

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

Problema di classificazione

I dati

Valori mancanti

Analisi esplorativa

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

$X = (X_1, \dots, X_p)^\top$ 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

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

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

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

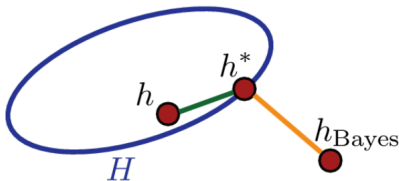
Errore stocastico e di approssimazione

Sia

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

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* $h^* - h_{\text{Bayes}}$



Errore di training e di test

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

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

Errore di classificazione (training set)

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

Errore di classificazione (test set)

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

Accuratezza (test set)

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

Table of Contents

Problema di classificazione

I dati

Valori mancanti

Analisi esplorativa

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 ($1 =$ sopravvissuto, $0 =$ 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

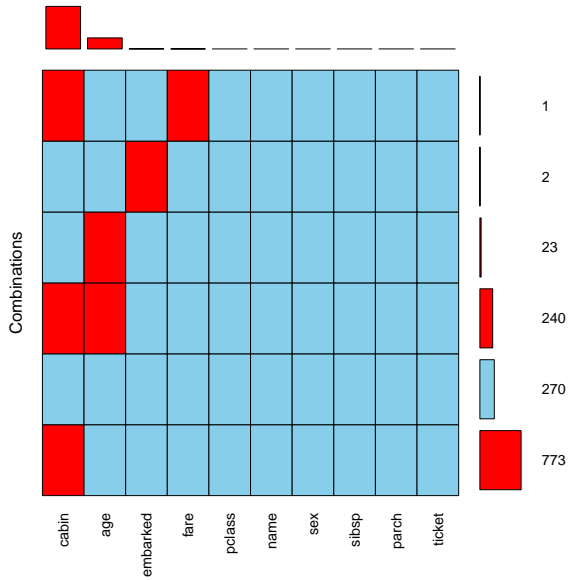
Table of Contents

Problema di classificazione

I dati

Valori mancanti

Analisi esplorativa



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

Problema di classificazione

I dati

Valori mancanti

Analisi esplorativa

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"

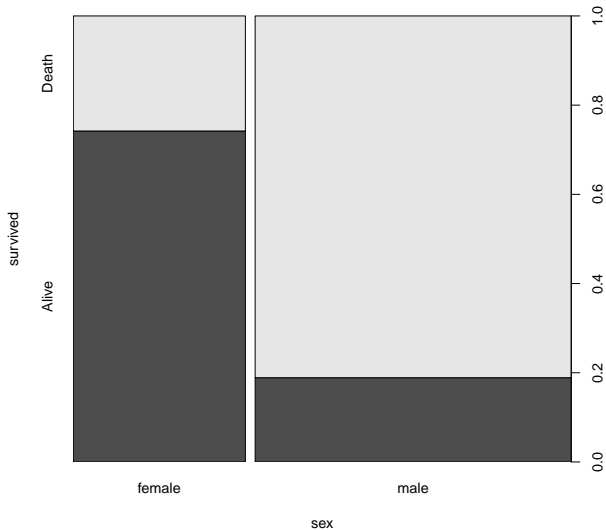
Accuratezza delle previsioni sul test set : 62.2%

Tabella di confusione :

	Death	Alive
Death	260	158

Genere (sex)

Prima le donne?



Modello con solo il genere

Questo modello classifica i passeggeri del test set in funzione del sesso: se donna, sopravvissuta, se uomo, deceduto.

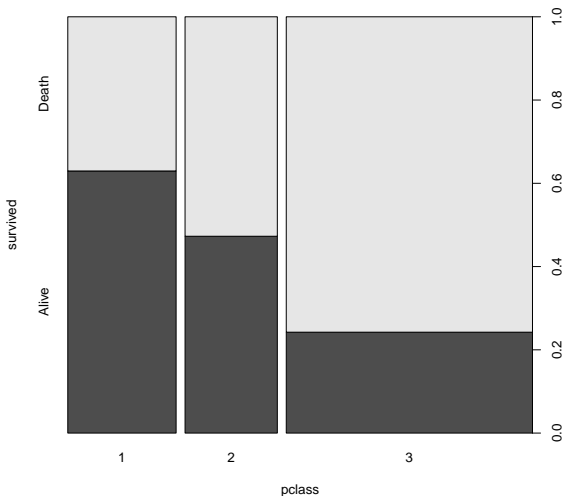
Tabella di confusione :

	Death	Alive
Alive	46	106
Death	214	52

Accuratezza : 76.5%

Classe (pclass)

I passeggeri che viaggiavano in prima classe hanno avuto maggiore probabilità di sopravvivenza?



