10-708: Probabilistic Graphical Models 10-708, Spring 2014

6: The Exponential Family and Generalized Linear Models

Lecturer: Eric P. Xing *Scribes: Alnur Ali (lecture slides 1-23), Yipei Wang (slides 24-37)*

1 The exponential family

A distribution over a random variable $\mathbf{X}$ is *in the exponential family* if you can write it as

$$P(\mathbf{X} = \mathbf{x}; \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})\right).$$

Here, $\boldsymbol{\eta}$ is the vector of *natural parameters*, $\mathbf{T}$ is the vector of *sufficient statistics*, and A is *the log partition function*[1]

1.1 Examples

Here are some examples of distributions that are in the exponential family.

1.1.1 Multivariate Gaussian

Let $\mathbf{X}$ be $\in \mathbb{R}^p$.

Then we have:

$$\begin{aligned} P(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(\operatorname{tr} \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \ln|\boldsymbol{\Sigma}|)\right) \\ &= \underbrace{\frac{1}{(2\pi)^{p/2}}}_{h(\mathbf{x})} \exp\left(-\frac{1}{2}\underbrace{\operatorname{tr}\, \boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^T}_{\operatorname{vec}(\boldsymbol{\Sigma}^{-1})^T \operatorname{vec}(\mathbf{x}\mathbf{x}^T)} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \underbrace{\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\boldsymbol{\Sigma}|)}_{A(\boldsymbol{\eta})}\right), \end{aligned}$$

where vec$(\cdot)$ is the vectorization operator.

[1] It's called this, since in order for P to normalize, we need $\exp(A(\boldsymbol{\eta}))$ to equal $\int_{\mathbf{x}} h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})) \Rightarrow A(\boldsymbol{\eta}) = \ln\left(\int_{\mathbf{x}} h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}))\right)$, which is the log of the usual normalizer, which is the partition function.

This implies that:

$$\begin{aligned}
\boldsymbol{\eta} &= \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})\right) \\
\mathbf{T}(\mathbf{x}) &= \left(\mathbf{x}, \text{vec}(\mathbf{x}\mathbf{x}^T)\right) \\
A(\boldsymbol{\eta}) &= \frac{1}{2}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \ln|\boldsymbol{\Sigma}|) \\
h(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2}}.
\end{aligned}$$

1.1.2 Bernoulli

A common way to take distributions and show they're exponential family is to take the $\exp(\ln(\cdot))$ of (at least part of) their functional form.

E.g.:

$$\begin{aligned}
P(x;p) &= p^x(1-p)^{1-x} \\
\Rightarrow \ln P(x;p) &= x\ln(p) + (1-x)\ln(1-p) \\
&= x\ln(p) - x\ln(1-p) + \ln(1-p) \\
&= x\left(\ln(p) - \ln(1-p)\right) + \ln(1-p) \\
&= x\ln(\frac{p}{1-p}) + \ln(1-p) \\
\Rightarrow \exp(\ln P(x;p)) &= \exp\left(x\ln(\frac{p}{1-p}) + \ln(1-p)\right)
\end{aligned}$$

This implies that:

$$\begin{aligned}
\eta &= \ln(\frac{p}{1-p}) \\
T(x) &= x \\
A(\eta) &= -\ln(1-p) \\
h(x) &= 1.
\end{aligned}$$

1.1.3 Others

Just to list a few: the univariate Gaussian, Poisson, gamma, multinomial, linear regression, Ising model, restricted Boltzmann machines, and conditional random fields (CRFs) are all in the exponential family.

Multinomial If you try to follow this same logic as with the Bernoulli in order to write the multinomial as exponential family, you'll end up with $A(\boldsymbol{\eta}) = 0$, which is a problem. To avoid this, we'll require that $\boldsymbol{\eta}$ be an open rectangle (i.e. all possible combinations of one open interval for each component in $\boldsymbol{\eta}$), which is called a *full rank* exponential family. This lets us write the multinomial as an exponential family distribution by allowing the first $p-1$ parameters to form an open rectangle, and constraining the last parameter to be one minus the sum of the first $p-1$.

