Lecture 9: Other GLMs

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

Read: SLS Chs. 3.1–3.4

Outline

1. Poisson Regression

2. Multiclass Logistic Regression

3. Sparse Generalized Linear Models

The next section would be . . . . . .

1 Poisson Regression

2 Multiclass Logistic Regression

3 Sparse Generalized Linear Models

Poisson GLM for count data

- When the response variable Y is nonnegative and represents a count, its mean will be positive and the Poisson likelihood is often used for inference.
- Use the log-linear model to enforce the positivity.
- Assume that for each $X = x$, the response Y follows a Poisson distribution with mean μ satisfying

$$\log \mu = \beta_0 + \boldsymbol{\beta}^T x$$

Poisson distribution

- pmf:

$$\frac{\mu^y e^{-\mu}}{y!} = e^{y \log(\mu) - \mu - \log(y!)}$$

- As a member of exponential family:

$$\theta = \log(\mu),\ b(\theta) = e^{\theta},\ b'(\theta) = e^{\theta}$$

- If use canonical link:

$$g(\mu) = (b')^{-1}(\mu) = \log(\mu)$$

- We will use a linear function in $\boldsymbol{\beta}$ to model $\log(\mu)$

Maximum Log-likelihood

$$\max \sum_i y_i \log(\mu_i) - \mu_i + C$$

Or equivalently

$$\min \sum_i \{-y_i f(\boldsymbol{x}_i) + e^{f(\boldsymbol{x}_i)}\}$$

where $f(x) = \beta_0 + \boldsymbol{\beta}^T x$.

Recall loss function for logistic regression:

$$\{-y_i f(\boldsymbol{x}_i) + \log(1 + e^{f(\boldsymbol{x}_i)})\}$$

Note the Poisson approximation of Binomial distribution and $\log(1 + x) \approx x$

Example - Horseshoe Crabs and Satellites

A study of nesting horseshoe crabs. Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. Explanatory variables that are thought to affect this included the female crab's color (C), spine condition (S), weight (Wt), and carapace width (W). The response outcome for each female crab is her number of satellites (Sa). There are 173 females in this study.

```
> crabs
    Satellites Width Dark GoodSpine Rep1 Rep2
1            8  28.3   no        no    2    2
2            0  22.5  yes        no    4    5
3            9  26.0   no       yes    5    6
4            0  24.8  yes        no    6    6
5            4  26.0  yes        no    6    8
6            0  23.8   no        no    8    8
```

```
> fit.log.linear <- glm(Satellites˜., data = crabs, family
> summary(fit.log.linear)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.655015   0.574247  -4.623 3.77e-06 ***
Width         0.154174   0.021058   7.321 2.45e-13 ***
Darkyes      -0.240107   0.105431  -2.277  0.02276 *
GoodSpineyes -0.014044   0.098517  -0.143  0.88664
Rep1          0.018124   0.007266   2.494  0.01262 *
Rep2         -0.020216   0.007516  -2.690  0.00715 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.61  on 167  degrees of freedom
AIC: 918.91

Number of Fisher Scoring iterations: 6
```

Interpretation??

The next section would be ......

1 Poisson Regression

2 Multiclass Logistic Regression

3 Sparse Generalized Linear Models

Multinomial Regression

- Also known as Multiclass Logistic Regression. Model the posterior probabilities of K classes using x ($K > 2$).
- We may model $K - 1$ pairwise binary logistic regression after choosing one class as the reference class (say the Kth class.)

$$
\begin{gathered}
\log \frac{P(Y=1|X=x)}{P(Y=K|X=x)} = \beta_{01} + \boldsymbol{\beta}_1^T x \\
\log \frac{P(Y=2|X=x)}{P(Y=K|X=x)} = \beta_{02} + \boldsymbol{\beta}_2^T x \\
\cdots \\
\log \frac{P(Y=K-1|X=x)}{P(Y=K|X=x)} = \beta_{0,K-1} + \boldsymbol{\beta}_{K-1}^T x
\end{gathered}
$$

Posterior Probabilities

Note that

$$\sum_{k=1}^{K} P(Y = k|X = x) = 1$$

Some simple calculation leads to

$$P(Y = k|X = x) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{0\ell} + \boldsymbol{\beta}_\ell^T x)} \text{ for } k \neq K$$