Modello con solo la classe

Questo modello classifica i passeggeri del test set in funzione della classe: se prima, sopravvissuto, se seconda o terza, deceduto.

Tabella di confusione :

	Death	Alive
Alive	43	64
Death	217	94

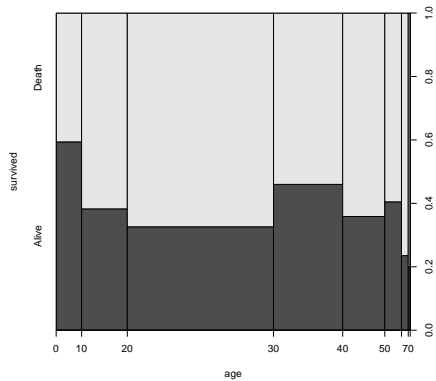
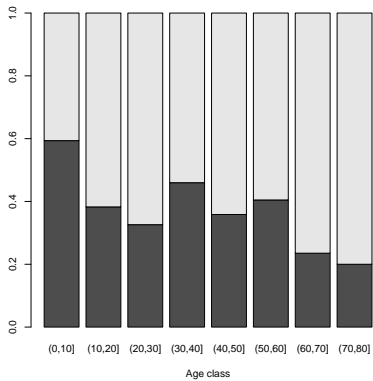
Accuratezza : 67.2%

Età (age)

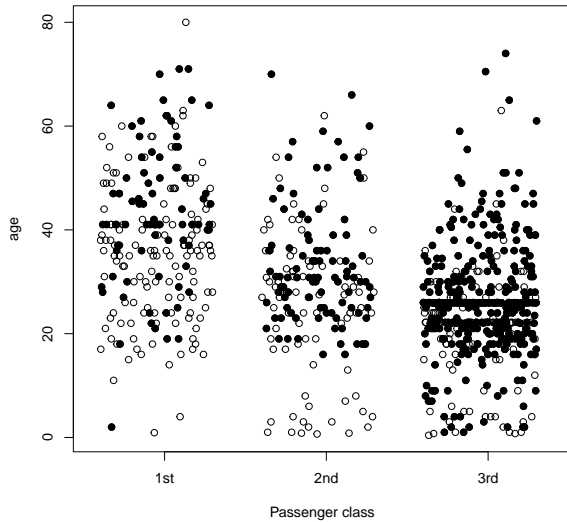
Prima i bambini? Qual è la relazione tra età e sopravvivenza?

```
glm(survived ~ age, train, family="binomial")
```

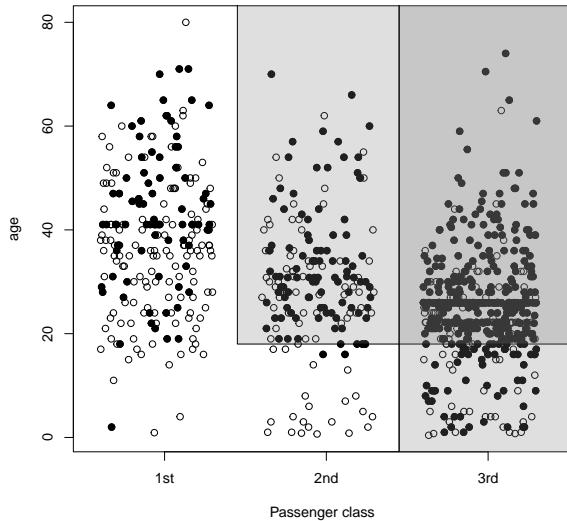
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.19	0.17	-1.16	0.25
age	-0.01	0.01	-1.83	0.07



Età e classe

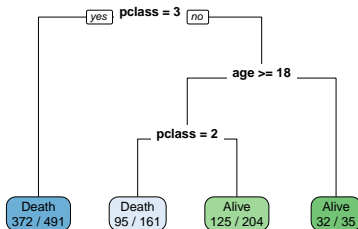


Albero di classificazione



Regola decisionale

Status	Pr(Death)	Prediction
Class 3	76%	Death
Class 1-2, younger than 18	9%	Alive
Class 2, older than 18	56%	Death
Class 1, older than 18	39%	Alive



Albero di classificazione con età e classe

Tabella di confusione :

	Death	Alive
Death	215	86
Alive	45	72

Accuratezza : 68.6%

Modello logistico

Predictors	Acc.Tr	Acc.Te
1	61.6%	62.2%
age	61.6%	62.2%
pclass	67.9%	67.2%
sex	78.7%	76.6%
age + pclass	69.1%	65.3%
age + sex	78.7%	76.6%
pclass + sex	78.7%	76.5%
age + pclass + sex	79.5%	75.6%

