Lec 4: Generalized Linear Model

Yanjun Han

Sept 24, 2024

Generalized linear model.

Setting. For $i = 1, 2, \cdots, n$, let $y_i \overset{ind}{\sim} p_{\theta_i}(y_i) = \exp(\langle \theta_i, T(y_i)\rangle - A(\theta_i))h(y_i)$.
where $\theta_i = (\langle x_i, \beta_1\rangle, \langle x_i, \beta_2\rangle, \cdots, \langle x_i, \beta_d\rangle) \in \mathbb{R}^d$

- $x_i \in \mathbb{R}^p$: feature / covariate
- $(\beta_1, \cdots, \beta_d) \in \mathbb{R}^{p\times d}$: regression coefficients
- written in matrix form : $\theta_i = \beta^T x_i$

MLE.

$$\begin{aligned}\hat{\beta} &= \arg\max_\beta \prod_{i=1}^n p_{\theta_i}(y_i) \\ &= \arg\max_\beta \sum_{i=1}^n \left(\langle \beta^T x_i, T(y_i)\rangle - A(\beta^T x_i)\right) \\ &= \arg\max_\beta \underbrace{\mathrm{Tr}\left(\sum_{i=1}^n T(y_i) x_i^T \cdot \beta\right)}_{\text{linear in } \beta} - \underbrace{\sum_{i=1}^n A(\beta^T x_i)}_{\text{convex in } \beta}\end{aligned}$$

Estimating equation $(d=1)$: $\sum_{i=1}^n T(y_i) x_i = \sum_{i=1}^n A'(\hat{\beta}^T x_i) x_i$.

The computation of MLE is a convex problem, thus efficient.

In R: model <- glm(y ~ X , family).

Examples. 1. Linear regression.

$$y_i \sim N(\theta_i, 1) = N(\beta^T x_i, 1)$$

$$\Rightarrow \hat{\beta} = \arg\min_\beta \sum_{i=1}^n (y_i - \beta^T x_i)^2 = \arg\min_\beta \| y - \underset{\mathbb{R}^{n\times p}}{X}\beta\|_2^2$$

2. Logistic regression.

$$y_i \sim \mathrm{Bern}\left(\frac{1}{1+e^{-\theta_i}}\right) = \mathrm{Bern}\left(\frac{1}{1+e^{-\beta^T x_i}}\right)$$

$$\begin{aligned}\Rightarrow \hat{\beta} &= \arg\max_\beta \sum_{i=1}^n \left(y_i \log\frac{1}{1+e^{-\beta^T x_i}} + (1-y_i)\log\frac{e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}}\right) \\ &= \arg\max_\beta \sum_{i=1}^n \left(y_i \beta^T x_i - \log(1+e^{\beta^T x_i})\right)\end{aligned}$$

2'. Probit model

$$Y_i \sim \mathrm{Bern}(\Phi(\theta_i)) = \mathrm{Bern}(\Phi(\beta^T x_i)),$$

where Φ is the standard normal CDF:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

MLE: $\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \left(y_i \log \Phi(\beta^T x_i) + (1-y_i) \log(1-\Phi(\beta^T x_i)) \right)$

Lemma. The above objective is concave in β.

Pf. For $f(x) = \log \Phi(x)$:

$$f'(x) = \frac{\varphi(x)}{\Phi(x)}, \quad f''(x) = \frac{\varphi' \Phi - \varphi^2}{\Phi^2} = -\frac{(x\Phi + \varphi)\varphi}{\Phi^2}.$$

Gaussian Mills ratio:

$$1 - \Phi(x) < \frac{\varphi(x)}{x}, \quad x > 0$$

$$\Rightarrow x\Phi(x) + \varphi(x) > 0,\ x < 0 \Rightarrow f''(x) < 0.$$

(See HW for an alternative proof)

> In an exponential family, there could be more than one parametrizations such that the MLE computation in the corresponding GLM is a convex problem.

3. Poisson regression.

$$Y_i \sim \mathrm{Poi}(e^{\theta_i}) = \mathrm{Poi}(e^{\beta^T x_i})$$

$$\Rightarrow \hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \left(T(y_i) \beta^T x_i - A(\beta^T x_i) \right)$$

$$= \arg\max_{\beta} \sum_{i=1}^{n} \left(y_i \beta^T x_i - e^{\beta^T x_i} \right).$$

4. Multinomial logit regression.

Recall that $\theta = (\theta_1, \cdots, \theta_k)$

$$T(y) = (1(y=1), 1(y=2), \cdots, 1(y=k))$$

$$A(\theta) = \log(e^{\theta_1} + \cdots + e^{\theta_k})$$

Model: $\mathbb{P}(y_i = j \mid x_i) = \dfrac{e^{\beta_j^T x_i}}{e^{\beta_1^T x_i} + e^{\beta_2^T x_i} + \cdots + e^{\beta_k^T x_i}}$.

