Data Analysis

Faiella Ciro, Giannino Pio Roberto, Scovotto Luigi and Tortora Francesco

{c.faiella8, p.giannino, l.scovotto1, f.tortora21}@studenti.unisa.it

Jan, 2023
Department of Computer Engineering, Electrical Engineering and Applied Mathematics (DIEM), University of Salerno, Fisciano, Italy

1 Pt. 1 - R

1.1 Analisi preliminare

1.1.1 Dataset

Il dataset è composto da 70 osservazioni *(n=70)* di una variabile dipendente Y e di 50 regressori *(p = 50)*.

1.1.2 Correlazione

Per cominciare definiamo il termine *correlazione*:

La correlazione è una misura della relazione lineare tra due variabili quantitative. La correlazione varia da -1 a 1, dove -1 indica una relazione negativa perfetta, 1 indica una relazione positiva perfetta e 0 indica l'assenza di una relazione lineare tra le variabili. La correlazione è spesso utilizzata per identificare la presenza di collinearità tra le variabili.

Nel nostro caso i dati hanno una correlazione non troppo alta; ciò può essere notato dalla seguente matrice di correlazione:

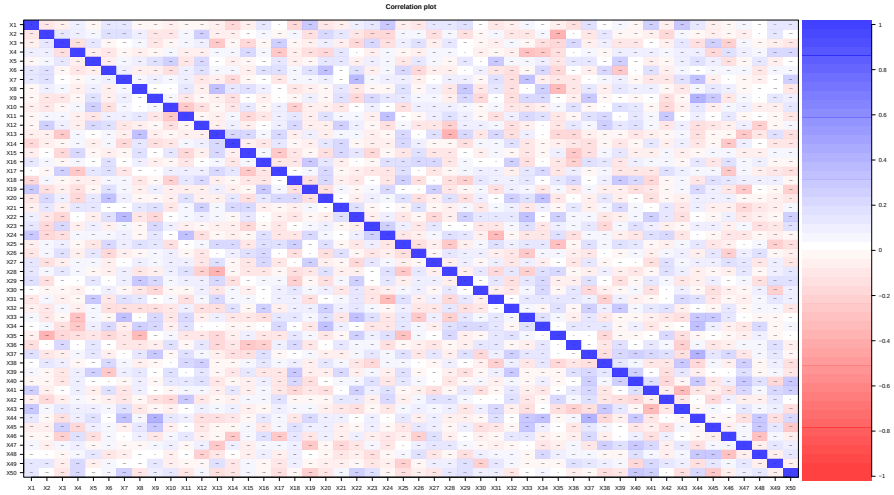

Figure 1: Matrice di Correlazione

1.1.3 Tecniche analizzate

Linear Model Applicare un modello di regressione lineare a dati ad alta dimensionalità e poca correlazione può essere problematico in quanto il modello potrebbe avere difficoltà a catturare le relazioni significative tra le variabili. In queste situazioni, è probabile che molti coefficienti siano molto piccoli o nulli, rendendo il modello poco interpretabile e meno preciso.

Best Subset Selection L'approccio BSS risulta non applicabile in contesti dove i dati sono ad alta dimensionalità; in particolare, l'algoritmo di BSS non è computazionalmente efficiente con insiemi di dati che hanno una p maggiore di 30 o 40 circa. Il nostro dataset ha 50 regressori (p=50), quindi l'analisi verrebbe effettuata su 2^{50} modelli, ciò porta a scartare il metodo.

Stepwise approach In generale, un enorme spazio di ricerca può portare a un overfitting e a un'elevata varianza delle stime dei coefficienti, per questo al BSS vengono preferiti approcci stepwise che esplorano un insieme di modelli molto più ristretto. Per questo motivo, e per ovviare ai problemi di efficienza del BSS, valutiamo alcuni approcci stepwise: possiamo andare ad utilizzare il forward selection o il backward selection, oppure un ibrido tra i due. I due metodi nel nostro caso andranno a ricercare il migliore tra un numero di modelli finito che sarà: $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$, essendo p=50, possiamo dire che i due approcci ricercheranno tra 1276 modelli. Non è garantito però che entrambi i metodi producano il miglior modello contenente un sottoinsieme di predittori p.

