

Insegnamento di "Data mining" Prova d'esame del 26 marzo 2013

1. Spiegare perché la valutazione di una certa tecnica viene comunemente effettuata su un insieme di dati diverso da quello su cui si è compiuto l'addestramento (cioè la "stima") della tecnica.
2. La curva "lift" rappresenta un criterio di valutazione più articolato della semplice frazione totale di errore. Chiarire quali informazioni aggiuntive offre il "lift".
3. Presentare in modo sintetico gli elementi essenziali delle "spline di regressione".
4. Si spieghi cosa si intende per livello di significatività osservato.

5. In una indagine sul consumo di sigarette condotta negli Stati Uniti, per ciascuno degli stati e per il distretto di Washington sono state osservate le seguenti variabili: consumo pro capite di sigarette (y), età media dei residenti (age), frazione di popolazione che ha completato gli studi superiori (edu), reddito annuo pro capite (inc), frazione di popolazione di colore (blk), frazione di popolazione femminile (fem), prezzo medio di un pacchetto di sigarette (prc). Scopo dell'indagine è spiegare come il consumo dipenda dalle variabili considerate.

- (a) Un primo modello di regressione tra il consumo e la percentuale di popolazione femminile ha fornito il seguente output.

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	-93.4260	207.8126	-0.4496
fem	4.2191	4.0777	1.0347

Residual standard error: 32.05 on 49 degrees of freedom

Multiple R-Squared: 0.02138

- i. Si scriva il modello statistico corrispondente.
- ii. Si calcolino le stime dell'intercetta e del coefficiente angolare che si sarebbero ottenute utilizzando come variabile esplicativa la frazione di popolazione maschile, ovvero la variabile 1-fem.

iii. Si calcoli il coefficiente di correlazione lineare semplice tra y e fem.

- (b) Un secondo modello, costruito utilizzando tutte le variabili esplicative, ha fornito il seguente output.

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	109.6822	245.7455	0.4463
age	4.6464	3.2034	1.4505
edu	-0.0580	0.8120	-0.0714
inc	0.0189	0.0102	1.8584
blk	0.3699	0.4871	0.7595
fem	-1.2411	5.5645	-0.2230
prc	-3.2629	1.0302	-3.1674

Residual standard error: 28.13 on 44 degrees of freedom

Multiple R-Squared: 0.3228

iv. Si scriva il modello statistico corrispondente.

v. Si specifichi l'ipotesi statistica relativa all'affermazione: "Le variabili age, edu, inc, blk e prc non hanno effetto sul consumo".

vi. Come si può condurre il test per verificare l'ipotesi statistica di cui al punto precedente? Puoi fornire il valore della statistica test?

vii. Si dica quali variabili hanno un effetto statisticamente significativo sul consumo di sigarette.

6. (facoltativo)

Sia X una variabile osservata su n unità. Si dimostri il seguente risultato

$$D = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 = 2 \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = 2\sigma^2$$

dove m e σ^2 sono rispettivamente media aritmetica e varianza della variabile X.