Data Science

Esercizi primo compitino - serie 2 - 2019

(serie risolta)

Esercizio 1

Una squadra professionistica di pallavolo ha implementato un sistema di raccolta dati in uso durante le partite della stagione. Per ogni tocco di palla di uno dei propri giocatori, una persona dotata di un apposito software registra:

- L'istante esatto del tocco (data e ora, con precisione al millisecondo)
- Il numero del giocatore
- Il tipo di tocco (battuta, ricezione, alzata, schiacciata, muro)
- Una valutazione numerica da 0 (pessimo) a 5 (ottimo) relativa all'abilita' mostrata dal giocatore in quel tocco.

Che aspetto ha una tabella che contiene questo dataset? Definisci cosa sono le osservazioni e quali sono gli attributi; Se il dataset fosse salvato su un file CSV, mostra un esempio delle prime tre righe.

```
timestamp,numero,tocco,valutazione
10012,1,battuta,3.5
21033,5,ricezione,3.0
22161,4,alzata,3.0
...
```

Stima a grandi linee quanto spazio di memoria occuperanno i dati di una singola partita.

10 byte per osservazione * 10 tocchi per punto * 100 punti/partita = circa 10kByte

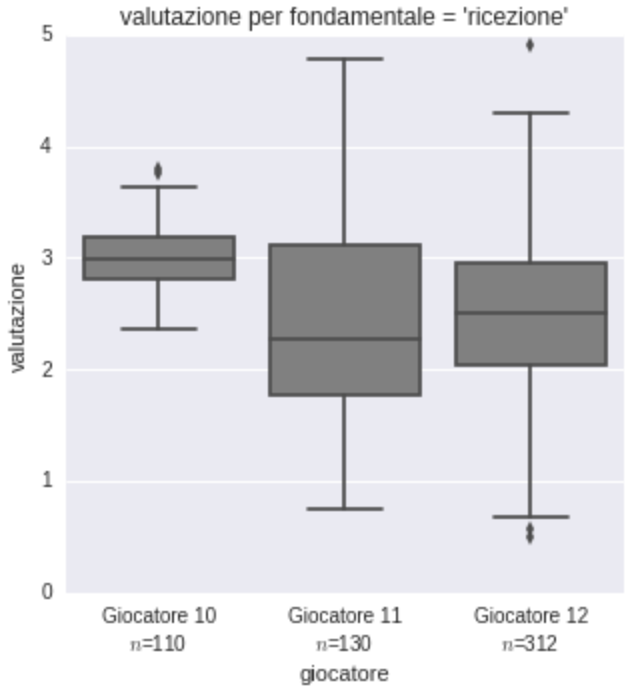
Descrivi brevemente e in modo intuitivo come l'allenatore puo' elaborare i dati in modo da identificare quale tra i propri giocatori e' il miglior battitore.

L’allenatore puo’ filtrare il dataset sull’attributo “tocco” = “battuta”. Sul dataset filtrato risultante, puo’ raggruppare i dati per giocatore, e calcolare la media della valutazione per ciascuno. In alternativa, puo’ risultare informativo un boxplot che invece che la sola media riporti informazioni piu’ ricche sulla distribuzione delle valutazioni per ciascun giocatore

Il prossimo weekend, l’allenatore ha la possibilita’ di mandare tre dei propri giocatori a un workshop in cui viene migliorata l’abilita’ in ricezione. Descrivi e motiva un possibile modo con cui l’allenatore puo’ scegliere quali tra i suoi giocatori mandare.

Assumendo che l’allenatore voglia uniformare le abilita’ della squadra: puo’ selezionare i tre giocatori con la valutazione media in ricezione piu’ bassa.

Il software produce il seguente boxplot relativo alle valutazioni di ciascun giocatore per il fondamentale “ricezione”. I baffi (whisker) si estendono fino a max 1.5 volte il range interquartile (convenzione di Tukey, come definito nelle slide), e tutti i datapoint che al di fuori di questo range sono rappresentati con dei piccoli rombi (fliers). Il numero di osservazioni per ciascun giocatore e’ riportato nelle etichette dell’asse x.



