Introduction to Statistical Learning

Omid Safarzadeh

February 2, 2022

Table of contents

Logistic regression

- $Y \in \{0, 1\}$. Ex: $0 = ebola$, $1 =$ no ebola
- $X \in \mathscr{R}$

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \text{ (logistic function)}$$

- $\lim_{X\to-\infty} \pi(X)$? $\lim_{X\to+\infty} \pi(X)$?
- $\pi(X)$ models $Pr(Y = 1|X)$
- Odds:

$$\frac{\pi(X)}{1 - \pi(X)} = e^{\beta_0+\beta_1 X}$$

- Log-odds (logit):

$$logit(\pi(X)) = log(\frac{\pi(X)}{1 - \pi(X)}) = \beta_0 + \beta_1 X$$

- logit is linear in X!

MLE for simple logistic regression

- Data: $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^n$
- Model: $Y_1, ..., Y_n$ are independent. $Y_i \sim$Bernoulli$(\pi(x_i))$

Likelihood

$$L(\beta_0, \beta_1) = p(\mathscr{D}|\beta_1, \beta_0) = \prod_{i:y_i=1} \pi(x_i) \prod_{i':y_{i'}=0} (1 - \pi(x_{i'}))$$

Log-likelihood

$$l(\beta_0, \beta_1) = \log p(\mathscr{D}|\beta_1, \beta_0) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - log(1 + e^{\beta_0 + \beta_1 x_i})]$$

MLE

$$(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}) = \arg_{\beta_0, \beta_1} \max L(\beta_0, \beta_1) = \arg_{\beta_0, \beta_1} \max l(\beta_0, \beta_1)$$

MLE for simple logistic regression

- No closed form solution for $(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE})$
- MLE can be found by **Newton-Raphson method**

Multiple logistic regression

- Response: $Y \in \{0, 1\}$
- Predictors: $\mathbf{X} = [1, X_1, ..., X_p]^T$
- Parameters: $\beta = [\beta_0, ..., \beta_p]^T$
- Logistic function:

$$\pi(\mathbf{X}; \beta) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} = \frac{e^{\beta^t \mathbf{X}}}{1 + e^{\beta^T \mathbf{X}}}$$

- $\pi(\mathbf{X}; \beta)$ models $Pr(Y = 1|X_1, ..., X_p; \beta)$
- Odds:

$$\frac{\pi(\mathbf{X}; \beta)}{1 - \pi(\mathbf{X}; \beta)} = e^{\beta^T \mathbf{X}}$$

- Log-odds (logit):

$$logit(\pi(\mathbf{X}; \beta)) = log(\frac{\pi(\mathbf{X}; \beta)}{1 - \pi(\mathbf{X}; \beta)}) = \beta^T \mathbf{X}$$

- logit is linear in X!

MLE for multiple logistic regression

- Data: $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \mathbf{x}_i = [1, x_{i1}, ..., x_{ip}]^T$
- Model: $Y_1, ..., Y_n$ are independent.

$$Y_i \sim \text{Bernoulli}(\pi(\mathbf{X}_i))$$

- Log-likelihood

$$l(\beta) = \log p(\mathscr{D}|\beta) = \sum_{i=1}^{n} [y_i \beta^T \mathbf{x}_i - log(1 + e^{\beta^T \mathbf{x}_i})]$$

- MLE

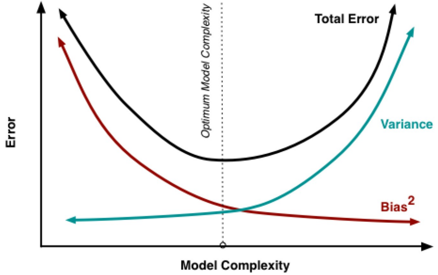
$$\hat{\beta}^{MLE} = \underset{\beta \in \mathscr{R}^{p+1}}{\arg} \max l(\beta)$$

Regularization

Properties of the least squares estimate:

- When relation between Y and $X = [X_1, ..., X_p]^T$ is almost linear, least squares estimate have low bias
- But it can have high variance. Ex: when $p \approx n$ or $p \geq n$
- Shrinking regression coefficients results in better fit

Reducing the complexity of linear regression



Two method for regularization

Ordinary leas squares:

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

Ridge regression:

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

$$= RSS(\beta) + \lambda \sum_{j=1}^{p}\beta_j^2$$

Lasso:

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

$$= RSS(\beta) + \lambda \sum_{j=1}^{p}|\beta_j|$$

Ridge regression

$$Loss_R(\beta,\lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \underbrace{\lambda}_{\text{tuning parameter}} \underbrace{\sum_{j=1}^{p}\beta_j^2}_{penalty}$$

$$\hat{\beta}^R = \arg_\beta \min Loss_R(\beta,\lambda)$$

$$\underset{\beta}{minimize}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2\} \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \leq s$$

What happens when

- $\lambda \to 0$
- $\lambda \to \infty$

How to select λ?