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{0\ell} + \boldsymbol{\beta}_\ell^T x)}$$

Alternative Formula

Alternatively, we may model the log-probability as a linear function in x, subtract a constant.

$$\log P(Y=k|X=x) = \beta_{0k} + \boldsymbol{\beta}_k^T x - \log(Z)$$

where Z is chosen so that the probabilities sum up to 1, that is

$$P(Y=k|X=x) = \exp(\beta_{0k} + \boldsymbol{\beta}_k^T x)/Z$$

and

$$\sum_{\ell=1}^{K} [\exp(\beta_{0\ell} + \boldsymbol{\beta}_\ell^T x)]/Z = 1$$

Hence

$$P(Y=k|X=x) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^T x)}{\sum_{\ell=1}^{K}[\exp(\beta_{0\ell} + \boldsymbol{\beta}_\ell^T x)]}$$

Note that here we have K formulas versus $K-1$ formulas in the previous model. It is easy to see that they are equivalent.

- This model is over specified (not identifiable), since we can add the linear term $\gamma_0 + \boldsymbol{\gamma}^T \boldsymbol{x}$ to the linear model for each class, and the probabilities are unchanged.
- Sometimes we prefer this redundant but symmetric approach if we want to add regularization terms (later).
- In those cases, we can show that the redundancy is effectively eliminated.

The next section would be . . . . . .

1 Poisson Regression

2 Multiclass Logistic Regression

3 Sparse Generalized Linear Models

- Sparse Binary Logistic Regression
- Sparse Poisson Regression
- Sparse Multiclass Logistic Regression

Penalized log-likelihood.

- Each GLM can be written as maximum log-likelihood (or equivalently minimum negative log-likelihood)
- For sparsity as in lasso, we can impose some sparsity penalties to the coefficient vector.

$$\min_{\beta_0, \boldsymbol{\beta}} \quad \underbrace{-\frac{1}{N}\mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{X}, \boldsymbol{y})}_{\text{negative log-likelihood, similar to RSS}} \quad + \lambda \cdot P(\boldsymbol{\beta})$$

Sparse Binary Logistic Regression

- 0/1 coding:

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\}+\lambda\|\beta\|_1$$

- ±1 coding (more common in machine learning community):

$$\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{-y_i(\beta_0+\beta^T x_i)})+\lambda\|\beta\|_1$$

- ℓ_2 (ridge) penalty are possible too, but they leads to no sparsity solution.

Fitting Sparse Logistic Models.

- For regularized logistic regression, one could apply coordinate descent directly to the criterion.
- Disadvantage: optimizing values along each coordinate are not explicitly available and require a line search
- Recall that for unpenalized (traditional) logistic regression, the Newton algorithm is about iterative reweighted least square, which can be written as the following Quadratic Programing:

$$\ell_Q := \sum_{i=1}^{N} w_i (z_i - \beta_0 - \beta^T x_i)^2 \text{ where}$$

1. $\tilde{\beta}_0, \tilde{\beta}$ are current estimates
2. $\tilde{p}_i = p(x_i; \tilde{\beta}_0, \tilde{\beta}) = \frac{\exp[\tilde{\beta}_0 + \tilde{\beta}^T x_i]}{1+\exp[\tilde{\beta}_0 + \tilde{\beta}^T x_i]}$
3. $w_i = \tilde{p}_i(1 - \tilde{p}_i)$
4. $z_i = \tilde{\beta}_0 + \tilde{\beta}^T x_i + y_i w_i^{-1}$

- Therefore, we consider iteratively solve the following

$$\min_{\beta_0,\beta} \ell_Q + \lambda\|\beta\|_1$$

  which is lasso-penalized weighted least-squares problem and it can be solved (quickly) using coordinate descent. Solution to this is called proximal Newton map.

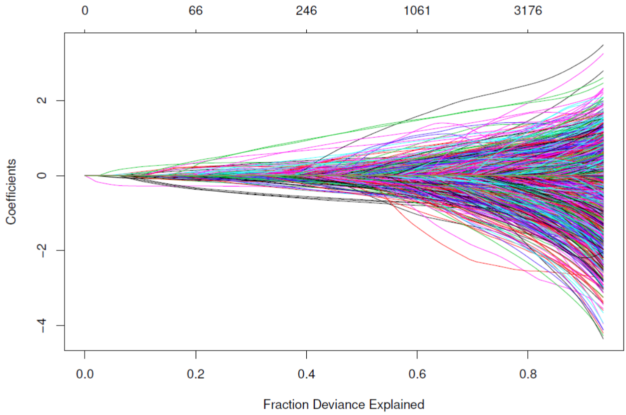
- The whole algorithm is called **generalized** Newton method.

