STATS 769

Model Selection and Regularisation

Yong Wang
Department of Statistics
The University of Auckland

2021

Improvements on Linear Regression

Two major approaches:

- Subset selection — Coefficients obtained by least squares
- Regularisation/shrinkage methods — Coefficients constrained or regularised

Ordinary Linear Regression

- Linear regression that includes all variables is often known as ordinary least squares (OLS).
- Including all variables in the linear regression model does not imply best prediction.
- It does imply that the RSS is the smallest. In fact, it decreases as each variable is added to the model.
- In other words, $\mathrm{MSE}(\widehat{f}; \mathbf{X}, \mathbf{y})$ decreases as each variable is added to the model.
- This does not mean that $\mathrm{MSE}(\widehat{f}; \mathbf{X}', \mathbf{y}')$ or $\mathrm{MSE}(\widehat{f}; X, Y)$ behaves in the same way.

Model Selection

- Hence we might just want to use a subset of variables that are available.
- Model selection is to find the “optimal” model out of all possible subset models.
- That are a number of model selection criteria that have been developed based on the likelihood theory.

Maximum Likelihood Estimation

- To use the likelihood method, one needs to assume a probability density (or mass) function for the relationship between X and Y. Let's denote it by $g(y; x, \beta)$.
- Then the likelihood function is simply

$$L(\beta) = \prod_{i=1}^{n} g(y_i; x_i, \beta).$$

- One is to use the maximum likelihood estimate (MLE) β, which maximises $L(\beta)$. This is a maximisation problem.
- Often, the log-likelihood function is used:

$$\ell(\ell) = \log[L(\beta)] = \sum_{i=1}^{n} \log[g(y_i; x_i, \beta)]$$

  It provides many benefits in computation and analysis.
- Note that maximising $L(\beta)$ is equivalent to maximising $\ell(\beta)$, since log is a strictly increasing function.

Normality Assumption for Linear Regression

- Typically, a normal distribution assumption is made for linear regression as follows:

  $$g(y; x, \beta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y-f(x)]^2}{2\sigma^2}},$$

  where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$,

  $$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j,$$

  and σ^2 is the variance of the noise term ϵ.
- It can be shown that the MLE $\widehat{\beta}$ is exactly the same as the LSE $\widehat{\beta}$.
- Note that the maximum likelihood approach also provides an estimate $\widehat{\sigma}$ of σ.

Model Selection Criteria

- Model selection as a penalised likelihood approach:

$$\mathrm{Criterion}(k) = -2\ell(\widehat{\beta}) + ck,$$

  where k is the number of free parameters (coefficients) and c is a term used by a criterion to penalise the (log-)likelihood function.

- Note that $\widehat{\beta}$ here denotes the MLE for the model with k parameters.
- One is to minimise $\mathrm{Criterion}(k)$ over all possible k-values or subset models.
- Two popular model selection criteria:
  - Akaike Information Criterion (AIC):

$$\mathrm{AIC}(k) = -2\ell(\widehat{\beta}) + 2k.$$

  - Bayesian Information Criterion (BIC):

$$\mathrm{BIC}(k) = -2\ell(\widehat{\beta}) + log(n)k,$$

    where n is the number of observations.

Using RSS

- For linear regression, AIC and BIC can also be written in terms of RSS.
- The MSE $\widehat{\sigma}^2$ of σ^2 is

$$\widehat{\sigma}^2 = \mathrm{RSS}/n$$

- Thus the fitted log-likelihood can be written

$$\ell(\widehat{\beta}) = -\frac{n}{2}\log(\mathrm{RSS}) + \mathit{Constant}.$$

- Hence

$$\begin{aligned}\mathrm{AIC}(k) &= n\log(\mathrm{RSS}) + 2k, \\ \mathrm{BIC}(k) &= n\log(\mathrm{RSS}) + \log(n)k,\end{aligned}$$

where the *Constant* term is ignored.