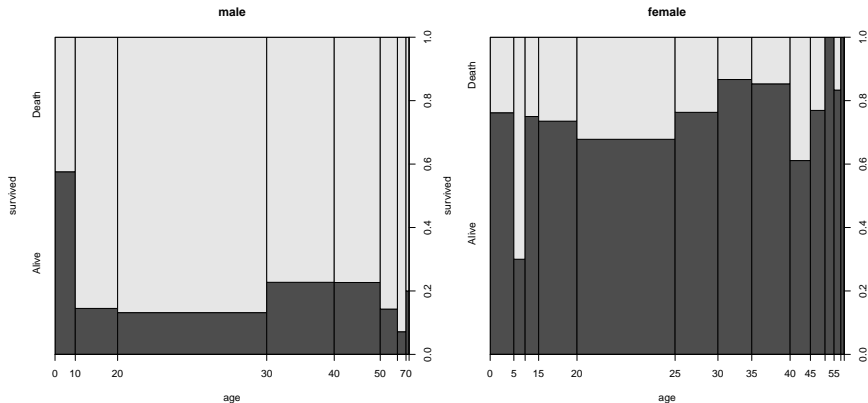
Come migliorare il modello gender-only?

Per migliorare il modello sex-only, dobbiamo capire

- quali maschi sopravvivono
- quali femmine muoiono

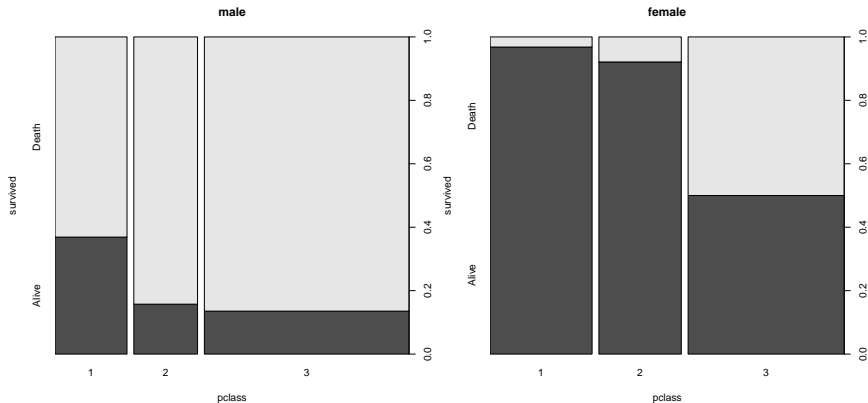
Età e genere

Molti maschi sopravvissuti sono tra i giovani



Classe e genere

La maggior parte delle femmine che muoiono viaggiano in terza classe

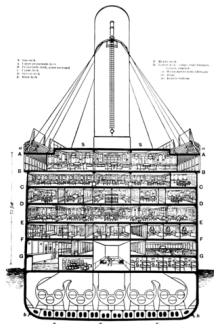


```
# 21/40 maschi sotto i 16 anni sopravvivono
table(train$survived[train$sex=='male' & train$age<16])
Death Alive
    19    21
```

```
# 72/144 femmine che viaggiano in III classe non sopravvivono
table(train$survived[train$sex=='female'&train$pclass==3])
Death Alive
    72    72
```

Cabina

```
# the first character of cabin is the deck  
table(substr(combi$cabin, 1, 1))
```



Titolo

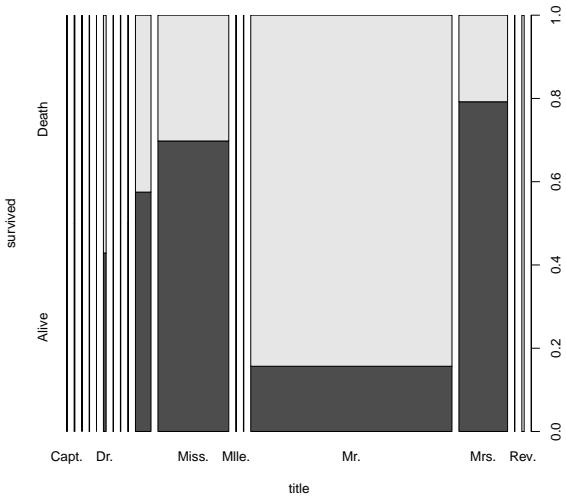
```
combi$name[1]
```

```
"Braund, Mr. Owen Harris"
```

