Lineær og logistisk regression: Shrinkage, variabelselektion og $p >> N$-problemstillingen

Peter N. Bakker

08-06-2023

Lineær regression

- Lineær regression er en metode til at forudsige en kontinuerlig responsvariabel baseret på en eller flere prædiktorer. Fungerer som benchmark for nærmest alle former af regressionsanalyse.
- Modellen: $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$
- Formålet er at estimere de ukendte koefficienter $\beta_0, \beta_1, \ldots, \beta_p$ baseret på observationer af Y og $X_1, X_2, \ldots, X_p$.
- Estimering af koefficienter: Mindste kvadraters metode (OLS) og maximum likelihood estimation (MLE). Hvis Gauss-Markow antagelserne er opfyldte er $\hat{\beta}$ BLUE. Hvis ϵ'erne er normalfordelte er $\hat{\beta}$ identisk med MLE.
- Simpel, fortolkelig og ikke særlig beregningstung. Er dårlig til at prædiktere ikke lineære sammenhænge.

Logistisk regression

- Logistisk regression er en metode til at modellere en binær responsvariabel baseret på en eller flere prædiktorer.
- Modellen (for 2 kategorier): $\log \frac{P(G=1|X=x)}{P(G=0|X=x)} = \beta_0 + \beta^T x$
- Formålet er at estimere de ukendte koefficienter $\beta_0, \beta_1, \ldots, \beta_p$ baseret på observationer af Y og $X_1, X_2, \ldots, X_p$. Igennem likelihood-funktionen:
  $\ell(\beta) = \sum_{i=1}^{N} (g_i(\beta_0 + \beta^T x_i) - \log(1 + \exp(\beta_0 + \beta^T x_i)))$.
- Estimering af koefficienter: Maximum likelihood estimation (MLE) og gradient descent.
- G_i antages at være uafhængige givet X_i.

Udfordringer ved høj dimension

- Når antallet af prædiktorer (p) er større end antallet af observationer (N), kan lineær og logistisk regression støde på udfordringer som overfitting og lav præcision.
- Shrinkage-metoder som ridge regression, lasso og elastic net kan hjælpe med at reducere varians og forbedre prædiktioner.
- Shrinkage-metoder indfører en strafparameter, der straffer store koefficienter og favoriserer modeller med mindre koefficienter.
- Variabelselektion er processen med at identificere de mest informative prædiktorer og ignorere de irrelevante.
- Metoderne kan fungere, selv hvis $p >> N$.

Shrinkage-metoder

- Ridge regression:
  - Formel: $\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$
  - Trækker β-koefficienterne mod 0, men ikke helt til at være 0.
  - $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$, minder meget om alm. lin. reg.
- Lasso:
  - Formel: $\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$
  - Trækker β-koefficienterne til at være 0. Er dog mindre hård ved store koefficienter end Ridge
- Elastic net:
  - Formel: $\hat{\beta}_{\text{enet}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} ((1-\alpha)\beta_j^2 + \alpha|\beta_j|) \right\}$
  - En kombination af Ridge og Lasso, hvor både λ og α er hyperparametre. Hvilket også gør den markant mere beregningstung.
- Der findes tilsvarende for logistisk regression.

Variabelselektion: Best Subset Selection

- Best Subset Selection er en metode, der undersøger alle mulige kombinationer af prædiktorer for at finde den bedste model med et subset af prædiktorer.
- Fordel:
  - Identificerer den bedst mulige model baseret på en givet kriterium.
- Ulemper:
  - Beregningstung metode, da den kræver at undersøge alle mulige kombinationer af prædiktorer.
  - Risiko for overfitting, især ved store antal prædiktorer.
- Formel for Best Subset Selection-algoritmen:
  - Find det subset af prædiktorer, der minimerer eller maksimerer det ønskede kriterium.

Variabelselektion: Forward Stepwise Selection

- Forward Stepwise Selection er en metode, der starter med en tom model og tilføjer én prædiktor ad gangen baseret på et kriterium (f.eks. laveste AIC, BIC eller højeste R^2).
- Fordel:
  - Reducerer beregningskompleksiteten sammenlignet med Best Subset Selection.
- Ulemper:
  - Kan ikke garantere at finde den bedst mulige model, da den tager beslutninger trinvist baseret på et kriterium.
- Formel for Forward Stepwise Selection-algoritmen:
  - Start med en tom model og tilføj den prædiktor, der giver den største forbedring af kriteriet.
  - Gentag processen ved at tilføje én prædiktor ad gangen, indtil et stopkriterium er opfyldt.

Variabelselektion: Backward Stepwise Selection

- Backward Stepwise Selection er en metode, der starter med en fuld model og fjerner én prædiktor ad gangen baseret på et kriterium.
- Fordel:
  - Kan reducere kompleksiteten af modellen og identificere de mest væsentlige prædiktorer.
- Ulemper:
  - Kan ikke garantere at finde den bedst mulige model.
  - Kan være følsom over for støj i data.
- Formel for Backward Stepwise Selection-algoritmen:
  - Start med en fuld model og fjern den prædiktor, der har påvirker et ønsket kriterium bedst (f.eks. ved fjernelse af en parameter hvad sænker AIC mest muligt?).
  - Gentag processen ved at fjerne én prædiktor ad gangen, indtil et stopkriterium er opfyldt.

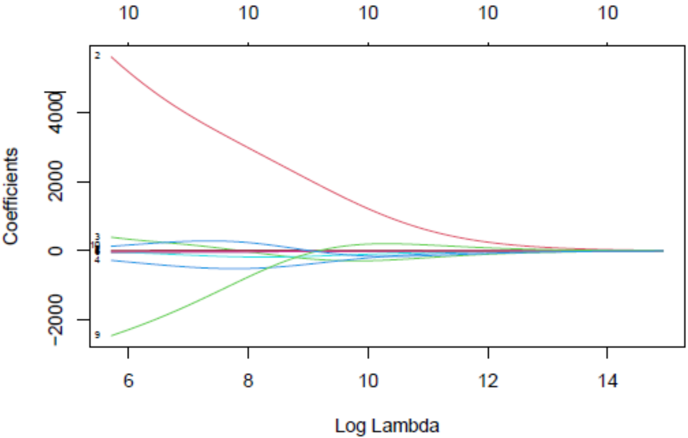
Evalueringskriterier for variabelselektion

- Ofte anvendte kriterier:
  - AIC (Akaike's Information Criterion)
  - BIC (Bayesian Information Criterion)
  - Justeret R^2
  - Krydsvalidering (f.eks. K-fold krydsvalidering)
- Valg af evalueringskriterium afhænger af det specifikke problem og mål.

Dataeksempel Huspriser

- Vi ønsker at prædiktere Kvadratmeterpris baseret på en række prædiktorer.
- Jeg sammenligner alm. lin. reg, Best subset, Ridge og Lasso.
  - RMSE_OLS: 5719.674
  - RMSE_BSS: **5483.507**
  - RMSE_Ridge: 5977.366
  - RMSE_Lasso: 5733.575
- Både Lasso og Best subset virker til at prioritere *Kvadratmeter* og postnummer 2720 i regulariseringen og variabelselktionen. Det er dog ingen garanti for at de vil prædiktere det samme, hvis de brugte det samme variable (som også ses i RMSE). Da de to metoder optimeres anderledes.

Dataeksempel Ridge



Dataeksempel Lasso