Selection Strategies

- Best subset selection
  - Consider all subset models.
  - The total number of all subset models is

$$\binom{p}{0} + \binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{0} = 2^p.$$

  - Computationally infeasible for p large.
- Forward stepwise selection
  - Start with the null model (with no predictor)
  - Add one predictor at a time: the most additional improvement
  - Computationally feasible
  - Predictors which may later become insignificant remain in the model

Selection Strategies II

- Backward stepwise selection
  - Start with the full model (with all predictors)
  - Remove one predictor at a time: the least significant
  - Good to remove insignificant predictors in order
  - Computationally feasible if $p \leq n$
  - Not feasible if $p > n$.
- Hybrid stepwise selection
  - Start with the null model
  - Adds one predictor at a time: the least significant
  - Possibly removes one predictor at a time
  - Computationally feasible

Regularisation Methods

- Alternatively, we can fit a model with all predictors, with their coefficients being constrained or regularised.
- This is equivalent to shrinking the coefficients towards 0.
- Two best-known methods:
    - Ridge Regression
    - Lasso (least absolute shrinkage selection method)

Ridge Regression

- Instead of minimising RSS, ridge regression minimises

$$\begin{aligned}\mathrm{RSS}^{\mathrm{ridge}}(\beta;\lambda) &= \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2,\end{aligned}$$

  for each fixed value of the tuning parameter $\lambda \geq 0$.
- The value of a tuning parameter needs to be determined by some other method, e.g., cross-validation (to be studied later).
- We may also write $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$, where

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$$

  is the length of vector β in a Euclidean space (and is known as the ℓ_2 norm of β).

Ridge Penalty

- The second term $\lambda\|\beta\|_2^2$ penalises RSS.
- For a fixed value of λ, the larger the value of $\|\beta\|_2$, the higher the penalty.
- It thus tends to choose a model with a smaller $\|\beta\|_2$, i.e., a vector β closer to the origin, hence shrinking the OLS estimate $\widehat{\beta}$ towards 0.
- The value of λ controls the severity of the shrinkage.
    - When $\lambda = 0$, there is no shrinkage and the estimate $\widehat{\beta}_\lambda^{\text{ridge}}$ is just the OLS one.
    - When $\lambda = \infty$, it shrinks β completely and the estimate is just the origin, i.e., $\widehat{\beta}_\lambda^{\text{ridge}} = 0$
    - As λ increases, $\|\widehat{\beta}_\lambda^{\text{ridge}}\|_2$ always decreases.
    - $\|\widehat{\beta}_\lambda^{\text{ridge}}\|_2 / \|\widehat{\beta}\|_2$ ranges from 1 (when $\lambda = 0$) to 0 (when $\lambda = \infty$)

Ridge Regression: Discussion

- Better standardise all predictors first (to have mean 0 and unit variance).
- Can improve upon the least squares estimate in terms of prediction, especially when variables are highly correlated.
- No need to consider combinations of the variables as in subset selection.
- Computationally efficient: Can find solutions for all λ-values at the same computation cost as the OLS.
- Model does not get “simpler”: All predictors are still needed in the final, fitted model (no coefficient is 0 for $\lambda > 0$).
- Not feasible when p large.

Lasso

- Instead of minimising RSS, Lasso minimises

$$\mathrm{RSS}^{\mathrm{lasso}}(\beta; \lambda) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

  for each fixed value of the tuning parameter $\lambda \geq 0$.
- We can also write

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

  (which is known as the ℓ_1 norm of vector β).
- It differs from ridge regression by replacing $\|\beta\|_2^2$ with $\|\beta\|_1$.
- $\|\beta\|_1$ is also a “length” of vector β, but measured in terms of the sum of absolute coordinate values.

Lasso Penalty

- We can pretty much repeat our description for ridge penalty, with $\|\beta\|_2^2$ being replaced with $\|\beta\|_1$.
- In particular, $\|\widehat{\beta}_\lambda^{\text{lasso}}\|_1 / \|\widehat{\beta}\|$ ranges from 1 (when $\lambda = 0$) to 0 (when $\lambda = \infty$).

Lasso: Discussion

The most important advantage of Lasso over ridge regression is that, for $\lambda > 0$, some coefficients can become exactly 0. Hence these variables are “eliminated” from the model.

- Better standardise all predictors first.
- Can improve upon the least squares estimate in terms of prediction, especially when variables are highly correlated.
- No need to consider combinations of the variables as in subset selection.
- Computationally efficient: Can find solutions for all λ-values at the same computation cost as the OLS.
- Model does get “simpler”: Some predictors are eliminated for $\lambda > 0$.
- Can be used for p large, even if $p > n$.

Some Comments

- Linear regression is the most fundamental, classical topic in statistical modelling.
- As classical as it is, there still remain many interesting questions to be answered and many potential usages to be discovered.
- It has to be more true to many newly-invented models and methods in data mining/machine learning/statistical learning.