Data Science / Data Analysis - 30 gennaio 2025

Laurea Magistrale in Ingegneria Informatica - Università degli Studi di Salerno

Si analizzi il data set `RegressionDSDA250130.csv` che contiene $n = 100$ osservazioni di una variabile dipendente Y e di $p = 25$ regressori X_j $(j = 1, 2, \ldots, p)$, tutti potenzialmente utili alla predizione di Y. A tal fine, si richiede di utilizzare l'ambiente **R**.

1. Determinare i modelli lineari che minimizzano il Mean Squared Error (MSE), individuando i regressori significativi per la predizione di Y e stimando i loro coefficienti β_j, tramite le seguenti strategie:

   i) **best subset selection (BSS)** basata sul Bayesian Information Criterion (BIC),

   ii) **backward stepwise** che utilizza **5-fold cross-validation**,

   iii) **ridge regression**,

   iv) **LASSO regression**.

2. Valutare l'MSE di test dei modelli lineari individuati al punto precedente e selezionare la strategia di regressione che permette di costruire il modello empirico lineare che minimizza, tra quelle esaminate, l'MSE di test.

Si richiede che il 70% delle osservazioni del data set per la regressione venga utilizzato per il training dei modelli e la scelta dei loro parametri, mentre il test set sia costituito dal restante 30% dei dati forniti.