

$X = (X_1, \ldots, X_p)^\top$ sono p predittori

Un classificatore è una funzione $\hat{h} : \mathcal{X} \mapsto \{0, 1\}$. L'errore di classificazione di $\hat{h}$ è definito da

$$\mathrm{Err}(\hat{h}) = \mathbb{P}(Y \neq \hat{h}(X))$$

E' possibile mostrare che l'errore di classificazione è minimizzato dal classificatore di Bayes

$$h_{\text{Bayes}}(x) = \begin{cases} 1 & \text{se } \mathbb{P}(Y = 1 | X = x) > 1/2 \\ 0 & \text{altrimenti} \end{cases}$$

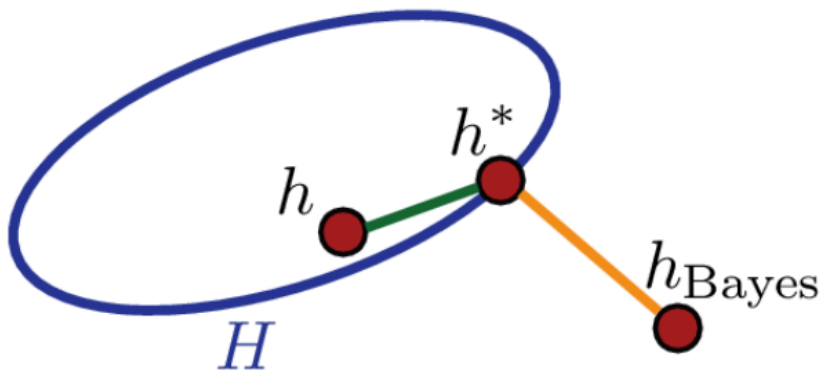
Errore stocastico e di approssimazione

Sia

$$h^* = \arg\min_{h \in \mathcal{H}} \mathrm{Err}(h)$$

dove $\mathcal{H}$ è la classe di classificatori considerata.

L'errore di previsione si può scomporre in *errore stocastico* $\hat{h} - h^*$ ed *errore di approssimazione* $h^* - h_{\mathrm{Bayes}}$



Errore di training e di test

Training set: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Test set: $(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_m^*, y_m^*)$

Errore di classificazione (training set)

$$\mathrm{Err}_{\mathrm{Tr}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{y_i \neq \hat{h}(x_i)\}$$

Errore di classificazione (test set)

$$\mathrm{Err}_{\mathrm{Te}} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{y_i^* \neq \hat{h}(x_i^*)\}$$

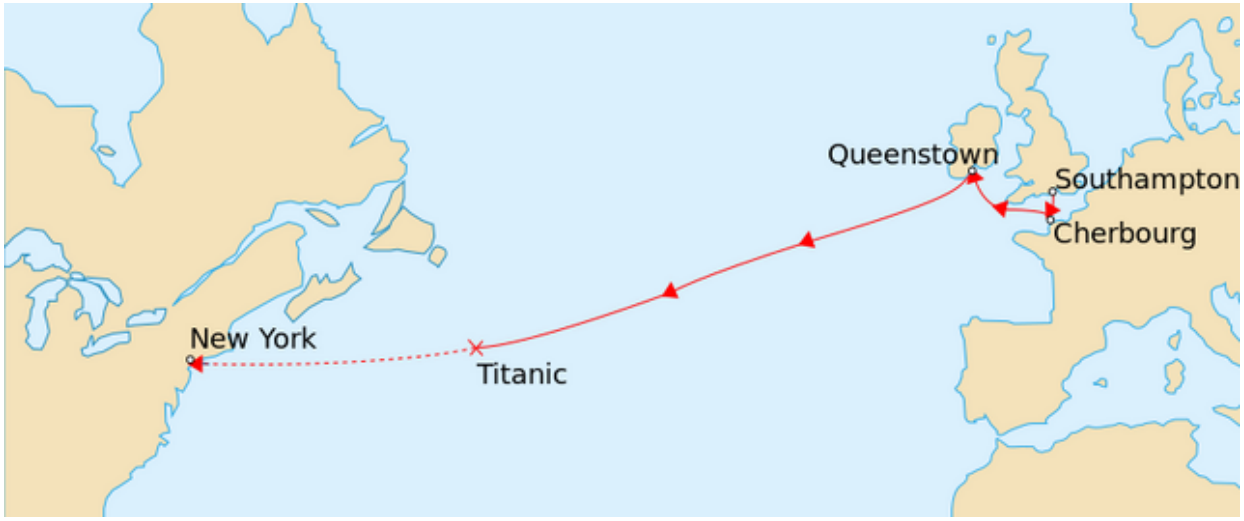
Accuratezza (test set)

$$\mathrm{Acc}_{\mathrm{Te}} = 1 - \mathrm{Err}_{\mathrm{Te}}$$

Table of Contents

Il disastro

Il 15 aprile 1912, durante il suo viaggio inaugurale, il Titanic affondò dopo essersi scontrato con un iceberg, causando la morte di 1502 persone (su 2224 tra passeggeri ed equipaggio)



