10-708: Probabilistic Graphical Models 10-708, Spring 2016

5 : Exponential Family and Generalized Linear Models

Lecturer: Matthew Gormley *Scribes: Yuan Li, Yichong Xu, Silun Wang*

1 Exponential Family

Probability density functions that are in exponential family can be expressed in the following form.

$$p(x|\eta) = h(x)exp\{\eta^T T(x) - A(\eta)\}$$

$$A(\eta) = log \int h(x)exp\{\eta^T T(x)\}dx$$

One example of exponential family is multinomial distribution. For given data $\mathbf{x} = (x_1, x_2, ..., x_k)$, $x_i \sim$ Multi(1, π_i) and $\Sigma\ \pi_i = 1$, we can write the probability of the data as follows.

$$\begin{aligned} p(\mathbf{x}|\pi) &= \pi_1^{x_1}\pi_2^{x_2}...\pi_k^{x_k} \\ &= e^{log(\pi_1^{x_1}\pi_2^{x_2}...\pi_k^{x_k})} \\ &= e^{\Sigma_{i=1}^k x_i log\pi_i} \end{aligned}$$

Thus, in the corresponding exponential form, η = [log π_1, $log\pi_2$, ..., $log\pi_k$], $\mathbf{x}$ = [x_1, x_2, ..., x_k], T($\mathbf{x}$) = $\mathbf{x}$, A(η) = 0 and h($\mathbf{x}$) = 1.

Another example is Dirichlet distribution. Let α_1, α_2, ... , $\alpha_k > 0$. The probability function of such distribution can be represented as an exponential function.

$$\begin{aligned} p(\pi|\alpha) &= \frac{1}{B(\alpha)}\Pi_{i=1}^K \pi_i^{\alpha_i - 1} \\ &= e^{\Sigma_{i=1}^K (\alpha_i - 1)log\pi_i - logB(\alpha)} \end{aligned}$$

where A(η) = $logB(\alpha)$, $\eta = [\alpha_1, \alpha_2, ..., \alpha_k]$, $T(\mathbf{x}) = [log\pi_1, log\pi_2, ..., \pi_k]$ and $h((x)) = 1$.

2 Cumulant Generating Property

Notice that one appealing feature of the exponential family is that we can easily compute moments of the distribution by taking derivatives of the log normalizer $A(\eta)$.

2.1 First cumulant a.k.a Mean

The first derivative of $A(\eta)$ is equal to the mean of sufficient statistics $T(X)$.

$$\begin{aligned}
\frac{dA}{d\eta} &= \frac{d}{d\eta} logZ(\eta) \\
&= \frac{1}{Z(\eta)} \frac{d}{d\eta} Z(\eta) \\
&= \frac{1}{Z(\eta)} \frac{d}{d\eta} \left\{ \int h(x) exp\{\eta^T T(x)\} dx \right\} \\
&= \int T(x) \frac{h(x) exp\{\eta^T T(x)\}}{Z(\eta)} dx \\
&= \int T(x) p(x|\eta) dx \\
&= E[T(X)]
\end{aligned}$$

2.2 Second cumulant a.k.a Variance

The second derivative of $A(\eta)$ is equal to the variance or first central moment of sufficient statistics $T(X)$.

$$\begin{aligned}
\frac{d^2A}{d\eta^2} &= \int T(x) exp\{\eta^T T(x) - A(\eta)\} (T(x) - \frac{dA}{d\eta}) h(x) dx \\
&= \int T^2(x) \frac{h(x) exp\{\eta^T T(x)\}}{Z(\eta)} dx - \frac{dA}{d\eta} \int T(x) \frac{h(x) exp\{\eta^T T(x)\}}{Z(\eta)} dx \\
&= \int T^2(x) p(x|\eta) dx - \frac{dA}{d\eta} \int T(x) p(x|\eta) dx \\
&= E[T^2(X)] - (E[T(X)])^2 \\
&= Var[T(X)]
\end{aligned}$$

2.3 Moment estimation

Accordingly, the q^{th} derivative gives the q^{th} centered moment. When the sufficient statistic is a stacked vector, partial derivatives need to be considered.