MLE:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \Big(1(y_i=1)\beta_1^T x_i + 1(y_i=2)\beta_2^T x_i + \cdots + 1(y_i=k)\beta_k^T x_i - \log\big(\sum_{j=1}^{k} e^{\beta_j^T x_i}\big)\Big)$$

$$= \arg\max_{\beta} \sum_{j=1}^{k} \beta_j^T \sum_{i:\, y_i=j} x_i - n \log\Big(\sum_{j=1}^{k} e^{\beta_j^T x_i}\Big).$$

Note: the MLE is not unique, as $(\beta_1, \cdots, \beta_k)$ and $(\beta_1 + c, \cdots, \beta_k + c)$ give the same objective. So we can assume that $\beta_1 = 0$.

4' Ordered logit model (ordinal regression).

Suppose y_i could take k values with ordered relationship.

Model:

$$\log \frac{\mathbb{P}(y_i \le j)}{\mathbb{P}(y_i > j)} = \alpha_j + \beta^T x_i \quad (j = 1, 2, \cdots, k-1)$$

or equivalently,

$$\mathbb{P}(y_i \le j) = \frac{1}{1 + e^{-(\alpha_j + \beta^T x_i)}}.$$

Proportional odds assumption: the difference in the log-odds

$$\log \frac{\mathbb{P}(y_i \le j+1)}{\mathbb{P}(y_i > j+1)} - \log \frac{\mathbb{P}(y_i \le j)}{\mathbb{P}(y_i > j)}$$

is independent of x. More on this in Lecture 5.

MLE:

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{(\alpha,\beta)} \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} 1(y_i = j) \log \mathbb{P}(y_i = j) \Big)$$

$$= \arg\max_{(\alpha,\beta)} \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} 1(y_i = j) \cdot \log\Big(\frac{1}{1 + e^{-(\alpha_j + \beta^T x_i)}} - \frac{1}{1 + e^{-(\alpha_{j-1} + \beta^T x_i)}} \Big)\Big)$$

where $\alpha_0 \triangleq 0$, $\alpha_k \triangleq +\infty$.

$\boxed{\text{Exercise (HW): show that the log-likelihood is concave in } (\alpha, \beta).}$

Variance of MLE.

In the sequel we assume that $d=1$ for simplicity, i.e. $\beta \in \mathbb{R}^1$.

$$\text{F.O.C. for MLE: } 0 = \sum_{i=1}^{n} \left(T(y_i) - A'(x_i^T \hat{\beta}^{MLE}) \right) x_i$$
$$= \sum_{i=1}^{n} \left(A'(x_i^T \beta) - A'(x_i^T \hat{\beta}^{MLE}) \right) x_i$$
$$+ \underbrace{\sum_{i=1}^{n} \left(T(y_i) - A'(x_i^T \beta) \right) x_i}_{\mathrm{Cov}(\cdot) = \sum_{i=1}^{n} A''(x_i^T \beta) x_i x_i^T}$$

Delta method (Taylor expansion):

$$\text{first term} \approx \left(\sum_{i=1}^{n} A''(x_i^T \beta) x_i x_i^T \right) (\beta - \hat{\beta}^{MLE})$$

$$\boxed{\mathrm{Cov}_\beta(\hat{\beta}^{MLE}) \approx \left(\sum_{i=1}^{n} A''(x_i^T \beta) x_i x_i^T \right)^{-1}.}$$

Fisher information.

