**Modelli Statistici e Statistical Learning - 2022/2023**

RELAZIONE ESERCIZIO 2

Gruppo SparadAIs

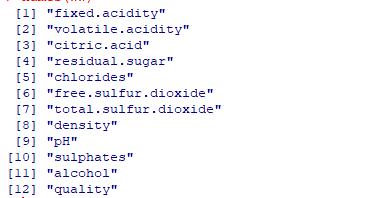


PRESENTAZIONE

L’esercizio consiste nel realizzare un modello predittivo sulla qualità del vino bianco. Per la realizzazione ci siamo serviti dei metodi di regolarizzazione studiati a lezione: Ridge Regression, LASSO ed Elastic Net. Nella relazione sono descritti i passaggi nella realizzazione di vari modelli basati sui metodi sopra citati. Al termine di ciò, valutando l’MSE minore, si è selezionato il modello migliore.

IL DATASET

Il dataset è composto da 4898 osservazioni che descrivono la qualità e le caratteristiche di un certo tipo di vino bianco. Di seguito sono mostrate le caratteristiche:



In particolare, consideriamo ‘quality’ come la nostra variabile indipendente, la quale rappresenta la qualità del vino bianco come un punteggio da 0 a 10. Le altre 11 caratteristiche costituiscono invece i regressori del modello da stimare.

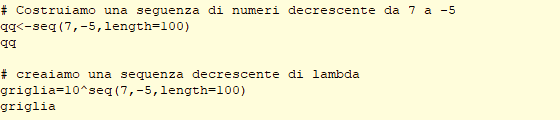
Innanzitutto, leggiamo i dati in R e costruiamo le strutture necessarie.

Immagine che contiene testo

Descrizione generata automaticamente

La scelta del migliore modello predittivo si basa sull’errore quadratico medio minimo misurato sul test set. Tale modello viene selezionato anche attraverso l’utilizzo delle tecniche di cross validation, le quali permettono di ottenere diversi modelli di apprendimento in base a come si suddivide l’insieme delle osservazioni in test set e training set. Inoltre, quando si utilizzano le tecniche di regolarizzazione, la selezione del modello migliore per un data set dipende anche dalla scelta del parametro di penalità . Fissato il valore di , scegliamo il modello che minimizza MSE. Per questo discretizziamo e lo facciamo variare in un range di valori fissati.

Definiamo quindi una sequenza decrescente di 100 valori di .



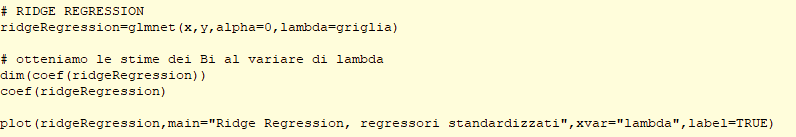
Questi valori di lambda saranno i valori tenuti in considerazione per tutti i successivi metodi di regolarizzazione.

La standardizzazione dei dati viene eseguita in automatico dal comando glmnet.

Abbiamo scelto di settare il seed a 31.

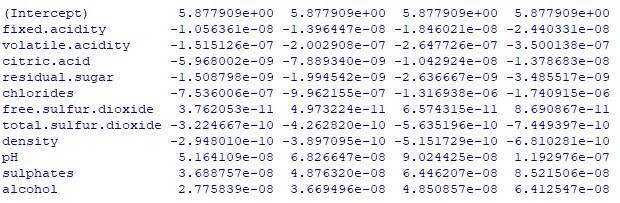
RIDGE REGRESSION

Il primo metodo di regolarizzazione sperimentato è la Ridge Regression. Essa opera un restringimento della stima dei parametri rispetto alla stima ottenuta con i minimi quadrati a causa del parametro di penalità considerato.



Per ogni valore di si ottiene un vettore di stime dei parametri. Per i primi quattro valori di , pari a 

si ottengono le seguenti stime:



Il grafico seguente riassume l’andamento dei valori dei parametri al variare del valore di lambda (nel caso del grafico si fa riferimento al log(lambda)):



Ogni curva descrive la variazione del coefficiente al variare di lambda. Si può notare come lo stimatore Reidge Regression operi un restringimento delle stime rispetto alle stime ottenute con lo stimatore ai minimi quadrati. Quando le stime dei coefficienti tendono a zero, mentre quando si ottengono le stime ai minimi quadrati.

A questo punto siamo a interessati a trovare quel valore di , tale per cui l’MSE del modello è minimo. Per fare ciò ci serviamo di due metodologie di Cross-Validation: il K-Fold CV e il LOOCV.

Immagine che contiene testo

Descrizione generata automaticamente

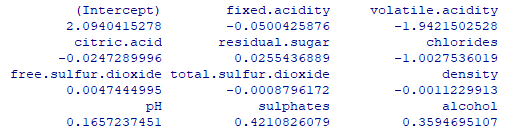
I risultati ottenuti con K-Fold sono i seguenti:



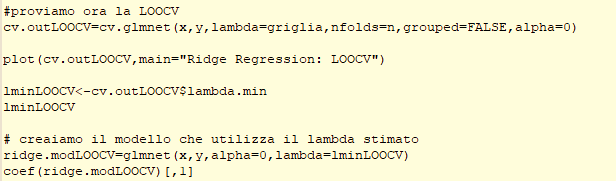
Il lambda minimo ottenuto è pari a 0.004641589.



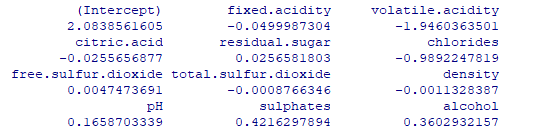
I coefficienti ottenuti stimando il modello con un lambda pari a 0.004641589 sono:



Con la tecnica LOOCV otteniamo i seguenti risultati:



In questo caso il minimo valore di lambda ottenuto è 0.003511192 .