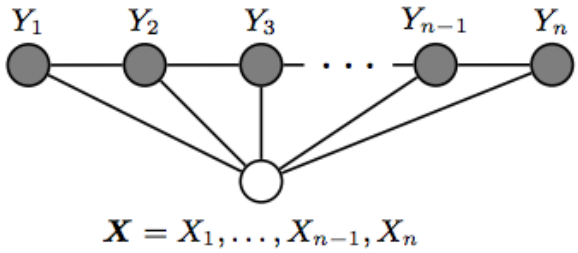

Figure 1: The CRF.

Conditional Random Fields (CRFs) Think of part of speech tagging: given a string of words $X \triangleq (X_1, \ldots, X_n)$, we want to predict the part of speech of each word $Y \triangleq (Y_1, \ldots, Y_n)$. HMMs are used for this, but are limited in (at least) two ways: (1) they require the joint: i.e. you must know every possible combination of $X \times Y$ (2) no context can be used during decision-making: any output Y_i is only a function of the corresponding input X_i. CRFs tries to work around these limitations by modeling the sought-after conditional $P(y|x)$ directly.

CRFs are a log linear model based on the undirected graphical model in Figure 1. Potentials are defined on pairwise outputs Y_i and Y_{i-1}, as well as on each output Y_i, and we always condition on all the inputs X. This gives the following conditional probability:

$$P(y|x) = \frac{1}{Z} \exp\left(\sum_{i=2}^{n} \lambda_i t_i(y_{i-1}, y_i, x, i) + \sum_{i=1}^{n} \mu_i s_i(y_i, x, i)\right).$$

Here, the λ's and μ's are parameters that need to be learned, and the t's and s's are feature functions.

1.2 Properties

1. The exponential family has the following property (called *the moment generating property*): the d'th derivative of the log partition equals the d'th centered moment of the sufficient statistic (if you have a vector of sufficient statistics, then $\partial^d A / \partial \eta_i^d = \mathrm{E}[\mathbf{T}(\mathbf{x})_i^d]$).

   E.g., the first derivative of the log partition function is the mean of $T(X)$; the 2nd is its variance.

2. This implies that the log partition function is convex, because its second derivative must be positive, since variance is always non-negative.

3. This further implies that: we can write the first derivative of the log partition function as a function of the natural parameter (aka *the canonical parameter*), set it equal to the mean, and then invert[2] to solve for the natural parameter in terms of the mean (aka *the moment parameter*). In symbols: $\eta = \psi(\mu)$.

4. Doing MLE on the exponential family is the same as doing moment matching. This follows by:

   (a) Writing down the log likelihood of a generic exponential family member: $\text{const} + \boldsymbol{\eta}^T(\sum_{i=1}^n \mathbf{T}(\mathbf{x}_i)) - nA(\boldsymbol{\eta})$.

   (b) Taking the gradient w.r.t. η: $\sum_{i=1}^n \mathbf{T}(\mathbf{x}_i) - n\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta})$.

[2]The first derivative of a convex function is strictly monotone up, which implies that the function is invertible.

Figure 2: Bayesian POV on sufficient statistics.

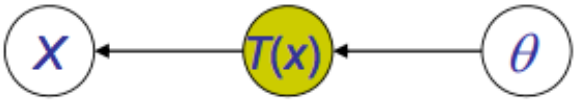

Figure 3: Frequentist POV on sufficient statistics.

(c) Setting equal to zero and solving for $\nabla_{\boldsymbol{\eta}} A$:
$\nabla_{\boldsymbol{\eta}} A = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}(\mathbf{x}_i) \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}(\mathbf{x}_i) \Rightarrow$ estimated moment = sample moment.

1.2.1 More on sufficient statistics

Bayesian point of view on sufficient statistics: if T tells us all we need to know to predict the population parameter θ, (i.e. T is a sufficient statistic), then $\theta \perp X \mid T \Rightarrow P(\theta|X,T) = P(\theta|T)$ (see the causal trail in Figure 2).

Frequentist POV: if T tells us all we need to know to generate data, then $X \perp \theta \mid T \Rightarrow P(X|T;\theta) = P(X|T)$ (see the evidential trail in Figure 3).