Ridge La regressione di Ridge è un modello di regressione lineare che introduce una penalità ℓ_2 sui coefficienti per prevenire l'overfitting. Tuttavia, se i dati hanno una alta dimensionalità con poca correlazione tra le variabili,

la penalità ℓ_2 può non essere efficace nel ridurre la complessità del modello. In questo caso, il modello potrebbe avere una penalizzazione troppo bassa e perdere informazioni importanti dai dati. Quando la penalizzazione risulta essere prossima allo 0, avremo che il modello Ridge andrà a comportarsi come il modello lineare. In queste situazioni, potrebbero essere preferibili altri modelli di regressione come la Lasso o Stepwise selection.

LASSO La regressione Lasso è un altro modello di regressione lineare che introduce una penalità ℓ_1 sui coefficienti invece che una penalità ℓ_2 come nella Ridge Regression. La penalità ℓ_1 ha la proprietà di produrre coefficienti di regolarizzazione nulli per alcune variabili, effettuando quindi *variable selection*. Questo può essere utile quando si ha una alta dimensionalità con molte variabili che non sono correlate o che hanno un effetto debole sulla variabile dipendente. In questo caso, la penalità ℓ_1 può essere più adatta a selezionare solo le variabili più rilevanti, riducendo la complessità del modello e migliorandone la capacità di generalizzazione.

Elastic NET È un metodo di regressione regolarizzato che combina linearmente le penalità ℓ_1 e ℓ_2 dei metodi Lasso e Ridge, quindi fondamentalmente mette insieme caratteristiche di Lasso e di Ridge. Con i dati a nostra disposizione, possiamo ipotizzare che il modello con MSE più basso andrà a tendere verso Lasso, mentre Ridge, come abbiamo già sottolineato, potrebbe avere un termine di penalizzazione ℓ_2 troppo basso per via della composizione dei dati.

1.2 Linear Model

Una volta calcolato il Linear Model, possiamo verificare dal summary che vi sono 17 predittori signficativi, con un p-value prossimo allo 0 (quindi <1%/5%). L'MSE calcolato in base alla predizione sul test-set è circa 9 400. Un risultato accettabile in quanto i valori delle Y sono molto grandi e quindi un errore del genere consente comunque di ottenere dei giusti valori sui coefficienti dei regressori, in modo da arrivare alla frase corretta, che è stata calcolata anche con altri metodi più efficienti.

1.3 Stepwise approach

La selezione stepwise è un approccio per la selezione automatica di variabili in un modello statistico. In questo approccio, le variabili vengono aggiunte o eliminate dal modello in base a criteri statistici.

1.3.1 Backward selection

I risultati ottenuti in questo caso sono i più promettenti; la scelta dei regressori avviene attraverso il calcolo del minimo MSE. Oltre ad avere il minimo MSE, il modello scelto dalla backward selection riesce a prendere come migliori predittori i 17 che risultano poi essere i regressori che vanno a formare la frase.

1.3.2 Forward selection

Nel caso forward i risultati combaciano con quanto definito con backward: l'MSE risulta basso e i regressori scelti sono quelli d'interesse.

1.3.3 Hybrid selection

Questo caso porta agli stessi risultati di backward e forward, infatti trova il miglior MSE e determina i 17 regressori di interesse.

1.4 Metodi di Shrinkage

In questo caso si vanno a determinare modelli più robusti che saranno caratterizzati da un termine di penalizzazione (shrinkage), il quale assegnarà delle penalità ℓ_1, ℓ_2 o una combinazione di essi.

1.4.1 Ridge

Ricordiamo che il caso Ridge andrà ad utilizzare la penalizzazione ℓ_2. I risultati in questo caso sono peggiori rispetto alle altre tecniche utilizzate, avendo un MSE maggiore. Ciò è dovuto al fatto che il valore ottimo di lambda, scelto mediante la tecnica di cross validation, è molto basso, quindi il termine di penalizzazione raggiunge valori prossimi allo 0, e quindi non viene effettuata una penalizzazione sufficiente. In questo caso, i regressori che avranno valori significativi saranno i 17 di nostro interesse, mentre i rimanenti saranno valori molto bassi, diversi da zero.

1.4.2 Lasso

Lasso, a differenza di Ridge, va ad effettuare variable selection, andando a porre a 0 i valori dei coefficienti dei regressori che non sono significativi. I risultati sono stati abbastanza superiori a Ridge: si hanno una decina di regressori che hanno valori nulli mentre tutti i non significativi hanno comunque valori prossimi allo 0, i valori significativi, invece, hanno valori molto alti. Portando quindi a un MSE molto più basso.

