Linear model for regression

Peida Wu

IMS

2025 年 6 月 3 日

Review

In this term, we introduce classical linear regression model.

1. Least square
2. Subset Selection
    1. Best subset selection
    2. Forward subset selection
    3. Backward subset selection
3. Shinkage method
    1. Lasso regression
    2. Ridge regression
    3. Least angle regression
4. Derived input directions
    1. Principal component regression
    2. Partial least square

Input vector: $X^T = (x_1, \ldots, x_p)$.
The linear regrssion model has the form:

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

For a set of training data:$(x_1, y_1), \ldots, (x_N, y_N)$ to estimate β, where $x_i = (x_{i1}, \ldots, x_{ip})^T$

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{1}$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{2}$$

In vector notation, we have

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

Gauss Markov Theorem: the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates.

Subset selection

Benifits: Better prediction accuracy and better interpretation.

1. With subset selection we retain only a subset of the variables, and eliminate the rest from the model.
2. Least squares regression is used to estimate the coefficients of the inputs that are retained.

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \dots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

图 1: Best subset selection

1. simple and conceptually appealing
2. computational limitations

Subset selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \dots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

图 2: forward stepwise subset selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \dots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

图 3: backward stepwise subset selection

Ridge regression:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Equivalent version:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\},$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Solution: $\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$

Lasso regression:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Equivalent version:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t.$$

Lasso vs Ridge

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

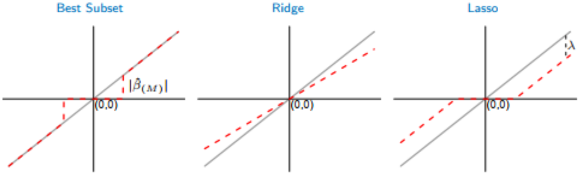

图 4: Enter Caption

Least angle regression

A kind of "democratic" version of forward stepwise regression.

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

图 5: Least angle regression

Lasso and Least angle regression

Algorithm 3.2a *Least Angle Regression: Lasso Modification.*

4a. If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

图 6: Lasso modification

Least angle regression:

$$x_j^T(y - X\beta) = r \cdot s_j$$

Lasso regression:

$$R(\beta) = \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$$

$$x_j^T(y - X\beta) = \lambda \cdot \text{sign}(\beta_j)$$

Principal component regression

Input matrix $X \in R^{N \times p}$

$$X = UDV^T$$

$$X^T X = VD^2 V^T$$

The first principal component has the property $z_1 = Xv_1$

$$Var(z_1) = Var(Xv_1) = \frac{d_1^2}{N}$$

$z_1 = Xv_1 = u_1 d_1$, u_1 is normalized first principal component.

Partial least square

Unlike Principal component regression, PLS also consider input's relationship with output.
Consider m predictors $X_1, \ldots, X_m$, p response $Y_1, \ldots, Y_p$
The first principal component T_1, U_1 is linear combination of $X = (X_1, \ldots, X_m), Y = (Y_1, \ldots, Y_p)$.

$$T_1 = w_{11}X_1 + \ldots, w_{1m}X_m = w_1'X$$

$$U_1 = v_{11}Y_1 + \ldots + v_{1p}Y_p = v_1'Y$$

The score of T_1, U_1 denote as t_1, u_1, where

$$t_1 = X_0 w_1 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1m} \end{bmatrix} = \begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix}$$

$$u_1 = Y_0 v_1 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1p} \end{bmatrix} = \begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{bmatrix}$$

$$\max < t_1, u_1 >=< X_0 w_1, Y_0 v_1 >= w_1^T X_0^T Y_0 v_1$$

subject to$||w_1||^2 = ||v_1||^2 = 1$

By lagrange:

$$L = w_1^T X_0^T Y_0 v_1 - \frac{\lambda}{2}(||w_1||^2 - 1) - \frac{\theta}{2}(||v_1||^2 - 1)$$

$$\begin{cases} \frac{\partial L}{\partial w_1} = X_0^T Y_0 v_1 - \lambda w_1 = 0 \\ \frac{\partial L}{\partial v_1} = Y_0^T X_0 w_1 - \theta v_1 = 0 \end{cases} \Rightarrow \begin{cases} Y_0^T X_0 X_0^T Y_0 v_1 = \lambda^2 v_1 \\ X_0^T Y_0 Y_0^T X_0 w_1 = \lambda^2 w_1 \end{cases}$$

Then construct regression function of $X_1, \dots, X_m, Y_1, \dots, Y_p$ to T_1,

$$\begin{cases} X_0 = t_1\alpha_1' + E_1 \\ Y_0 = t_1\beta_1' + F_1 \end{cases}$$

where E_1, F_1 are residue matrix of size $n \times m, n \times p. \alpha_1' = (\alpha_{11}, \dots, \alpha_{1m}), \beta_1' = (\beta_{11}, \dots, \beta_{1p})$.

By least square, $\begin{cases} \alpha_1 = X_0^T t_1/||t_1||^2 \\ \beta_1 = Y_0^T t_1/||t_1||^2 \end{cases}$

Repeat r times, we get

$$X_0 = t_1\alpha_1' + \cdots + t_r\alpha_r' + E_r,$$
$$Y_0 = t_1\beta_1' + \cdots + t_r\beta_r' + F_r.$$

Partial least square

$$\max_{\alpha} \mathrm{Corr}^2(y, X\alpha)\,\mathrm{Var}(X\alpha)$$

$$\text{subject to} \|\alpha\| = 1, \alpha^T S\hat{\varphi}_\ell = 0, \ell = 1, \ldots, m-1.$$