Rispondi alle seguenti domande e motiva brevemente:

- Quale giocatore ha fatto la ricezione migliore in assoluto? E quella peggiore in assoluto?

Entrambe il giocatore 12

- Quali giocatori hanno ottenuto una valutazione inferiore a 2 per piu' della meta' delle loro ricezioni?

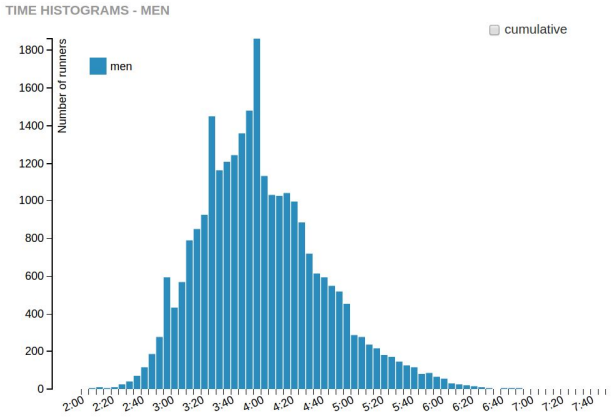
Nessuno

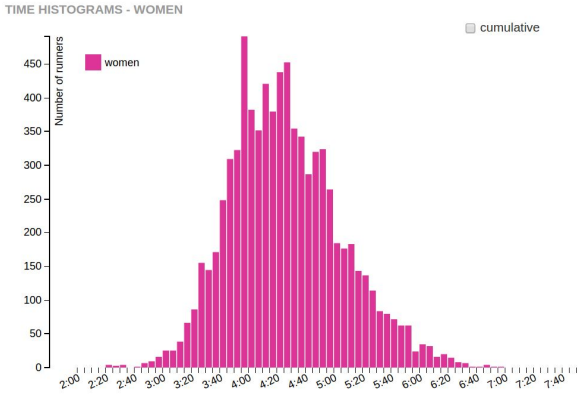
- Approssimativamente quante ricezioni con valutazione superiore al 3 ha fatto il giocatore 12?

Circa ¼ delle proprie ricezioni (78)

Esercizio 2

Si considerino i seguenti istogrammi, che riportano il tempo ottenuto durante la maratona di Berlino del 2012 da atleti maschi e femmine.





- Quanti sono approssimativamente i maschi e le femmine che hanno concluso la gara? (serve solo un ordine di grandezza, non un valore esatto).

Nell'ordine dei 20000 i maschi, 6000 le femmine

- Quanto e' ampio ciascun intervallo (bin) dell'istogramma?

5 minuti

- Considera solo gli atleti maschi: rispondi alle seguenti domande (dove e' possibile farlo con i dati a disposizione) e motiva:
    - Sono arrivati piu' atleti con un tempo inferiore a 3:00 di quanti ne siano arrivati con un tempo superiore
    - Sono arrivati piu' atleti con un tempo compreso tra 2:55 e 3:00 di quanti ne siano arrivati in un tempo compreso tra 3:00 e 3:05
    - Piu' di 300 atleti sono arrivati con un tempo compreso tra 3:00 e 3:01

- No, perche' la somma delle barre a sinistra di x=3:00 e' minore della somma delle barre a destra
- Si', perche' la barra subito a sinistra di x=3:00 e' piu' alto di quella subito a destra
- Non possiamo dirlo dai dati a disposizione

- Rispondi alle stesse domande considerando gli atleti indipendentemente dal sesso

- No, perche' la somma delle barre a sinistra di x=3:00 e' minore della somma delle barre a destra tanto per i maschi quanto per le femmine
- Si': il contributo delle atlete femmine e' trascurabile in questo caso (sono poche decine)
- Non possiamo dirlo dai dati a disposizione

- Sia per gli atleti maschi che per le atlete femmine, il grafico mostra un notevole cambiamento in corripondenza del tempo di 4:00 e di 5:00. Proponi una sintetica spiegazione del fenomeno.

Una possibile spiegazione e' che gli atleti si allenano (e pianificano la propria gara) per battere un tempo "intero" simbolico.
Lo stesso fenomeno e' visibile per il tempo di 3:00 e 3:30

Esercizio 3

Ci viene fornito un dataset contenente la concentrazione nel sangue di una determinata sostanza in 25 persone diverse.