1. Outer loop: Decrement λ (similar to the same step in coordinate descent)
2. Middle loop: Update ℓ_Q using the current estimates (similar to the loop in regular logistic reg.)
3. Inner loop: Run the coordinate descent algorithm on the penalized weighted-least-squares problem above (essentially a lasso calculation.)

Example: Document Classification

- Document classification using the 20-Newsgroups corpus (Lang 1995)
- There are N = 11,314 documents and p = 777,811 features, with 52% in the positive class. Only 0.05% of the features are nonzero for any given document.
- Two classes: positive class = 10 groups with names of the form sci.*, comp.* and misc.forsale, and the rest are the negative class.
- The feature set consists of trigrams, with message headers skipped, no stoplist, and features with less than two documents omitted.

ℓ_1 logistic regression solution path



Newsgroup: ℓ_1 logistic regression

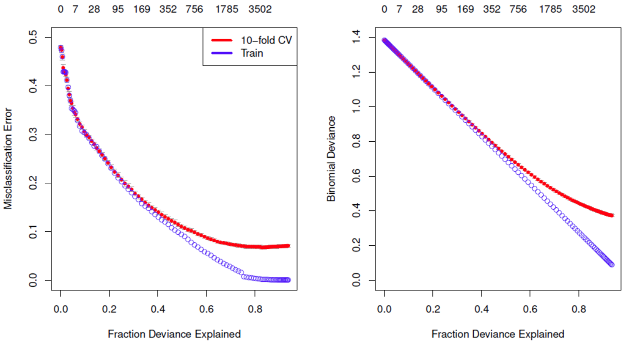


Figure 3.2 *Lasso (ℓ_1)-penalized logistic regression. Tenfold cross-validation curves for the Newsgroup data are shown in red, along with pointwise standard-error bands (not visible). The left plot shows misclassification error; the right plot shows deviance. Also shown in blue is the training error for each of these measures. The number of nonzero coefficients in each model is shown along the top of each plot.*

- Misclassification error: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i f(\boldsymbol{x}_i) < 0\}$
- Deviance loss: $\frac{1}{n}\sum_{i=1}^{n}\{-y_i f(\boldsymbol{x}_i) + \log(1 + e^{f(\boldsymbol{x}_i)})\}$

Newsgroup: ℓ_2 logistic regression

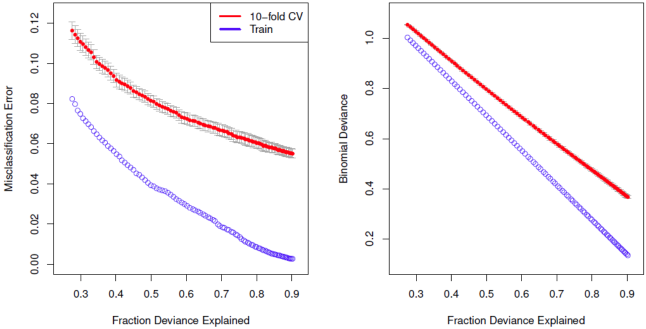


Figure 3.3 *Ridge (ℓ_2)-penalized logistic regression: tenfold cross validation curves for the Newsgroup data are shown in red, along with pointwise standard-error bands. The left plot shows misclassification error; the right plot shows deviance. Also shown in blue is the training error for each of these measures.*

- Misclassification error: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i f(\boldsymbol{x}_i) < 0\}$
- Deviance loss: $\frac{1}{n}\sum_{i=1}^{n}\{-y_i f(\boldsymbol{x}_i) + \log(1 + e^{f(\boldsymbol{x}_i)})\}$

- CV errors are about the same between ℓ_1 and ℓ_2 penalties.
- # of nonzeross in every model is p $=$ 777,811 with ℓ_2 penalty compared to a maximum of 5,277 with ℓ_1.
- Ridge logistic regression might be more costly. For ridge the 10-fold CV took 8.3 minutes, while for lasso under one minute.