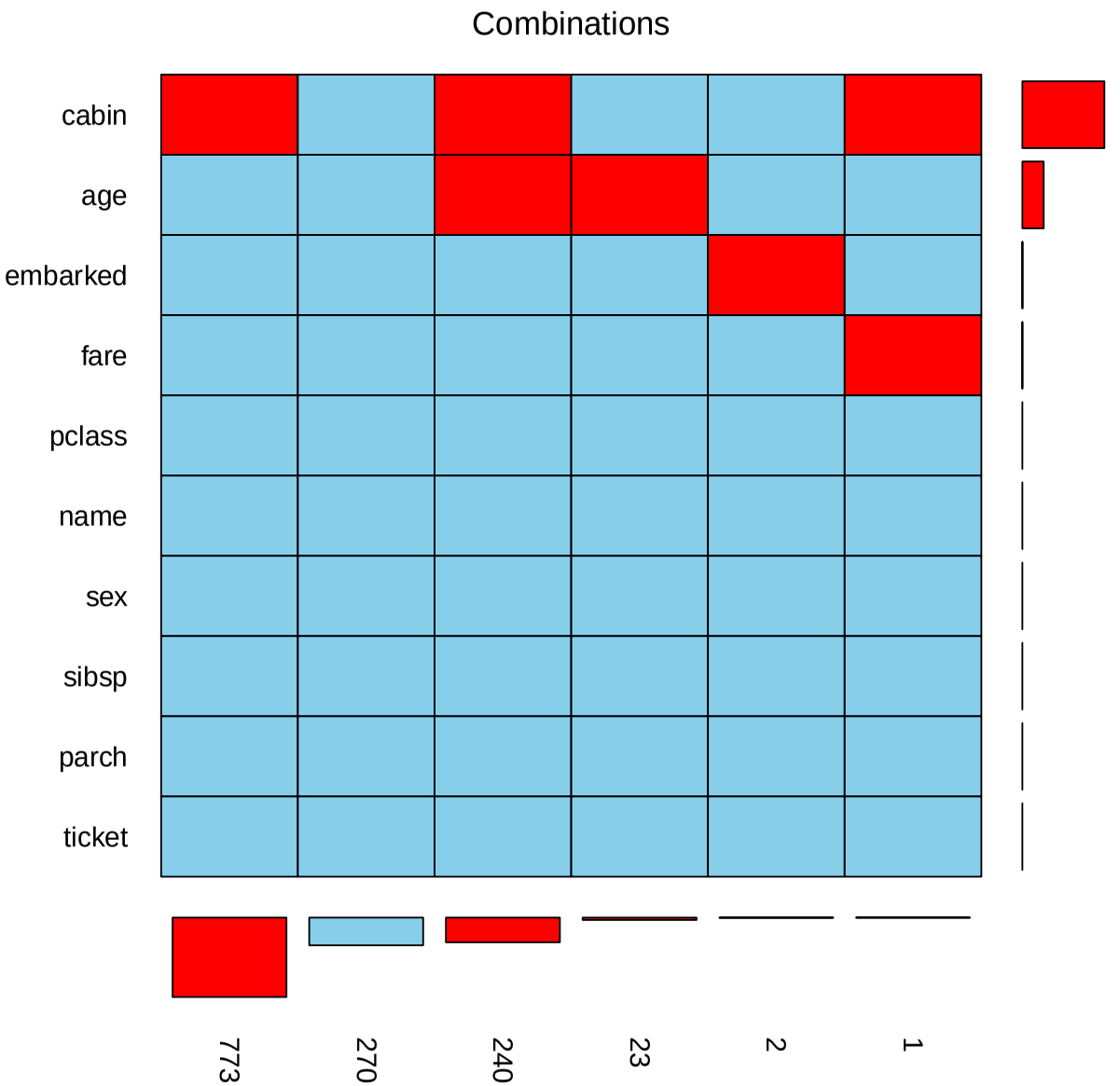
Training set di $n = 891$ passeggeri, sui quali sono state misurate 10 variabili (predittori)

L'obiettivo è prevedere la sorte (1 = sopravissuto, 0 = deceduto) di $m = 418$ passeggeri del test set

```
$ pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
$ survived: int  0 1 1 1 0 0 0 0 1 1 ...
$ name    : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs.
$ sex     : chr  "male" "female" "female" "female" ...
$ age     : num  22 38 26 35 35 NA 54 2 27 14 ...
$ sibsp   : int  1 1 0 1 0 0 0 3 0 1 ...
$ parch   : int  0 0 0 0 0 0 0 1 2 0 ...
$ ticket  : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282"
$ fare    : num  7.25 71.28 7.92 53.1 8.05 ...
$ cabin   : chr  "" "C85" "" "C123" ...
$ embarked: chr  "S" "C" "S" "S" ...
```

Si veda questo file di informazioni sulle variabili

Table of Contents


Combinations
cabin
age
embarked
fare
pclass
name
sex
sibsp
parch
ticket
773
270
240
23
2
1

Tariffa (fare)

```
     pclass survived                 name
1282      3     <NA> Storey, Mr. Thomas

      sex  age sibsp parch ticket fare
1282 male 60.5     0     0   3701   NA

     cabin embarked survived01
1282  <NA>        S          NA
```

Sostituzione del valore mancante

	pclass	embarked	fare
1	1	C	76.7292
2	2	C	15.3146
3	3	C	7.8958
4	1	Q	90.0000
5	2	Q	12.3500
6	3	Q	7.7500
7	1	S	52.0000
8	2	S	15.3750
9	3	S	8.0500

Porto di imbarcazione (embarked)

```
    pclass survived
62       1    Alive
830      1    Alive

                                                name
62                             Icard, Miss. Amelie
830 Stone, Mrs. George Nelson (Martha Evelyn)

       sex age sibsp parch ticket fare
62  female  38     0     0 113572   80
830 female  62     0     0 113572   80

    cabin embarked survived01
62    B28     <NA>           1
830   B28     <NA>           1
```

Table of Contents

Modello nullo

Training set: il 38.38% dei passeggeri è sopravvissuto

Il modello nullo utilizza solo y e prevede tutti i passeggeri del test set nella classe ”non sopravvissuto”

Accuratezza delle previsioni sul test set : 62.2%

Genere (sex)