Classification III: Sparse Logistic Regression

-Applied Multivariate Analysis-

Lecturer: Darren Homrighausen, PhD

Reminder: Generalized Linear Models (GLMs)

Logistic regression (with logit link):

Let $\pi(X_i) = Pr(Y_i = 1|X_i)$,

$$\log\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) = X_i^\top \beta$$

It is differentiable, maps [0,1] to $\mathbb{R}$, and is invertible. Its inverse is:

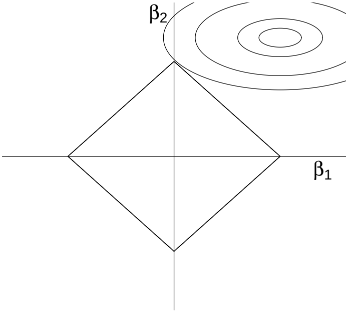
$$\pi(X_i) = \frac{\exp\{X_i^\top \beta\}}{1+\exp\{X_i^\top \beta\}}$$

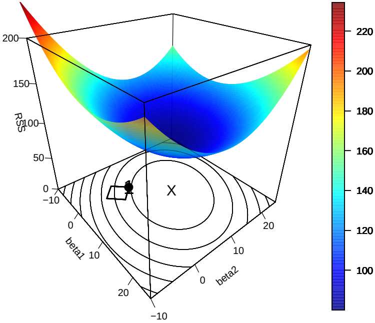
Important: We can

- ... estimate: $\hat{\beta}$ through maximum likelihood
- ... estimate: $\hat{\pi}(X) = \frac{\exp\{\hat{\beta}^\top X\}}{1+\exp\{\hat{\beta}^\top X\}}$
- ... classify: $\hat{Y} = \mathbf{1}(\hat{\pi}(X) > \text{threshold})$

  (For instance, $\text{threshold} = 0.5$)

Reminder: The lasso





This regularization set...

... is convex (computationally efficient)

... has corners (performs model selection)

Best of both worlds?

Summary: We have

- Logistic regression: Useful in the classification problem by providing an estimate of $\mathbb{P}(Y = 1|X)$
- Lasso: Useful for prediction/inference when p is large, but Y is continuous

We can combine these methods, but we need to think of lasso in terms of maximum likelihood

Maximum likelihood

Recall the Gaussian likelihood

$$L(\mu, \sigma; Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y-\mu)^2}$$

If we take the log, we get

$$\ell(\mu, \sigma; Y) = \log(L(\mu, \sigma; Y)) = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2}(Y-\mu)^2$$

If we want to do maximum likelihood over μ, we can do:

(Now, we are considering σ^2 known and hence can eliminate it from the expression)

$$\hat{\mu}_{MLE} = \arg\max_{\mu} \ell(\mu, \sigma; x) = \arg\max_{\mu} -(Y-\mu)^2 = ?$$

Maximum likelihood

Recall the Gaussian likelihood

$$L(\mu, \sigma; Y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y-\mu)^2}$$

If we take the log, we get

$$\ell(\mu, \sigma; Y) = \log(L(\mu, \sigma; Y)) = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2}(Y-\mu)^2$$

If we want to do maximum likelihood over μ, we can do:

(Now, we are considering σ^2 known and hence can eliminate it from the expression)

$$\hat{\mu}_{MLE} = \arg\max_{\mu} \ell(\mu, \sigma; x) = \arg\max_{\mu} -(Y-\mu)^2 = ?$$

(Answer: $? = Y$)

Maximum likelihood

Now, suppose we have $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$

Then we get

$$\ell(\mu, \sigma; Y_1, \ldots, Y_n) = -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \mu)^2$$

If we want to do maximum likelihood over μ, we can do:

(Same as before, use calculus to maximize)

$$\hat{\mu}_{MLE} = \arg\max_{\mu} \ell(\mu, \sigma; x) = \arg\max_{\mu} \sum_{i=1}^{n} -(Y_i - \mu)^2 = ?$$

Maximum likelihood

Now, suppose we have $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$

Then we get

$$\ell(\mu, \sigma; Y_1, \ldots, Y_n) = -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \mu)^2$$

If we want to do maximum likelihood over μ, we can do:

(Same as before, use calculus to maximize)

$$\hat{\mu}_{MLE} = \arg\max_{\mu} \ell(\mu, \sigma; x) = \arg\max_{\mu} \sum_{i=1}^{n} -(Y_i - \mu)^2 = ?$$

(Answer: $? = \overline{Y}$)

Maximum likelihood

Now, suppose we have pairs of data$(X_1, Y_1), \ldots, (X_n, Y_n)$.