1.4.3 Elastic Net

Con Elastic Net si effettua una combinazioe di Ridge e Lasso, attraverso l'utilizzo di un valore α che pesa la percentuale di utilizzo dei due modelli. Con una $\alpha = 0$ avremo un modello completamente tendente a Ridge, mentre $\alpha = 1$ rappresenta Lasso. Effettuando varie prove, abbiamo notato che il modello migliore utlizzando Elastic Net ha una α molto tendente a 1, quindi a Lasso. L'MSE risulta poco più alto di Lasso di una quantità irrisoria: ciò ci porta a pensare che l'utilizzo del solo Lasso possa essere migliore dell'utilizzo di una combinazione tra Ridge e Lasso.

1.5 Conclusioni

L'MSE migliore è stato trovato tramite approcci stepwise: in particolare con l'utilizzo di qualsiasi selection method. Questi ultimi trovano perfettamente i 17 regressori che formano la frase e il loro MSE è abbastanza minore dell'MSE risultante dagli altri approcci visti.

2 Pt. 2 - Python

L'esercizio prevede la realizzazione del seguente sistema:

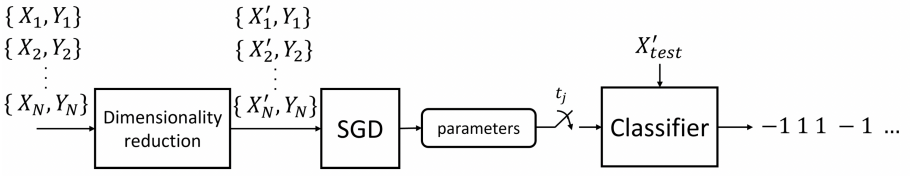

Figure 2

2.1 Analisi preliminare

2.1.1 Dataset

Il dataset è composto da un training set e da un test set. Il training set è composto da 24 000 righe e 21 colonne, mentre il test set è composto da 80 righe e 21 colonne. L'intero dataset viene standardizzato (a media 0 e varianza 1) per far sì che la PCA possa essere applicata correttamente.

2.1.2 Correlazione

Per valutare la correlazione tra i dati di training, viene fornita la relativa matrice:

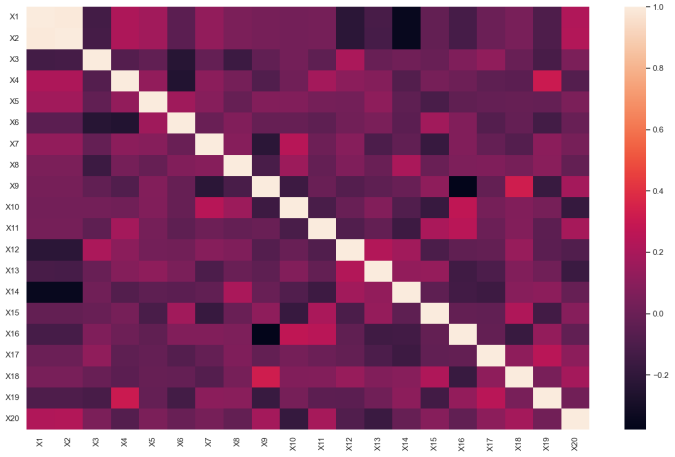

Figure 3: Matrice di correlazione sui dati scalati

Dall'immagine notiamo un'alta correlazione tra *X1* e *X2*, mentre una correlazione relativamente bassa caratterizza gli altri regressori.

2.2 Riduzione della dimensionalità - PCA

Valutata la correlazione tra i dati, andiamo ad effettuare una PCA senza ridurre la dimensionalità: in questo modo avremo dei dati non correlati tra loro e quindi una rappresentazione differente. A titolo di prova, ecco la matrice di correlazione calcolata dopo la PCA, mantenendo tutte le componenti:

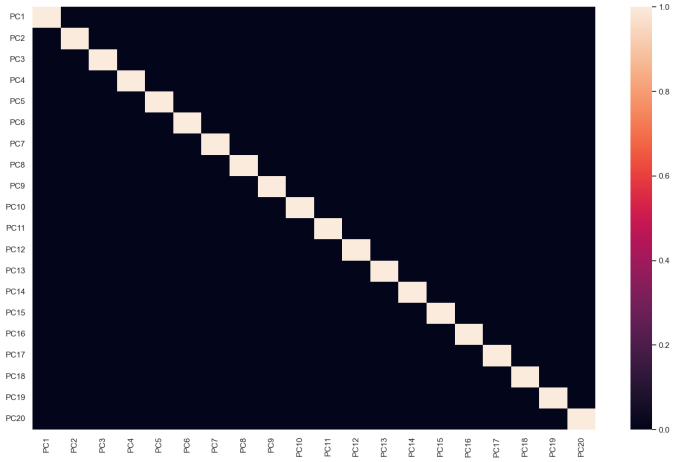

Figure 4: Matrice di correlazione successiva alla PCA

A questo punto, possiamo valutare il seguente grafico a barre (figura 5) ed andare a scegliere il numero di componenti che possono assicurarci di mantenere almeno il $90-95\%$ della varianza cumulativa ed evitare di perdere informazioni importanti.
Per mantenere una percentuale adeguata, si è deciso di prendere 16 componenti principali in modo da conservare una percentuale di poco inferiore al 95% della varianza cumulativa: in questo modo assicuriamo una rappresentazione dei dati affidabile. Possiamo inoltre notare che la componente principale 20 risulta dare un apporto significativamente basso, il che può essere spiegato andando a valutare l'unica correlazione molto forte presente nella nostra matrice di correlazione iniziale (figura 3).

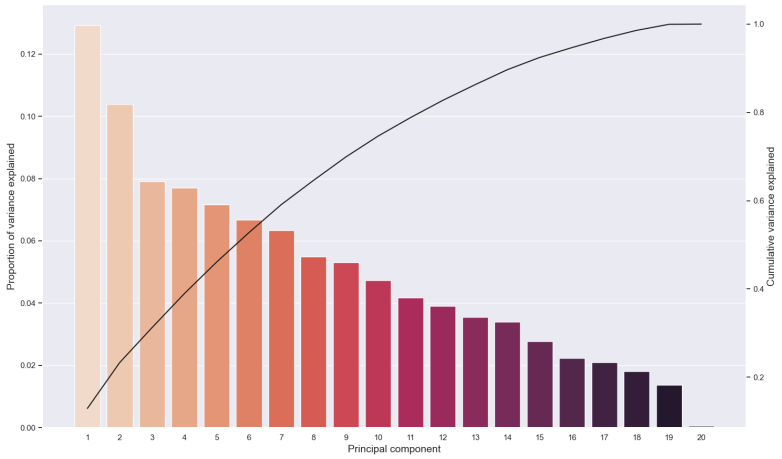

Figure 5: Varianza spiegata da ogni componente principale

2.3 Training - SDG

L'algoritmo del gradiente stocastico viene utilizzato per aggiornare i pesi del nostro modello. La scelta della funzione di loss è stata presa in quanto, secondo le specifiche, doveva essere addestrato un classificatore binario e la loss più comune in questi casi è la *logistic loss*. In aggiunta, quest'ultima si adatta meglio alle esigenze del problema, in quanto il nostro output dovrebbe contemplare i valori 1 e -1, quindi $Y \in \{-1, 1\}$, a differenza di altre loss inizialmente considerate (es. MSE), che invece considerano l'insieme $\mathbb{R}$. L'andamento trovato con l'utilizzo di questo algoritmo è stato il seguente:

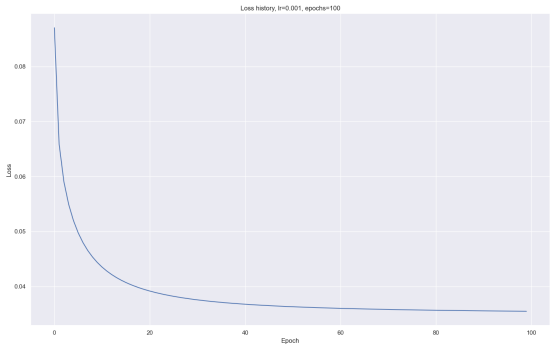

Figure 6: Valore della loss per ogni epoca

2.4 Classificatore

Il classificatore esegue la seguente operazione:

$$\begin{cases} 1 \text{ se } X'_{test}(j)\hat{\beta}(t_j) > 0 \\ -1 \text{ altrimenti} \end{cases}$$

Con $X'_{test}(j)$ si indica la feature a ridotta dimensionalità, con t_j l'instante di tempo e con $\hat{\beta}(t_j)$ il parametro stimato dall'algoritmo del gradiente stocatico al tempo t_j.

2.5 Conversione in ASCII

Infine, dato l'output del classificatore, tutti i valori -1 vengono posti a 0, e si prosegue alla conversione in ASCII.