LASSO

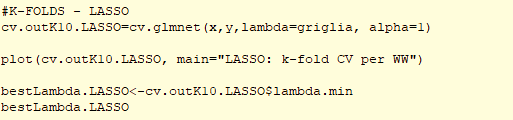
Proseguiamo la stima del modello con il metodo LASSO. A differenza del precedente (Ridge Regression), LASSO opera, non solo un restringimento sui parametri, ma anche una selezione dei regressori.

Immagine che contiene testo

Descrizione generata automaticamente



Come fatto in precedenza, per la stima del modello operiamo sia una K-Fold che una LOOCV:

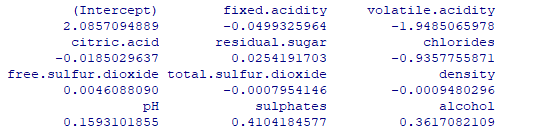


In questo caso il lambda minimo ottenuto è 0.001149757.



Immagine che contiene testo

Descrizione generata automaticamente

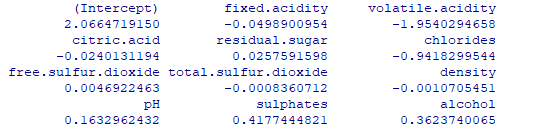


Ora proviamo con la LOOCV:

Con la LOOCV otteniamo il seguente lambda: 0.0004977024.



I parametri stimati sono i seguenti:



//todo confronto tra RIDGE k-fold e LASSO k-fold (con alcuni lambda si annullano i parametri density e citric.acid nella LASSO).

ELASTIC NET

Elastic Net opera una via di mezzo tra il metodo Ridge Regression e il metodo LASSO. Elastic Net si può anche vedere come una generalizzazione di entrambi gli approcci, infatti grazie al parametro “alpha” si attribuisce più peso alla penalità di un metodo e meno peso all’altro. In particolare, si osserva che per alpha=0, il modello si riduce alla Ridge Regression, mentre per alpha=1 ci si riconduce alla LASSO. Siccome i casi particolari per alpha=0 e alpha=1 sono già stati visti, operiamo una variazione dei valori di alfa che oscillano tra 0 e 1.

Abbiamo sperimentato degli alfa crescenti utilizzando sempre la K-fold:

Per alpha=0.1, quindi molto vicini alla Ridge Regression:

//plot parametri



//plot lambda



//valore di lambda

0.006135907.

Per alpha=0.3:

//plot parametri



//plot lambda



//valore di lambda

0.004641589.

Per alpha=0.5, quindi un modello intermedio tra Ridge Regression e LASSO:

//plot parametri



//plot lambda



//valore di lambda

0.002656088.

Per alpha=0.7:

//plot parametri



//plot lambda



//valore di lambda

0.002656088.

Per alpha=0.9, quindi molto vicini alla LASSO:

//plot parametri



//plot lambda



//valore di lambda

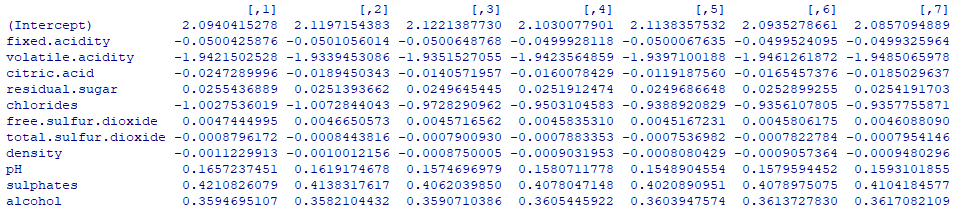
0.001519911.

SELEZIONE DEL MODELLO

Realizzati i modelli con i metodi di regolarizzazione citati, siamo pronti per la selezione del modello migliore andando a confrontare tutti i modelli ottenuti.

Di seguito riportiamo tutte le stime dei parametri ottenuti variando alpha:

//foto parametri ottenuti

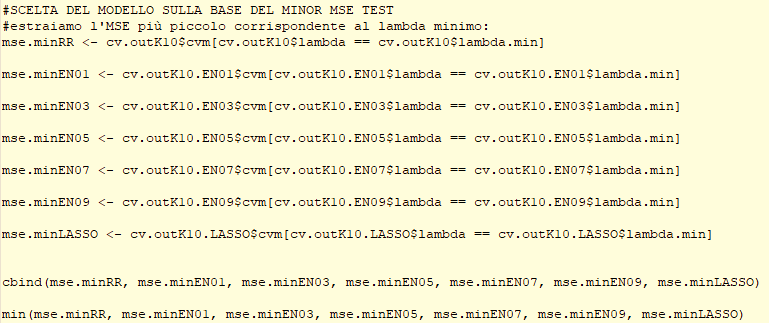


//da aggiungere dopo

Una cosa interessante che si può notare è che con Elastic Net emerge una operazione di selezione sui regressori “progressiva”. Infatti, per alfa=0.1, quindi vicino alla Ridge Regression non viene fatta alcuna opera di selezione, per alfa=0.3 invece emerge il primo regressore per cui il parametro associato si annulla (‘citric acid’). Infine, per alpha=0.7 emerge l’ultimo regressore da non selezionare, che è ‘density’, esattamente come ci aspettiamo dal metodo LASSO (alpha=1).

//

Per selezionare il modello migliore, a questo punto, confrontiamo tutti i valori degli MSE ottenuti dai rispettivi modelli per i rispettivi valori di lambda che minimizzano il proprio MSE:





Risulta evidente dai dati ottenuti, che il miglior metodo di regolarizzazione per questo specifico dataset è Ridge Regression.