ID Paziente	**Concentrazione**
0	5.194
1	5.114
2	5.066
3	5.166
4	4.853
5	5.096
6	5.088
7	5.159
8	4.964
9	5.006
10	4.994
11	4.952
12	5.145
13	5.196
14	5.162
15	5.107
16	5.142
17	5.048
18	5.111
19	5.025
20	5.058
21	5.012
22	5.058
23	5.611
24	5.501

La media campionaria e' 5.113 e la varianza campionaria 0.024746.

- Calcola scarto e z-score per le ultime 5 osservazioni e commenta il risultato.

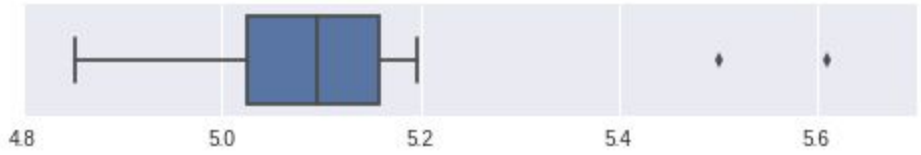
	20	21	22	23	24
scarto	-0.055	-0.101	-0.055	0.498	0.388
zscore	-0.350	-0.643	-0.350	3.165	2.466

I valori di zscore per le righe 23 e 24 sono entrambi molto alti e suggeriscono di verificare attentamente i dati

- Osserva il resto del dataset: senza fare calcoli, pensi che esistano altre osservazioni con uno z-score ancora piu' alto in valore assoluto? Perche'?

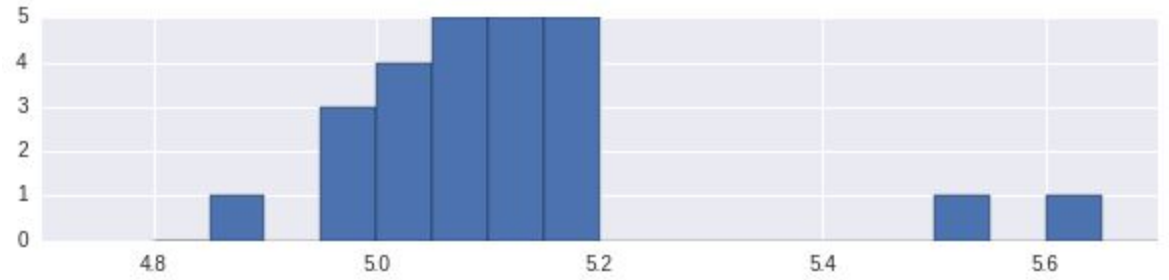
Possiamo escluderlo, perche' non vediamo valori piu' alti di 5.6, ne' piu' bassi di 5.113-0.498=4.617 (ovvero il valore che dista dalla media, per difetto, tanto quanto 5.6.

- Rappresenta tutti i dati tramite un boxplot con i baffi che si estendono per max 1.5 volte il range interquartile, ed eventuali outlier rappresentati tramite puntini (fliers).

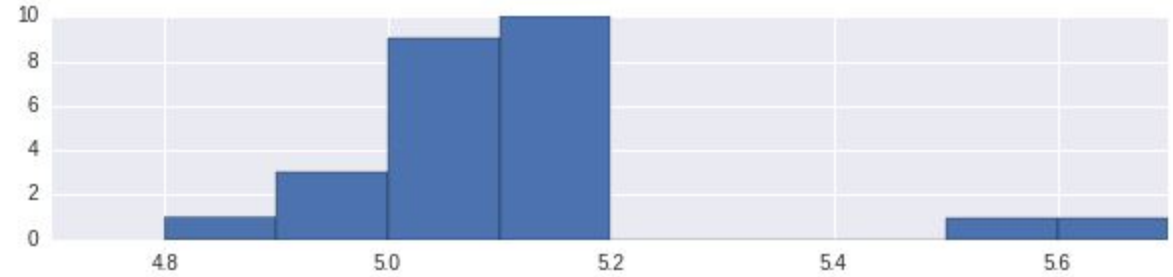


- Scegli un intervallo di discretizzazione ragionevole e rappresenta i dati tramite istogramma