Sparse Poisson Regression

The ℓ_1-penalized negative log-likelihood is given by

$$\min -\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0+\boldsymbol{\beta}^T x)+e^{f(\beta_0+\boldsymbol{\beta}^T x)}\}+\lambda\|\boldsymbol{\beta}\|_1$$

As with other GLMs, we can fit this model by **iteratively reweighted least squares**, which amounts to fitting a weighted lasso regression at each outer iteration.

Sparse Multiclass Logistic Regression

$$P(Y = k|X = x) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^T x)}{\sum_{\ell=1}^{K} [\exp(\beta_{0\ell} + \boldsymbol{\beta}_\ell^T x)]}$$

Here we prefer the redundant but symmetric approach, because

- since we regularize the coefficients, the regularized solutions are **not equivariant** under base changes, and
- the regularization automatically eliminates the redundancy (details below).

- Lasso penalized multinomial regression:

$$-\frac{1}{N}\sum_{i=1}^{N}\log\Pr(Y=y_i|X=x_i;\{\beta_{0k},\boldsymbol{\beta}_k\}_{k=1}^{K})+\lambda\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_1$$

- Denote by **R** the $N \times K$ indicator response matrix with $r_{ik}=\mathbb{1}\{y_i=k\}$, then the first term above can be written as

$$\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K}r_{ik}(\beta_{0k}+\boldsymbol{\beta}_k^T x_i)-\log\left\{\sum_{k=1}^{K}e^{\beta_{0k}+\boldsymbol{\beta}_k^T x_i}\right\}\right]$$

Elimination of the redundancy

- As mentioned earlier, $\{\beta_{jk} + c_j\}_{k=1}^K$ and $\{\beta_{jk}\}_{k=1}^K$ produce exactly the same probabilities. Hence it would have been an identifiable issue.
- However, the penalty force c_j to be a particular value (i.e. not just any arbitrary value.) For any candidate set $\{\tilde{\beta}_{jk}\}_{k=1}^K$

$$c_j = \underset{c \in \mathbb{R}}{\operatorname{argmin}} \sum_{k=1}^{K} |\tilde{\beta}_{jk} - c| = \text{the median of } \{\tilde{\beta}_{jk}\}_{k=1}^K$$

- β_{0k}'s are still undetermined. In package `glmnet`, they are constrained to sum to zero.

Grouped-Lasso Multinomial

- Lasso penalty disadvantage: although individual coefficient vectors are sparse, the overall model may not be. (WHY?)
- An alternative approach is to use a grouped-lasso penalty

$$-\frac{1}{N}\sum_{i=1}^{N}\log\Pr(Y=y_i|X=x_i;\{\beta_{0k},\boldsymbol{\beta}_k\}_{k=1}^{K})+\lambda\sum_{j=1}^{p}\|\vec{\beta}_j\|_2$$

- NOTE: not $\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_2$.
- Slight abuse of notation: **B** is $p \times K$ coefficient matrix. $\boldsymbol{\beta}_k \in \mathbb{R}^p$ all p coefficients for the kth category, kth column of **B**; $\vec{\beta}_j \in \mathbb{R}^K$ all K coefficients for the jth variable, jth row of **B**
- Block ℓ_1/ℓ_2 constraint: select all the coefficients for a particular variable to be in or out of the model simultaneously
- All the resulting coefficients for a particular variable (say variable j) satisfy $\sum_{k=1}^{K}\beta_{jk}=0$ (WHY?)

Block coordinate descent

- As before, coordinate descent techniques are one reasonable choice, in this case **block coordinate descent** on each vector β_j, holding all the others fixed.
- Next: more details on fitting binary sparse logistic regression and regular (non block version) coordinate descent applied to logistic regression.

Summary of R functions used

- Logistic regression. `glm` with binomial family
- Poisson regression (Log-linear model.) `glm` with poisson family
- Multinomial regression. `nnet::multinom`
- Sparse (ℓ_1, ℓ_2 or elastic net) Logistic Regression, Sparse (ℓ_1, ℓ_2 or elastic net) Multinomial Regression (incl. grouped lasso penalty). `glmnet` with family (binomial or multinomial), α (0, 1, or (0,1)), and type.multinomial (ungrouped,grouped)