2.4 Moment vs canonical parameters

Since the moment parameter μ can be derived from the natural (canonical) parameter η by:

$$\frac{dA(\eta)}{\eta} = E[T(x)] \stackrel{\text{def}}{=} \mu$$

Also notice that $A(\eta)$ is convex since:

$$\frac{d^2A(\eta)}{d\eta^2} = Var[T(x)] > 0$$

Hence we can invert the relationship and infer the canonical parameter from the moment parameter (1-to-1) by:

$$\eta \triangleq \psi(\mu)$$

which means a distribution in the exponential family can be parameterized not only by η (the canonical parameterization), but also by μ (the moment parameterization).

3 Sufficiency

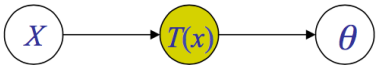
For $p(x|\theta)$, T(x) is sufficient for θ if there is no information in X regarding θ beyond that in T(x).

$$\theta \perp\!\!\!\perp X|T(X)$$

However, it is defined in different ways in the Bayesian and frequentist frameworks.

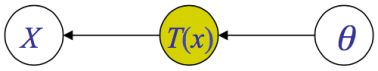
Bayesian view θ as a random variable. To estimate θ, $T(X)$ contains all the essential information in X.

$$p(\theta|T(x), x) = p(\theta|T(x))$$



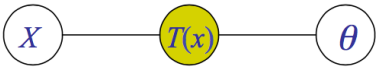
Frequentist view θ as a label rather than a random variable. $T(X)$ is sufficient for θ if the conditional distribution of X given $T(X)$ is not a function of θ.

$$p(x|T(x), \theta) = p(x|T(x))$$



For undirected models, we have

$$p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))$$



Since $T(x)$ is function of x, we can drop $T(x)$ on the left side, and then divide it by $p(\theta)$.

$$p(x|\theta) = g(T(x), \theta)h(x, T(x))$$

Another important feature of the exponential family is that one can obtain the sufficient statistics $T(X)$ simply by inspection. Once the distribution function is expressed in the standard form,

$$p(x|\eta) = h(x)exp\{\eta^T T(x) - A(\eta)\}$$

we can directly see $T(X)$ is sufficient for η.

4 MLE for Exponential Family

The reduction obtained by using a sufficient statistic $T(X)$ is particularly notable in the case of IID sampling. Suppose the dataset D is composed of N independent random variables, characterized by the same exponential family density. For these i.i.d data, the log-likelihood is

$$\begin{aligned} l(\eta; D) &= log \prod_{n=1}^{N} h(x_n) exp\{\eta^T T(x_n) - A(\eta)\} \\ &= \sum_{n=1}^{N} log(h(x_n)) + (\eta^T \sum_{n=1}^{N} T(x_n)) - NA(\eta) \end{aligned}$$

Take derivative and set it to zero, we can get

$$\begin{aligned} \frac{\partial l}{\partial \eta} &= \sum_{n=1}^{N} T(x_n) - N \frac{\partial A(\eta)}{\partial \eta} = 0 \\ \frac{\partial A(\eta)}{\partial \eta} &= \frac{1}{N} \sum_{n=1}^{N} T(x_n) \\ \hat{\mu}_{MLE} &= \frac{1}{N} \sum_{n=1}^{N} T(x_n) \\ \hat{\eta}_{MLE} &= \psi(\hat{\mu}_{MLE}) \end{aligned}$$

Our formula involves the data only via the sufficient statistic $\sum_{n=1}^{N} T(X_n)$. This means that to estimate MLE of η, we only need to maintain fixed dimensions of data. For Bernouilli, Poisson and multinomial distributions, it suffices to maintain a single value, the sum of the observations. Individual data points can be thrown away. While for the univariate Gaussian distribution, we need to maintain the sum $\sum_{n=1}^{N} x_n$ and the sum of squares $\sum_{n=1}^{N} x_n^2$.

