ASML: quinto laboratorio

Alberi di regressione

Leonardo Egidi

Novembre 2024

Università di Trieste

L'obiettivo è familiarizzare con gli alberi di regressione, capendone il partizionamento ricorsivo binario, la scelta dei nodi e le tecniche di potatura.

In classe:

- Importare i dati sulle assicurazioni nel file `TL.csv` (già usati per Laboratorio 2) e creare una partizione in training e test set a piacere (usare sempre: $LFACE = \log(FACE)$ e $LINCOME = \log(INCOME)$).
- Stimare un primo albero di regressione usando la funzione `tree` del medesimo pacchetto. Quali sono le variabili influenti? Quanti sono i nodi terminali? Produrre un grafico dell'albero. Interpretare i risultati e le stime. Confrontare con i risultati `lm`, `ridge` e `lasso`.
- Ora potare l'albero e calcolare la larghezza ottimale del sotto-albero tramite CV. Produrre un plot per l'albero potato.
- Ora prevedere e calcolare MSE. Produrre un grafico simile a quello di Figura 8.5 del libro ISL (anche contenuta nella slide 21 della lezione sugli alberi). Commentare i risultati.

Per la consegna:

- Prendere i dati `Hitters` del pacchetto `ISLR2` e considerare la variabile risposta `log(Salary)`. Stimare un albero di regressione, potare l'albero, interpretare gli splits dei nodi, e stimare un modello lineare con funzione `lm()` e un modello LASSO con la funzione `glmnet()` e fare un raffronto con le stime ottenute con l'albero di regressione. Calcolare anche l'MSE per tutti e tre i metodi.