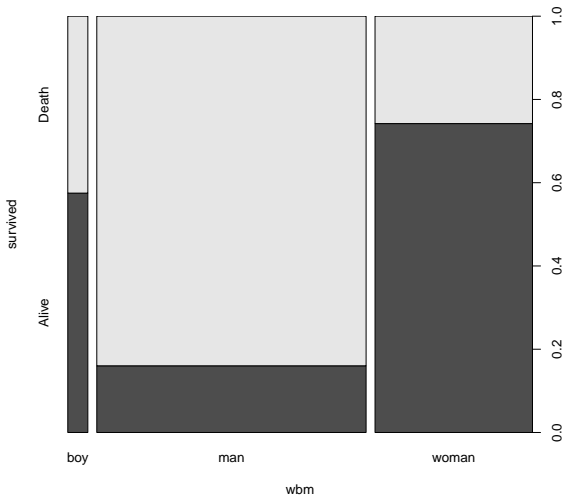
```
table(combi$title)
```

Capt.	Col.	Countess.
1	4	1
Don.	Dona.	Dr.
1	1	8
Jonkheer.	Lady.	Major.
1	1	2
Master.	Miss.	Mlle.
61	260	2
Mme.	Mr.	Mrs.
1	757	197
Ms.	Rev.	Sir.
2	8	1

Titolo



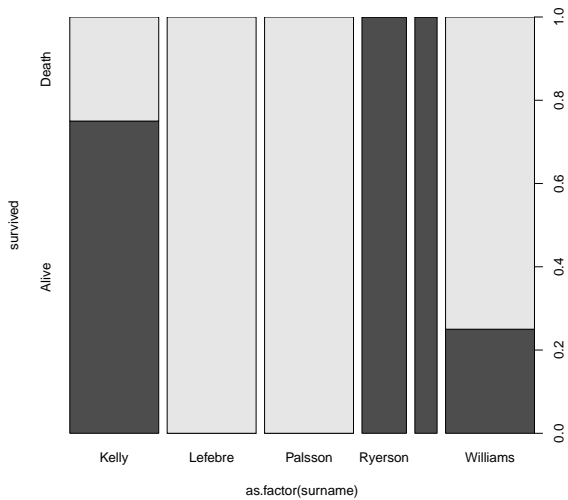
Uomo, ragazzo e donna



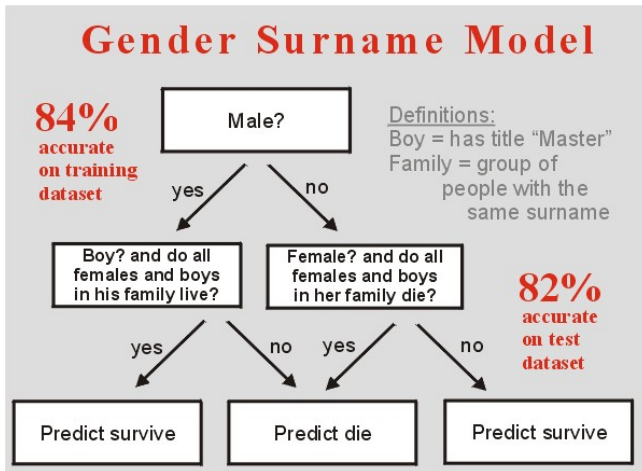
Distribuzione di frequenza dei cognomi

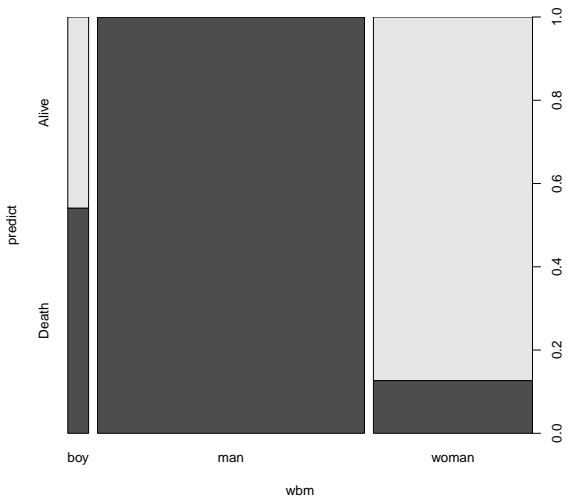
```
# famiglie con almeno 5 componenti  
table(combi$surname[combi$surnameFreq>4])
```

Andersson	Asplund	Brown
11	8	6
Carter	Davies	Ford
6	7	6
Fortune	Goodwin	Johnson
6	8	6
Kelly	Lefebvre	Palsson
5	5	5
Panula	Rice	Ryerson
6	6	5
Sage	Skoog	Smith
11	6	6
Thomas	Williams	
5	5	



Modello Gender Surname





Modello Gender Surname

Training set - tabella di confusione :

	Death	Alive
Alive	31	253
Death	518	89

Training set - Accuratezza : 86.5%

Test set - tabella di confusione :

	Death	Alive
Alive	38	113
Death	222	45

Test set - Accuratezza : 80.1%

Confronto tra modelli

Model	Acc.Tr	Acc.Te
All-dead	61.6%	62.2%
Gender-only	78.7%	76.6%
Gender surname	85.5%	80.1%