# Introduction to Statistical Learning

Omid Safarzadeh

February 2, 2022

### Table of contents

- Simple logistic regression
  - MLE for simple logistic regression
- Multiple logistic regression
  - MLE for General logistic regression
- Regularization
  - Two method for regularization
- A Ridge regression
  - Scale invariance
  - Bias-variance tradeoff
  - How to solve ridge regression?
  - Pros and cons of ridge regression
  - Geometric interpretation
- Seference

Omid Safarzadeh Installinston vo Sinsaled Republic February 2, 2022 2/20

### Logistic regression

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

$$\pi(X) = rac{e^{eta_0 + eta_1 X}}{1 + e^{eta_0 + eta_1 X}}$$
 (logistic function)

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

$$\frac{\pi(X)}{1-\pi(X)}=\mathrm{e}^{\beta_0+\beta_1X}$$

• Log-odds (logit):

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

logit is linear in X!

4□▶ 4□▶ 4□▶ 4□▶ □ 90○

3/20

Omid Safarzadeh Introduction to Statistical Learning February 2, 2022

# MLE for simple logistic regression

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

#### Likelihood

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

#### Log-likelihood

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

#### **MLE**

$$(\hat{\beta_0}^{\textit{MLE}}, \hat{\beta_1}^{\textit{MLE}}) = \arg_{\beta_0, \beta_1} \max L(\beta_0, \beta_1) = \arg_{\beta_0, \beta_1} \max I(\beta_0, \beta_1)$$

Omid Safarzadeh Buschisalara in Satisfiel Jeruma February 2, 2022 4/20

# 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

( D ) ( B ) ( E ) ( E ) ( C )

Omid Safarzadeh Intonnational Sanating February 2, 2022 5/2

# 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:

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

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

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

• 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!

eh anuccipation to Signification February 2, 2022 6/20

# MLE for multiple logistic regression

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

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

Log-likelihood

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

MLE

$$\hat{\beta}^{\textit{MLE}} = \mathop{\arg}_{\beta \in \mathscr{R}^{p+1}} \max I(\beta)$$

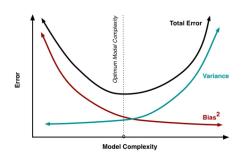
Omid Safarzadeh February 2, 2022

### 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



8 / 20

# Two method for regularization

### Ordinary leas squares:

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

### Ridge regression:

$$Loss_{R}(\beta, \lambda) = \sum_{i=1}^{n} (y_{i} - \beta_{0} - \sum_{j=1}^{p} \beta_{j} x_{ij})^{2} + \lambda \sum_{j=1}^{p} \beta_{j}^{2}$$
$$= RSS(\beta) + \lambda \sum_{i=1}^{p} \beta_{j}^{2}$$

#### Lasso:

$$Loss_{L}(\beta, \lambda) = \sum_{i=1}^{n} (y_{i} - \beta_{0} - \sum_{j=1}^{p} \beta_{j} x_{ij})^{2} + \lambda \sum_{j=1}^{p} |\beta_{j}|$$
$$= RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_{j}|$$

Omid Safarzadeh Burginston & Statistical Leanning February 2, 2022 9/2

## Ridge regression

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

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

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

#### What happens when

- $\bullet$   $\lambda \rightarrow 0$
- $\lambda \to \infty$

How to select  $\lambda$ ?

Omid Safarzadeh February 2, 2022 10 / 20

# Ridge regression

### 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.

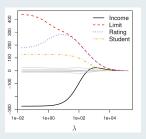


Figure: James et al., 2013

 ✓ □ > ✓ □ > ✓ □ > ✓ □ > ✓ □ > ✓ □ > ✓ □ > ✓ □ > ✓ □ 

 Omid Safarzadeh
 February 2, 2022
 11/20

### Scale invariance

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

Making ridge regression fair:

• Standardize the predictors:

$$\widetilde{X}_{ij} = \frac{X_{ij} - \overline{X}_{j}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \overline{X}_{j})^{2}}}$$

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

Properties of standardized predictors:

- ullet  $\frac{1}{n} \sum_{i=1}^{n} \chi_{ij}^{\sim} 2 = 1$  (unit variance)

Omid Safarzadeh Involvedan vo Sadistical Samina February 2, 2022

### Bias-variance tradeoff

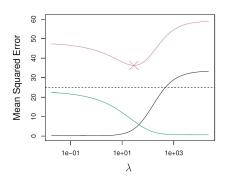


Figure: James et al., 2013

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

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

Omid Safarzadeh Insuransian o Sicilated Remitte

## How to solve ridge regression?

$$Loss_{R}(\beta, \lambda) = \sum_{i=1}^{n} (y_{i} - \beta_{0} - \sum_{j=1}^{n} 6p\beta_{j}X_{ij})^{2} + \lambda \sum_{j=1}^{n} 6p\beta_{j}^{2}$$
$$\hat{\beta}^{R} = \arg_{\beta} \min Loss_{R}(\beta, \lambda)$$

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



Omid Safarzadeh Introduction Statistical Legistra February 2, 2022 14 / 20

## How to solve ridge regression?

**Some notation:** y and X centered

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

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)  $\hat{\mathbf{y}} = \mathbf{x}^T \hat{\beta}^R$ Compare with least squares solution:

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

Omid Safarzadeh February 2, 2022

## 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

Omid Safarzadeh

$$\begin{aligned} Loss_L(\beta,\lambda) &= RSS(\beta) + \lambda \sum_{j=1}^p |\beta_j| \\ \hat{\beta}^L &= \arg_\beta \min Loss_L(\beta,\lambda) \\ &\textit{minimize}\{\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2\} \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s \end{aligned}$$

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

### What happens when

- ullet  $\lambda o 0$
- $\lambda \to \infty$



Omid Safarzadeh Invociosios oo Statistical Learning February 2, 2022

### Lasso

### 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)

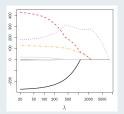


Figure: James et al., 2013

4 0 1 4 1 1 4

### Geometric interpretation

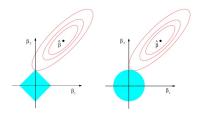


Figure: James et al., 2013

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

Omid Safarzadeh

### 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=qcl%5C\_AAAAQBAJ

Omid Safarzadeh Introducatora o Statistical Assenting February 2, 2022 20 / 20