Def. For a (regular) class of probability distributions $(p_\theta)_{\theta \in \mathbb{R}^d}$, the Fisher information at $\theta = \theta_0$ is defined as

$$I(\theta_0) = \mathbb{E}_{\theta_0}\left[- \nabla_\theta^2 \log p_\theta(y) \Big|_{\theta=\theta_0} \right]$$

Side note:

$$\dot{\ell}_{\theta_0}(y) = \nabla_\theta \log p_\theta(y) \Big|_{\theta=\theta_0} \quad \text{(score)}$$
$$\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}(y)] = 0$$
$$\mathrm{Cov}_{\theta_0}(\dot{\ell}_{\theta_0}(y)) = I(\theta_0)$$

In GLM: $\ell_\beta(x,y) = \sum_{i=1}^n \log p_{\theta_i}(y_i) = \sum_{i=1}^n \left(T(y_i)\beta^T x_i - A(\beta^T x_i)\right)$

$+ \text{const}(x,y)$

$$\dot{\ell}_\beta(x,y) = \nabla_\beta \ell_\beta(x,y) = \sum_{i=1}^n \underbrace{\left(T(y_i) - A'(\beta^T x_i)\right)}_{\text{has mean zero}} x_i$$

$$\ddot{\ell}_\beta(x,y) = \nabla_\beta \dot{\ell}_\beta(x,y) = -\sum_{i=1}^n A''(\beta^T x_i) x_i x_i^T$$

$$\Rightarrow I(\beta) = \mathbb{E}[-\ddot{\ell}_\beta(x,y)] = \sum_{i=1}^n A''(\beta^T x_i) x_i x_i^T .$$

$$\boxed{\text{(Asymptotic) Cramér-Rao bound: } I(\theta)^{-1} \text{ is the "best" covariance of any asymptotically unbiased estimator } \hat{\theta} \text{ for } \theta \text{ as } n \to \infty.}$$

$$\boxed{\text{Asymptotic efficiency of MLE: } \hat{\theta}^{MLE} \text{ asymptotically achieves the Cramér-Rao bound.}}$$

Bootstrap estimate for $\text{Cov}(\hat{\beta}^{MLE})$: same as Lecture 3.

Inference in GLM.

Recall: analysis of variance (ANOVA) in linear regression

Problem: fit $y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$, test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p_0} = 0$ vs. H_1: not H_0.

Idea: fit two models

- full model: $y_i = \hat{\beta}_1^{(F)} x_{i,1} + \cdots + \hat{\beta}_p^{(F)} x_{i,p}$, obtain

  residual sum of squares $RSS_{full} = \sum_i (y_i - \hat{\beta}_1^{(F)} x_{i,1} - \cdots - \hat{\beta}_p^{(F)} x_{i,p})^2$

- reduced model: $y_i = \hat{\beta}_{p_0+1}^{(R)} x_{i,p_0+1} + \cdots + \hat{\beta}_p^{(R)} x_{i,p}$ (i.e. pretending that H_0 holds),

  obtain $RSS_{reduced} = \sum_i (y_i - \hat{\beta}_{p_0+1}^{(R)} x_{i,p_0+1} - \cdots - \hat{\beta}_p^{(R)} x_{i,p})^2$

ANOVA table:

Model	RSS	degree of freedom	F-statistic	p-value
Full	RSS_{full}	$n-p$		
Reduced	$RSS_{reduced}$	$n-(p-p_0)$		
Difference	$\underbrace{RSS_{reduced} - RSS_{full}}_{=: \Delta RSS}$	p_0	$\frac{\Delta RSS/p_0}{RSS_{full}/(n-p)}$	calculated from $F_{p_0, n-p}$

Intuition: if $\Delta RSS/p_0$ is too large, then ignoring features $(X_{:,1}, \cdots, X_{:,p_0})$ incurs a too large loss in RSS, and we should reject H_0 (F-statistic will be large)

GLM: analysis of deviance

Problem: same hypothesis testing, with linear regression replaced by GLM

Idea: again, fit two models:

Full model: $y \sim glm(X, family)$, obtain fitted log-likelihood ℓ_{full}

Reduced model: $y \sim glm((X_{p_0+1}, \cdots, X_p), family)$, obtain $\ell_{reduced}$

Analysis of deviance table:

Model	$2\times$ log-likelihood	degree of freedom	p-value
Full	$2\ell_{full}$	$n-p$	
Reduced	$2\ell_{reduced}$	$n-(p-p_0)$	
Difference	$\underbrace{2(\ell_{full} - \ell_{reduced})}_{\text{deviance in GLM!!}}$	p_0	compare deviance with $\chi^2_{p_0}$

Justification: Wilks' Theorem states that under H_0,

$$2(\ell_{full} - \ell_{reduced}) \xrightarrow{d} \chi^2_{p_0} \quad \text{as } n \to \infty.$$

Compare with ANOVA table: in linear regression, can show

$$\text{deviance} = 2(\ell_{full} - \ell_{reduced}) = \frac{\Delta RSS}{\sigma^2}, \quad \text{with } \sigma^2 = \text{Var}(\varepsilon_i).$$

Statisticians use $\hat{\sigma}^2 = \frac{RSS_{full}}{n-p}$ to estimate σ^2, so the F-statistic is

$$\frac{\Delta RSS / p_0}{RSS_{full}/(n-p)} = \frac{\sigma^2}{\hat{\sigma}^2} \cdot \frac{\text{deviance}}{p_0} \approx \frac{\text{deviance}}{p_0} \sim \frac{\chi^2_{p_0}}{p_0} \approx F_{p_0, n-p} \quad \text{as } n \to \infty.$$

Model selection.

