Regressione e regolarizzazione

Regressione e Regolarizzazione

Perchè scelta diversa rispetto ai minimi quadrati?

- Accuratezza predittiva: Quando n é poco piú grande di p, il modello lineare avrá una alta variabilitá e quindi previsioni poco accurate. Se addirittura p é maggiore di n, allora il metodo dei minimi quadrati non puó neppure essere utilizzato ed é necessario ricorrere a metodi alternativi (vincoli o shrinking).
- Interpretabilitá del modello: in molti modelli di regressione multipla puó capitare che alcune variabili inserite non sono in realtá associate alla risposta. In questo caso c'é una complessitá del modello che puó essere ridotta con la scelta delle caratteristiche piú importanti.

Possibili procedure

- Scelta di un sottoinsieme di variabili: stima su un insieme ridotto di variabili col metodo dei minimi quadrati;
- Shrinkage: Si stimano tutti i coefficienti, ma alcuni di questi sono ridotti verso lo zero. Questa riduzione ha l'effetto di ridurre la varianza.
- Riduzione della dimensionalità: Predizione dei predittori in un sottospazio M-dimensionale con M più piccolo di p. Si calcolano M combinazioni lineari (proiezioni) delle variabili.

Scelta del miglior sottoinsieme

- Si parte dal modello nullo (M_0) che non ha predittori. Praticamente per ogni osservazione il modello prevede semplicemente la media campionaria.
- Per $k = 1, 2, \ldots, p$ si stimano tutti i modelli che contengono k predittori e fra ognuno di questi si individua il migliore in termini di RSS o di R^2. Lo si indica in genere con M_k.
- Fra tutti i modelli $M_0, \ldots, M_p$ si sceglie il miglior modello usando l'errore di previsione (con cross validation), l'AIC, il BIC o l'R^2 corretto.

La stessa procedura si può utilizzare anche per altri metodi, per esempio per la logistica. In quel caso non si considera RSS ma la devianza (calcolata come l'opposto della log-verosimiglianza massimizzata).

Il problema fondamentale di questi approcci è il costo computazionale. Con dataset in cui p è elevato, i modelli da provare sono molteplici. I modelli da valutare saranno infatti pari a 2^p. Se $p = 10$, si avranno 1000 modelli, ma se $p = 20$....oltre un milione di combinazioni.

La stessa procedura si può utilizzare anche per altri metodi, per esempio per la logistica. In quel caso non si considera RSS ma la devianza (calcolata come l'opposto della log-verosimiglianza massimizzata).

Il problema fondamentale di questi approcci è il costo computazionale. Con dataset in cui p è elevato, i modelli da provare sono molteplici. I modelli da valutare saranno infatti pari a 2^p. Se $p = 10$, si avranno 1000 modelli, ma se $p = 20$....oltre un milione di combinazioni.

Altre possibilità

Procedure iterative: forward stepwise, backward stepwise, stepwise mista

Forward

- M_0 modello nullo senza predittori.
- Per $k = 0, 1, \ldots, p-1$ si considerano tutti i $p-k$ modelli con un predittore in più rispetto a M_k e si sceglie il migliore fra questi modelli indicandolo con M_{k+1}
- Si seleziona tra tutti i modelli $M_0, \ldots, M_p$ il migliore utilizzando i classici criteri.

Backward

Backward

- M_p modello con tutti i predittori.
- Per $k = p, p-1, \ldots, 1$ si considerano tutti i k modelli con un predittore in meno rispetto a M_k e si sceglie il migliore fra questi modelli indicandolo con M_{k-1}
- Si seleziona tra tutti i modelli $M_0, \ldots, M_p$ il migliore utilizzando i classici criteri.

Vantaggio di queste due procedure è che si considerano solo $1+p(p+1)/2$ modelli. Possibile anche approccio misto (ibrido).

Metodi di Shrinkage

- Regressione ridge
- Regressione LASSO

Regressione Ridge

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_1 - \sum_{j=2}^{p} \beta_j x_{ij} \right)^2$$

La regressione ridge è simile ai minimi quadrati, ma i coefficienti sono stimati minimizzando una quantità leggermente differente.

$$\sum_{i=1}^{n} \left(y_i - \beta_1 - \sum_{j=2}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=2}^{p} \beta_j^2 = RSS + \lambda \sum_{j=2}^{p} \beta_j^2$$

$\lambda \geq 0$ parametro di tuning.