4.1 Examples

1. Gaussian distribution: We have

$$\begin{aligned} \eta &= \left[\Sigma^{-1}\mu; -\frac{1}{2}\text{vec}(\Sigma^{-1})\right] \\ T(x) &= \left[x; \text{vec}(xx^T)\right] \\ A(\eta) &= \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma| \\ h(x) &= (2\pi)^{-k/2}. \end{aligned}$$

So

$$\mu_{MLE} = \frac{1}{N}\sum_{n} T(x_n) = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

2. Multinomial distribution: We have

$$\begin{aligned} \eta &= \left[\ln \frac{\pi_k}{\pi_K}; 0\right] \\ T(x) &= [x] \\ A(\eta) &= -\ln\left(1 - \sum_{k=1}^{K-1} \pi_k\right) \\ h(x) &= 1. \end{aligned}$$

So

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

3. Poisson distribution: We have

$$\begin{aligned} \eta &= \log \lambda \\ T(x) &= x \\ A(\eta) &= \lambda = e^{\eta} \\ h(x) &= \frac{1}{x!}. \end{aligned}$$

So

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

5 Bayesian Estimation

5.1 Conjugacy

Prior

$$p(\eta|\phi) \propto exp\{\phi^T T(\eta) - A(\eta)\}$$

Likelihood

$$p(x|\eta) \propto exp\{\eta^T T(x) - A(\eta)\}$$

suppose $\eta = T(\eta)$, posterior

$$\begin{aligned} p(\eta|x,\phi) &\propto p(x|\eta)p(\eta|\phi) \\ &\propto exp\{\eta^T T(x) + \phi^T T(\eta) - A(\eta) - A(\phi)\} \\ &\propto exp\{ \underbrace{T(\eta)^T}_{\text{sufficient func}} (\underbrace{T(x) + \phi}_{\text{natural parameter}}) - (\underbrace{A(\eta) + A(\phi)}_{A(\eta,\phi)})\} \end{aligned}$$

6 Generalized Linear Model

GLIM is a generalized form of traditional linear regression. As in linear regression, the observed input x is assumed to enter the model via a linear combination of its elements $\xi = \theta^T x$. The output of the model, on the other hand, is assumed to have an exponential family distribution with conditional mean $\mu = f(\xi)$, where f is known as the response function. Note that for linear regression f is simply the identity function. Figure 1 is a graphical representation of GLIM. And Table 1 lists some correspondence between usual regression types and choice of f and Y.

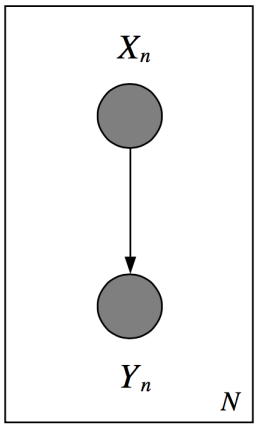


Figure 1: Graphical model of GLIM.

.

Regression Type	f	distribution of Y
Linear Regression	identity	$\mathcal{N}(\mu, \sigma^2)$
Logistic Regression	logistic	Bernoulli
Probit regression	cumulative Gaussian	Bernoulli
Multivariate Regression	logistic	Multivariate

Table 1: Examples of regression types and choice of f and Y.

6.1 Why GLIMs?

As a generalization of linear regression, logistic regression, probit regression, etc., GLIM provides a framework for creating new conditional distributions that comes with some convenient properties. Also GLIMs with the canonical response functions are easy to train with MLE.

However, Bayesian estimation of GLIMs doesn't have a closed form of posterior, so we have to turn to approximation techniques.

6.2 Properties of GLIM

Formally, we assume the output of GLIM has the following form:

$$p(y|\eta, \phi) = h(y, \phi) \exp\left[\frac{1}{\phi}(\eta^T(x)y - A(\eta))\right].$$

This is slightly different from the traditional definition of EF, where we include a new *scale parameter* ϕ; most distributions are naturally expressed in this form.