Now, we state there is a parameter β such that

$$Y|X \sim N(\mu, \sigma^2) \qquad \text{and} \qquad \mu = X^\top \beta$$

The log likelihood looks like before, but now we put $\mu = X^\top \beta$

$$\begin{aligned}\ell(\beta, \sigma; Y_1, \ldots, Y_n) &= -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \mu)^2 \\ &= -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - X_i^\top \beta)^2\end{aligned}$$

If we want to do maximum likelihood over β, we can do:

$$\begin{aligned}\hat{\beta}_{MLE} &= \underset{\beta}{\arg\max} - \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 \\ &=\end{aligned}$$

Maximum likelihood

Now, suppose we have pairs of data$(X_1, Y_1), \dots, (X_n, Y_n)$.

Now, we state there is a parameter β such that

$$Y|X \sim N(\mu, \sigma^2) \qquad \text{and} \qquad \mu = X^\top \beta$$

The log likelihood looks like before, but now we put $\mu = X^\top \beta$

$$\begin{aligned}\ell(\beta, \sigma; Y_1, \dots, Y_n) &= -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - \mu)^2 \\ &= -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - X_i^\top \beta)^2\end{aligned}$$

If we want to do maximum likelihood over β, we can do:

$$\begin{aligned}\hat\beta_{MLE} &= \underset{\beta}{\arg\max} - \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 \\ &= \underset{\beta}{\operatorname{argmin}} ||Y - \mathbb{X}\beta||_2^2 = (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top Y\end{aligned}$$

Lasso

Conclusion: We can think about the lasso as a regularized maximum likelihood estimator

For regression:

$$Y = X^\top \beta + \epsilon \sim N(X^\top \beta, \sigma^2)$$

For classification:

$$Y \sim \mathrm{Bernoulli}(\pi(X)) \qquad \text{and} \qquad \pi(X) = \mathrm{logistic}(X^\top \beta)$$

(We can do other likelihood-based methods such as negative binomial, Poisson, Cox proportional hazard, ... We won't be discussing these in this class, however)

Logistic lasso

From last slide: For classification:

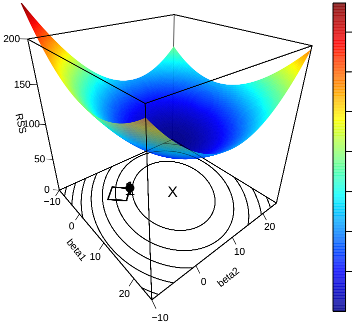
$$Y \sim \text{Bernoulli}(\pi(X)) \qquad \text{and} \qquad \pi(X) = \text{logistic}(X^\top \beta)$$

This is the same as writing the log-likelihood as

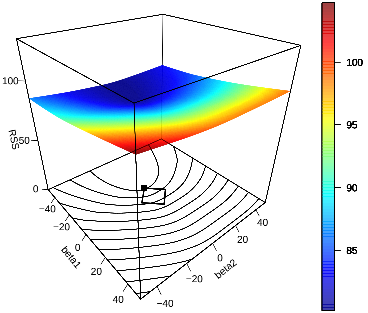
$$\ell(\beta) = \sum_{i=1}^{n} \left(Y_i \beta^\top X_i - \log(1 + e^{\beta^\top X_i}) \right)$$

Graphically it looks like...

Logistic lasso



Regression lasso



Logistic lasso

The problem is still...

... convex (computationally efficient)

... and has corners (performs model selection)

Logistic lasso

Now, as we have changed the likelihood, we want to solve

$$\min_{\beta} \sum_{i=1}^{n} - \left(Y_i \beta^\top X_i - \log(1 + e^{\beta^\top X_i}) \right) + \lambda \, ||\beta||_1$$

instead of the classic (Gaussian) lasso

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \beta^\top X_i \right)^2 + \lambda \, ||\beta||_1$$

Logistic lasso in R

Again, generous R developers have come to our rescue

We already know how to do logistic lasso

```
glmnet(x=X,y=Y,family='binomial',alpha=1)
#or
cv.glmnet(x=X,y=Y,family='binomial',alpha=1)
```

All the previous discussions apply

(Make sure the λ grid is appropriate, get predictions with predict or coefficients with coef, we can do elastic net (including logistic ridge regression) by setting alpha)

One slight difference is in getting the predictions

```
#For classifications
predict(out,X_0,s='lambda.min',type='class')
#For probabilities
predict(out,X_0,s='lambda.min',type='response')
```