Discretizzando con un intervallo di 0.05:

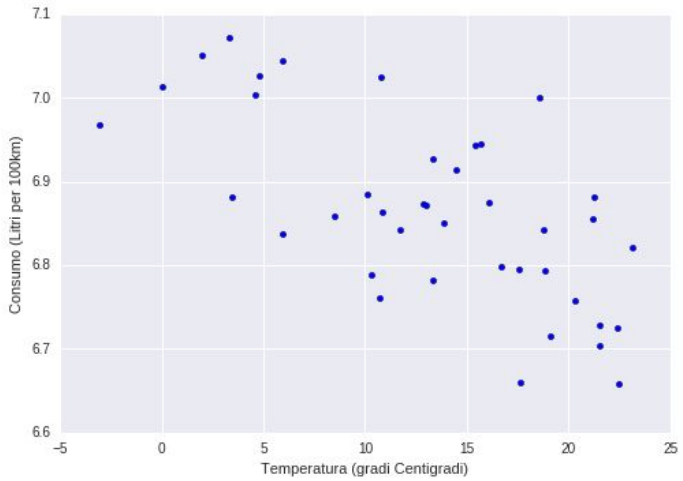


Discretizzando con un intervallo pari a 0.10:



Esercizio 4

Una casa automobilistica vuole verificare la relazione tra temperatura ambientale e consumo di carburante, svolgendo una serie di 50 test durante l'anno. Per ogni test viene registrata la temperatura media e il consumo medio su un percorso di 300km, sempre utilizzando lo stesso veicolo. I dati sono riassunti nel seguente scatter plot. (attenzione: i dati non sono identici a quelli presentati in classe, sono stati rigenerati in seguito)



- Discuti brevemente i risultati del test.

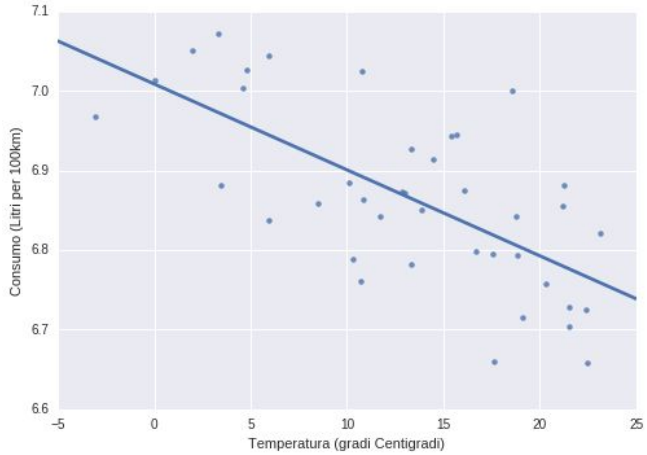
E’ evidente una correlazione negativa tra temperatura e consumo (all’aumentare dell’una decresce l’altro).

- Senza fare calcoli, stima il coefficiente di correlazione tra l’attributo temperatura e l’attributo consumo.

A occhio, il coefficiente parrebbe compreso tra -0.5 e -0.8. (Valore vero: -0.67)