Regressione Ridge

Il termine $\lambda \sum_{j=2}^{p} \beta_j^2$ è chiamato penalità shrinkage ed è piccolo quando $\beta_2, \ldots, \beta_P$ sono vicini a zero e ha l'effetto di ridurre le stime di β_j verso lo zero.λ controlla l'impatto di questi termine sulle stime. Se $\lambda = 0$ si ritorna ai minimi quadrati, invece se λ tende a infinito l'impatto del termine di shrinkage cresce.

Al variare di λ avremo diverse stime $\hat{\beta}_\lambda^R$

Regressione LASSO

$$\sum_{i=1}^{n}\left(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=2}^{p}|\beta_j| = RSS + \lambda\sum_{j=2}^{p}|\beta_j|$$

Anche in questo caso coefficienti ridotti verso lo zero, ma qualcuno portato esattamente a zero quando λ è sufficientemente grande.

Formulazione alternativa

LASSO

$$min \left[\sum_{i=1}^{n} \left(y_i - \beta_1 - \sum_{j=2}^{p} \beta_j x_{ij} \right)^2 \right]$$

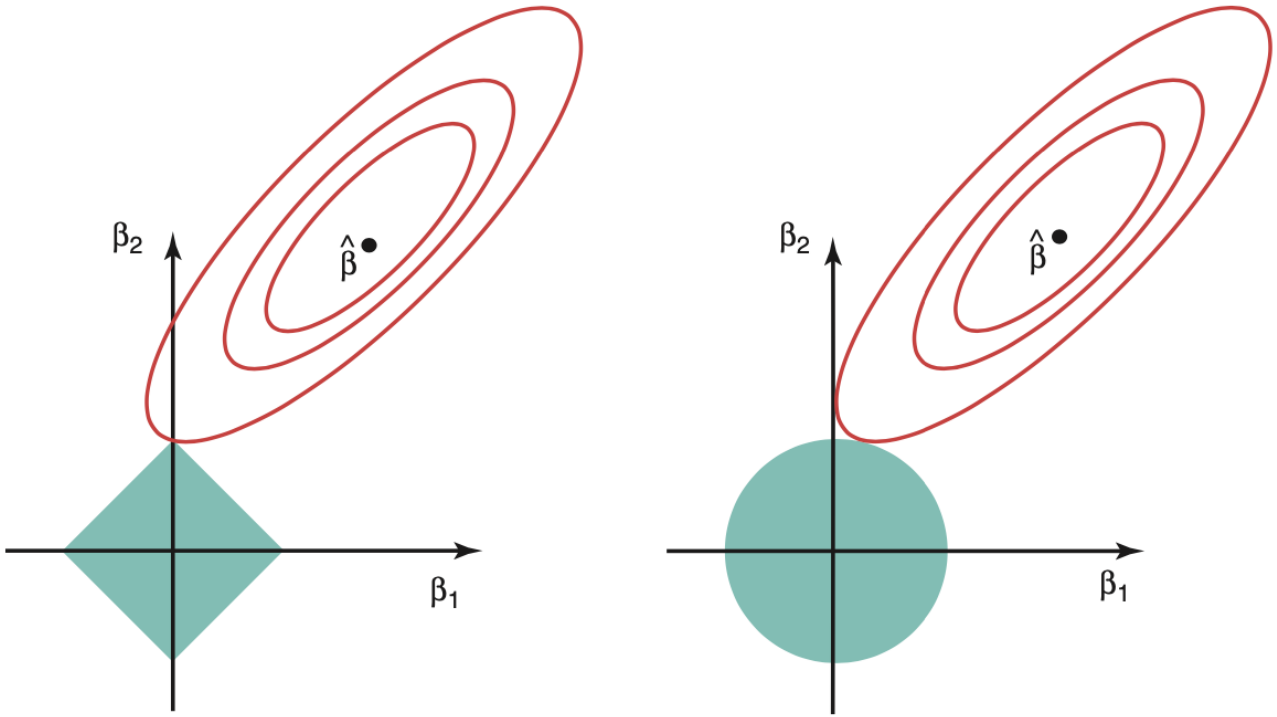
con vincolo $\sum_{j=2}^{p} |\beta_j| \leq s$

RIDGE

$$min \left[\sum_{i=1}^{n} \left(y_i - \beta_1 - \sum_{j=2}^{p} \beta_j x_{ij} \right)^2 \right]$$

con vincolo $\sum_{j=2}^{p} \beta_j^2 \leq s$

Ridge e LASSO



Ridge e LASSO