Example 4.1

Credit card balance prediction:

- $Y =$ card balance
- $X =$ (income, limit, rating, student, ...)
- Lines show estimated regression coefficients $\hat{\beta}^R$ by ridge regression.

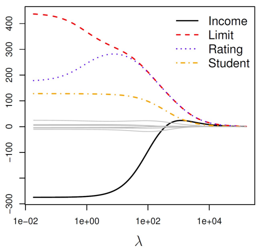


Figure: James et al., 2013

Scale invariance

- Least squares linear regression is scale invariant
- Is ridge regression scale invariant?

Making ridge regression fair:

- Standardize the predictors:

$$\underset{ij}{\tilde{X}} = \frac{X_{ij} - \bar{X}_j}{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}}$$

where $\bar{X}_j = \frac{1}{n}\Sigma_{i=1}^{n} X_{ij}$

Properties of standardized predictors:

1. $\frac{1}{n}\Sigma_{i=1}^{n} X_{ij}^{\sim} = 0$ (zero mean)
2. $\frac{1}{n}\Sigma_{i=1}^{n} X_{ij}^{\sim} 2 = 1$ (unit variance)

Bias-variance tradeoff

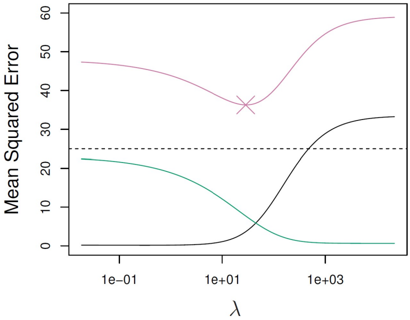


Figure: James et al., 2013

- bias: black, variance: green, MSE: red

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2$$

How to solve ridge regression?

$$Loss_R(\beta, \lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1} 6p\beta_j X_{ij})^2 + \lambda \sum_{j=1} 6p\beta_j^2$$

$$\hat{\beta}^R = \arg_\beta \min Loss_R(\beta, \lambda)$$

- Center the predictors and the response (centering makes the intercept $\hat{\beta}_0^R$)
- Standardize the predictors

How to solve ridge regression?

Some notation: y and $\mathbf{X}$ centered

$$\mathbf{y}_{n*1} \quad \beta_{p*1} \quad \mathbf{X}_{n*p}$$

Linear algebra and matrix calculus gives:

$$\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Hence given a new (centered and scaled) input $\mathbf{x}$, (centered prediction) $\hat{y} = \mathbf{x}^T\hat{\beta}^R$
Compare with least squares solution:

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

Pros and cons of ridge regression

Pros:

- Reduces variance
- $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$, $\lambda > 0$ is invertable even when $\mathbf{X}^T\mathbf{X}$ is not invertable.

Cons:

- Coefficients will be small but still almost all of them will be nonzero

Lasso

$$Loss_L(\beta,\lambda) = RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\hat{\beta}^L = \arg_\beta \min Loss_L(\beta,\lambda)$$

$$\underset{\beta}{minimize}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

- Bad news: no closed form solution like ridge regression
- Good news: no derivation

What happens when

- $\lambda \to 0$
- $\lambda \to \infty$

Example 4.2

Credit card balance prediction:

- $Y=$ card balance
- $X=$ (income, limit, rating, student, ...)
- Lines show estimated regression coefficients $\hat{\beta}^L$ by lasso.
- Lasso performs variable selection (results in a sparse model)

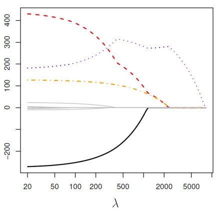


Figure: James et al., 2013

Geometric interpretation

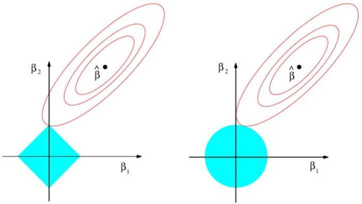


Figure: James et al., 2013

- Red lines: error contours for RSS (same error for all β values on the same contour)
- $\hat{\beta}$: least square solution
- Blue areas: region for which $|\beta_1| + |\beta_2| \leq S$ or $\beta_1^2 + \beta_2^2 \leq S$

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer New York. https://books.google.fr/books?id=qcI%5C_AAAAQBAJ