Problem: fit a GLM $y \sim \text{glm}(x_1 + x_2 + \cdots + x_j, \text{family})$, but don't know where to end (i.e. choose $j \in \{1, 2, \cdots, p\}$). How to find the best j?

Idea: for each $j \in \{1, 2, \cdots, p\}$, fit a GLM and compute the fitted log-likelihood ℓ_j

(note that $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_p$, and model j has j parameters)

1. AIC (Akaike information criterion)

$$\boxed{\hat{j}^{AIC} = \underset{j \in \{1, 2, \cdots, p\}}{\arg\min} \ \underbrace{2j - 2\ell_j}_{AIC_j}}$$

2. BIC (Bayesian information criterion).

$$\boxed{\hat{j}^{BIC} = \underset{j \in \{1, 2, \cdots, p\}}{\arg\min} \ \underbrace{j \log n - 2\ell_j}_{BIC_j}}$$

3. Lasso (without the need of fitting p+1 models in advance)

$$\boxed{\hat{\beta}^{Lasso} = \underset{\beta}{\arg\min} \ -\frac{1}{n} \sum_{i=1}^{n} \log P_{x_i^T \beta}(y_i) + \lambda \|\beta\|_1}$$

- λ is typically chosen by cross validation.

Application: Density estimation via Lindsey's method

Given i.i.d. $z_1, \cdots, z_n \sim p$, aim to fit

$$p \approx p_\theta = \exp(\langle \theta, T(z) \rangle - A(\theta)) h(z)$$

- known: $T(\cdot)$, $h(\ )$ • unknown: $\theta \in \mathbb{R}^d$.

Problem with MLE: log-partition function $A(\theta)$ untractable (more in Lec 6)

Lindsey's method.

- Suppose $\mathcal{Z} \subseteq \mathbb{R}$, and $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2 \cup \cdots \cup \mathcal{Z}_K$, with

$$\mathcal{Z}_k = [z_k - \tfrac{\Delta_k}{2}, z_k + \tfrac{\Delta_k}{2}].$$

- For small Δ_k,

$$\mathbb{P}(z \in \mathcal{Z}_k) = \int_{\mathcal{Z}_k} p_\theta(z) dz$$
$$\approx \exp(\langle \theta, T(z_k) \rangle - A(\theta)) h(z_k) \Delta_k =: p_k$$

- For $y_k = \#\{z_i \in \mathcal{Z}_k\}$, then

$$(y_1, \cdots, y_K) \sim \mathrm{Multi}(n; (p_1, \cdots, p_K))$$

- Poisson trick: fit

$$y_k \overset{\text{ind.}}{\sim} \mathrm{Poi}(e^{\langle \theta, T(z_k) \rangle + \log(h(z_k)\Delta_k) + \theta_0})$$

This is a Poisson GLM!

- Poisson conditioning property:

if $y_i \overset{\text{ind.}}{\sim} \mathrm{Poi}(\lambda_i)$, then
$$(y_1, \cdots, y_K) \mid \sum_{k=1}^K y_k = n \sim \mathrm{Multi}\left(n; \left(\frac{\lambda_1}{\sum_k \lambda_k}, \cdots, \frac{\lambda_K}{\sum_k \lambda_k}\right)\right)$$

Therefore, $(y_1, \cdots, y_K) \mid \sum_{k=1}^K y_k = n \sim \mathrm{Multi}(n; (q_1, \cdots, q_K))$, with

$$q_k = \frac{\exp(\langle \theta, T(z_k) \rangle + \log(h(z_k)\Delta_k) + \theta_0)}{\sum_j \exp(\langle \theta, T(z_j) \rangle + \log(h(z_j)\Delta_j) + \theta_0)}$$

$$\propto \exp(\langle \theta, T(z_k) \rangle) h(z_k) \Delta_k = p_k.$$

(alternative view in HW)

- Think: what does θ_0 represent?