IRLS in Exponential Families

Elvis Cui

MS in Biostatistics, UCLA

November 21, 2019

Overview

1. Canonical Exponential Families
   - Properties of CEF
   - MLE in CEF
2. Generalized Linear Models
   - Building low-dimensional CEF
   - Newton-Raphson as Banach fixed point theorem
   - Fisher scoring
3. Next Time: Variable Selection (requested by DaWangSh)

Canonical Exponential Families

Definition 1 (CEF as Radon-Nikodym Derivatives)

Let Q be any probability measure. Then the set of probability measures $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \mathcal{H} \subset \mathbb{R}^p\}$ is called a **canonical exponential family (CEF) generated by** Q iff its density w.r.t. Q (Radon-Nikodym derivative) has the following form:

$$\frac{d\mathbb{P}_\eta}{dQ}(x) = \exp\{\langle \eta, x \rangle - K(\eta)\} \; Q - a.s.$$

Where $K(\eta) = \log \int_\Omega \exp\{\langle \eta, x \rangle\} Q(dx)$ is known as **cumulant generating function** in statistics.

Properties of CEF

Theorem 2 (Moments of CEF)

Let $X \sim \mathbb{P}_\eta$, then we have

- $\mathbb{E}(X) = \dot{K}(\eta)$ *where* $\dot{K}$ *is the first order derivative.*
- $Var(X) = \ddot{K}(\eta)$
- $\mathbb{E}(e^{s^T X}) = \exp\{K(\eta + s) - K(\eta)\}$

Proof.

Use the definition of expectation and variance, then calculate them. □

Properties of CEF cont'd

Theorem 3 (MLE in CEF)

Suppose $X_1, \cdots, X_n \sim^{iid} \mathbb{P}_\eta$. *Then we have (as an exercise)*

- *the joint distribution of* $X = (X_1, \cdots, X_n) \sim \mathbb{P}_\eta$. *Explicitly,*

$$\frac{d\mathbb{P}}{d \otimes_1^n Q}(X) = \exp\{\langle \eta, \sum_{i=1}^n x_i \rangle - nK(\eta)\} \ \otimes_1^n Q - a.s.$$

- *the maximum likelihood estimation for* η *is the solution to the following equation:*

$$\frac{\sum_{i=1}^n x_i}{n} = \dot{K}(\eta) = \mathbb{E}(X_1)$$

- *If we wish to use other parameterization (say, mean), then the MLE stays the same due to the* **invariance principle**.

Generalized Linear Models

Definition 4 (GLM)

To define a GLM, we need 3 components (Rao, 2007):

- the **random component**: it specifies the probability distribution of the response variable (which is canonical exponential family).
- the **systematic component**: it specifies a linear function of the explanatory variables.
- the **link function**: it describes a functional relationship between the systematic component and the expectation of the random component.

For those who are not familiar with the structure of GLM, I recommend **Rao's Linear models and generalizations**, Chapter 10.

GLM for Binary Response

Model Set-up:

- $\mathbf{y} = (y_1, \cdots, y_n)$: responses, $y_i \in \{0, 1\}$.
- $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$: explanatory variables.
- $p_i, i = 1, \cdots, n$ s.t. $Y_i \sim Ber(p_i)$.
- $g(\cdot)$: link function s.t. $\eta_i = g(p_i) = x_i^T \beta$ where η_i is a function of p_i and β is the parameter that we want to estimate.

GLM for Binary Response cont'd

How do we choose $g(\cdot)$ or how do we specify the relation between p and η?

- **Probit model**: $\eta_i = \Phi^{-1}(p_i)$ where Φ is the cdf of standard normal and -1 denotes the inverse function.
- **Logistic model**: $\eta_i = \log \frac{p_i}{1-p_i}$. In this case, we have the famous relationship (logistic function):

$$p_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

- **Log-log model**: $\eta_i = \log(-\log(1 - p_i))$

Canonical Link Function

If we take η_i to be the canonical parameter in the distribution of y_i, then $g(p_i)$ is called **canonical link function**.

- Suppose we have data points $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$ and $X = (x_1, \cdots, x_n) \in \mathbb{R}^{n \times p}$, the the joint dist. of y (w.r.t. measure Q) is:

$$\frac{d\mathbb{P}}{d \otimes_1^n Q}(y|p_1, \cdots, p_n) = e^{\eta^T y - K(\eta)} \quad \otimes_1^n Q - a.s.$$

  which is a $n-$dimensional CEF.

- Subsititue η with $X\beta$, we have

$$\frac{d\mathbb{P}}{d \otimes_1^n Q}(y|p_1, \cdots, p_n) = e^{\beta^T X^T y - A(\beta)} \quad \otimes_1^n Q - a.s.$$

  Surpringly, this can be viewed a $p-$dimensional CEF !!!