Note that $\eta = \psi(\mu)$ and $\mu = f(\xi) = f(\theta^T x)$, so we have $\eta = \psi(f(\theta^T x))$. So the conditional distribution of y given x, θ and ϕ is

$$p(y|x, \theta\phi) = h(y, \phi) \exp\left[\frac{1}{\phi}(y^T \psi(f(\theta^T x)) - A(\psi(f(\theta^T x))))\right].$$

There're mostly 2 design choices of GLIM: the choice of exponential family and the choice of f. The choice of the exponential family is largely constrained by the nature of the data y. E.g., for continuous y we use multivariate Gaussian, where for discrete class labels we use Bernoulli or multinomial. Response function is usually chosen with some mild constraints, e.g., between $[0, 1]$ and being positive. There's a so-called *canonical response* function where we use $f = \psi^{-1}$; in this case the conditional probability is simplified to

$$p(y|x, \theta\phi) = h(y, \phi) \exp\left[\frac{1}{\phi}(\theta^T x \cdot y - A(\theta^T x))\right].$$

Figure 2 lists canonical response function for several distributions. Figure 3 and table 2 lists the relationship

Model	Canonical response function
Gaussian	$\mu = \eta$
Bernoulli	$\mu = 1/(1 + e^{-\eta})$
multinomial	$\mu_i = \eta_i / \sum_j e^{\eta_j}$
Poisson	$\mu = e^{\eta}$
gamma	$\mu = -\eta^{-1}$

Figure 2: Canonical response function for several distributions.

.

between variables and canonical functions.

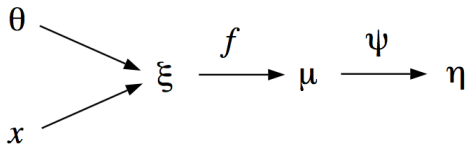


Figure 3: Relationship between variables and functions.

.

Regression Type	Canonical Response	$\mu = f(\xi)$	$\eta = f^{-1}(\mu)$	distribution of Y
Linear Regression	Y	$\mu = \xi$	$\eta = \mu$	$\mathcal{N}(\mu, \sigma^2)$
Logistic Regression	Y	$\mu = \frac{1}{1+\exp(-\xi)}$	$\eta = \log \frac{\mu}{1-\mu}$	Bernoulli(μ)
Probit regression	N	$\mu = \phi(\xi)$	$\eta = \phi^{-1}(\mu)$	Bernoulli(μ)

Table 2: Some regression types and their response/link functions.

6.3 MLE estimation for GLIMs with canonical response

Now we can compute the MLE estimation for canonical response functions: the log likelihood function is

$$l = \sum_n \log h(y_n) + \sum_n (\theta^T x_n y_n - A(\eta_n)).$$

Take derivative with respect to θ(note that θ is the only parameter for GLIMs with canonical response):

$$\frac{dl}{d\theta} = \sum_n \left(x_n y_n - \frac{dA(\eta_n)}{d\eta_n}\frac{d\eta_n}{d\theta} \right) = \sum_n (y_n - \mu_n)x_n = X^T(y-\mu).$$

So we can do stochastic gradient ascent with update rule

$$\theta^{(t+1)} = \theta^{(t)} + \rho(y_n - (\theta^{(t)})^T x_n)x_n$$

where ρ is the step size.

Another method is to use Newton-Raphson methods to obtain a batch-learning algorithm: The update rule is

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1}\nabla_\theta J$$

where J is the cost function and H is the Hessian matrix (second derivative). We have

$$\nabla_\theta J = X^T(y-\mu),$$

and

$$\begin{aligned} H &= \frac{\partial^2 l}{\partial\theta\partial\theta^T} = \frac{\partial}{\partial\theta^T}\sum_n (y_n - \mu_n)x_n = \sum_n x_n \frac{\partial\mu_n}{\partial\theta^T} \\ &= -\sum_n x_n \frac{\partial\mu_n}{\partial\eta_n}\frac{\partial\eta_n}{\partial\theta_T} \\ &= -\sum_n x_n \frac{\partial\mu_n}{\partial\eta_n} x_n^T \\ &= -X^T W X, \end{aligned}$$

where $W = \text{diag}\left(\frac{d\mu_1}{d\eta_1}, \frac{d\mu_2}{d\eta_2}, \cdots, \frac{d\mu_N}{d\eta_N}\right)$. So the update rule is

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1}\nabla_\theta J = (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)}$$

where the adjusted response is $z^{(t)} = X\theta^{(t)} + (W^{(t)})^{-1}(y - \mu^{(t)})$.