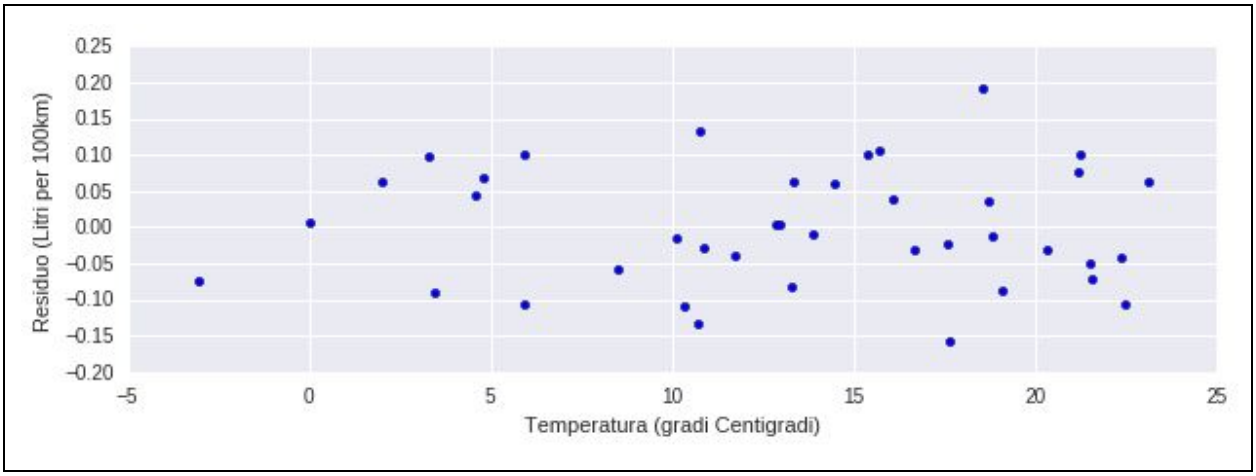
- Senza fare calcoli, stima intercetta e pendenza della retta che si otterrebbe effettuando una regressione lineare semplice con “Temperatura” come attributo esplicativo e “Consumo” come attributo target

Attezione: l’intercetta corrisponde con il valore della retta corrispondente a x=0 (ovvero, temperatura uguale a 0), e si puo’ esprimere nella stessa unita’ di misura presente sull’asse y. La pendenza si esprime in questo caso in (L per 100km / gradi celsius)
Intercetta: 7.01 L per 100km
Pendenza: -0.011 L per 100km / gradi celsius



- Descrivi brevemente che aspetto avrebbe il grafico dei residui.

I residui sono equamente distribuiti tra positivi e negativi. La distribuzione e’ apparentemente omoschedastica e non presenta particolari pattern.



- Come cambiano le risposte alle precedenti due domande se invece che in gradi centigradi, la temperatura venisse espressa in gradi Farenheit? Le due temperature sono legate da questa relazione: ℉ =℃ * 1.8000 + 32.00

Il coefficiente di correlazione non cambia, perche' e' adimensionale e non dipende dalle unita' di misura dei valori che si considerano (e' invariante a qualsiasi trasformazione lineare) dei dati.
Intercetta e pendenza cambiano. In particolare la pendenza verra' moltiplicata di un fattore (1/1.8).

Esercizio 5

Ci viene fornito un dataset contenente 1000 osservazioni e 101 attributi numerici. Uno di tali attributi rappresenta la variabile target, e gli altri 100 sono candidati ad essere variabili esplicative.

Vengono forniti tre modelli di regressione lineare multipla, ciascuno ottenuto utilizzando un diverso sottoinsieme dei 100 attributi esplicativi:

Modello	Coefficiente di determinazione R^2 multiplo
Ma	0.8120
Mb	0.`
Mc	0.8521

Domande:

- Spiega in modo intuitivo che relazione esiste tra la varianza della variabile target, la varianza dei residui ottenuti dal modello Ma e la varianza dei residui ottenuti dal modello Mc.

R^2 corrisponde alla frazione della varianza della variabile target spiegata dal modello. Per Mc, l'85.21% della varianza viene spiegata dal modello. Il rimanente 14.79% della varianza rimane nei residui. Quando $R^2 = 1$, il modello spiega perfettamente i dati, e infatti i residui hanno varianza pari a 0. Quando $R^2 = 0$, il modello non e' in grado di spiegare neanche un po' della varianza della variabile target; essa rimane invariata nei residui.

- Possiamo concludere che il modello Mc funzionerebbe meglio del modello Ma se lo applicassimo a un nuovo dataset diverso da quello utilizzato per calcolare i coefficienti?

No, non necessariamente. Il valore R^2 ci indica quanto bene il modello funziona sul dataset utilizzato per l'addestramento

- Possiamo concludere che Mb usa piu' variabili esplicative rispetto ad Ma?

No, non necessariamente: potrebbe essere che Mb utilizzi variabili esplicative piu' correlate con la variabile target rispetto a quelle utilizzate da Ma.

-