MLE for β

Since we have

$$\frac{d\mathbb{P}}{d \otimes_1^n Q}(y|p_1, \cdots, p_n) = e^{\beta^T X^T y - A(\beta)} \quad \otimes_1^n Q - a.s.$$

To derive the MLE for β, we could apply the theory for CEF. $\hat{\beta}_{MLE}$ is the unique solution to the following equation:

$$X^T y = \dot{A}(\beta) = \mathbb{E}(X^T y) = X^T \mu_y$$

In the logistic model, we have $\mu_y = p = (p_1, \cdots, p_n)$. However, even we know that μ_y is a function of β, it is a **non-linear** function so there is no closed-form solution. Thus, we have to use numerical method.

Newton-Raphson Method: contraction mapping

To solve for MLE, we recall the **Newton-Raphson Method**, which is a special case of **Banach fixed point theorem**. To state this theorem, we first need the definition of "contraction mapping" (from wikipedia).

Definition 5 (Contraction mapping)

Let (X, d) be a complete metric space. Then a map $T : X \to X$ is called a **contraction mapping** on X iff

$$\exists q \in [0, 1) \text{such that } d(T(x), T(y)) \leq qd(x, y) \forall x, y \in X$$

Newton-Raphson Method: Banach fixed point theorem

Theorem 6 (Banach fixed point theorem)

Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed-point X^* in X. That is, $T(x^*) = x^*$. In addition, generating $\{x_n\}$ by $x_n = T(x_{n-1})$ with $x_0 \in X$ will give us $x_n \to x^*$.

Newton-Raphson Method: Algorithm

Definition 7 (Newton-Raphson Method)

Suppose $f : \mathbb{R}^p \to \mathbb{R}^p$ and the matrix f' exists and we want to find the solution to the following equation:

$$f(x_0) = 0$$

Suppose the root is unique, then such solution can be derived by generating $\{x_n\}$ with an arbitrary initialization x_0:

$$x_{n+1} = x_n - [f'(x_n)]^{-1} f(x_n)$$

Then by Banach fixed point theorem, such algorithm are guaranteed to converge to some point $x^* \in \mathbb{R}^p$.

Newton-Raphson Method: Fisher Scoring

- Now consider finding MLE for β in GLM (see slide 10), we take f to be **the derivative of our log-likelihood function**, then f' will be **the Hessian matrix**.
- However, we know that the expectation of the Hessian of negative log-likelihood is the **Fisher information matrix**, we could replace Hessian by the estimation of Fisher information matrix since **they are equivalent asymptotically**.

Newton-Raphson Method: Fisher scoring cont'd

- To calculate $-\mathbb{E}\left[\frac{\partial^2}{\partial\beta\partial\beta^T}\log\mathbb{P}(y|\beta)\right]$, we use the theory of CEF to get:

$$\begin{aligned}\mathcal{I}_n(\beta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\beta\partial\beta^T}\log\mathbb{P}(y|\beta)\right] \\ &= \ddot{A}(\beta) \\ &= Var(X^T y) \\ &= X^T\Sigma X\end{aligned}$$

- Note that Σ is a diagonal matrix and the diagonal can be represented as a function of β.
- $\frac{\partial}{\partial\beta}\log\mathbb{P}(y|\beta) = X^T y - \dot{A}(\beta) = X^T(y-\mu_y)$ and μ_y is also a function of β.

Fisher Scoring Algorithm

Definition 8 (Fisher Scoring)

Fisher Scoring: Initialize $\beta^{(0)}$, then update it as the following:

$$\beta^{(t+1)} = \beta^{(t)} - \left[X^T \Sigma^{(t)} X\right]^{-1} (y - \mu_y^{(t)})$$

Rearrange this, we derive:

$$\beta^{(t+1)} = \left[X^T \Sigma^{(t)} X\right]^{-1} X^T \Sigma^{(t)} z^t$$

where

$$z^{(t)} = X\beta^{(t)} - (\Sigma^{(t)})^{-1}(\mu_y^{(t)}) - y$$

Thus, Fisher scoring is also known as **Iterative Reweighted Least Square method (IRLS)**.

Next Time: Variable Selection (requested by DaWangSh)

- Tikhonov regularization
  1. L_2 regularization (a.k.a. Ridge)
  2. Kernelized L_2
- Sparse regularization
  1. L_1 reg. (a.k.a. Lasso)
  2. Spectral norm reg. (a.k.a. Rank-reduced)
  3. SCAD (J. Fan and R. Li, 2001)
  4. Graphical Lasso
  5. CLIME (T. Cai et al, 2012)
- Other regularizations
  1. Frobenius norm regularization
  2. ESZSL (in computer vision)
  3. Elastic net