Markov random field POV:

1. Take the causal trail and drop its edges to get an undirected graph (see the MRF in Figure 4).
2. Write the joint $P(x,T,\theta)$ as $\psi_1(x,T)\psi_2(T,\theta)$ (use the 2 max clique potentials, and absorb $1/Z$ into either of them).
3. Notice that T always has the same value, so we can drop it from the lhs, giving: $P(X,\theta) = \psi_1(X,T)\psi_2(T,\theta)$.
4. Divide both sides by $P(\theta)$, and we get *the factorization theorem*: T is a sufficient statistic iff $P(X|\theta) = g(T,\theta)h(X,T)$, for some functions g and h.

One thing that's nice about the exponential family is that if X's distribution is in the exponential family, then instead of applying the factorization theorem to find a sufficient statistic for X, you can simply look at the T inside the $\exp(\cdot)$, and that's your sufficient statistic (e.g. x is a sufficient statistic for a Bernoulli rv, as discussed earlier). Another nice thing: a sufficient statistic for the sum of a bunch of iid rvs whose distribution is in the exponential family is just the sum of each rv's sufficient statistic.

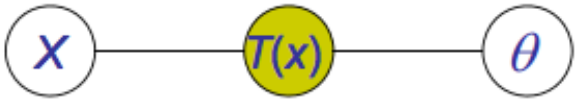

Figure 4: Markov random field POV on sufficient statistics.

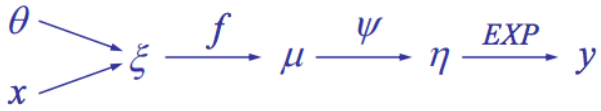

Figure 5: The GLIM framework.

1.3 Bayesian POV

Let's be Bayesian again for a second. After writing down the likelihood of the data given the natural parameter, we would want to pick a prior over the natural parameter, and then work out the posterior over the natural parameter.

E.g.:

$$\begin{aligned} P(\mathbf{x}|\boldsymbol{\eta}) &\propto \exp\left(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})\right) \\ P(\boldsymbol{\eta}) &\propto \exp\left(\boldsymbol{\xi}^T \mathbf{T}(\boldsymbol{\eta}) - A(\boldsymbol{\xi})\right) \\ P(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{\xi}) &\propto \exp\left(\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x}) + \boldsymbol{\xi}^T \mathbf{T}(\boldsymbol{\eta}) + A(\boldsymbol{\eta}) + A(\boldsymbol{\xi})\right) \end{aligned}$$

If $\boldsymbol{\eta} = \mathbf{T}(\boldsymbol{\eta})$, then the posterior is:

$$P(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{\xi}) \propto \exp\left(\mathbf{T}(\boldsymbol{\eta})(\mathbf{T}(\mathbf{x}) + \boldsymbol{\xi}) + A(\boldsymbol{\eta}) + A(\boldsymbol{\xi})\right),$$

which is exactly the same form as the prior.

So when $\boldsymbol{\eta} = \mathbf{T}(\boldsymbol{\eta})$ and we assert that $\boldsymbol{\eta} \sim$ exponentialFamily, then that prior will be a conjugate prior.

2 Generalized linear models (GLIMs)

Generalized linear models (GLIMs) are a statistical framework for unifying classification and regression. In this framework, we assume:

$$\begin{aligned} Y &\sim \text{exponentialFamily} \\ \eta &= \psi(\mu = f(\xi = \boldsymbol{\theta}^T \mathbf{x})), \end{aligned}$$

where Y are the responses, $\mathbf{x}$ are the fixed inputs, $\boldsymbol{\theta}$ are parameters we need to learn, and f (called *the response function*) and ψ give us added flexibility if we want it (f is often set to ψ^{-1}, in which it is called *the canonical response function*); see Figure 2.

Least squares can be viewed as GLIM with $Y \sim \mathcal{N}(\mu, \sigma^2)$, f set to the identity, and ψ set to the identity. Logistic regression can be viewed as a GLIM with $Y \sim$ Bernoulli(p), f set to a sigmoid, and ψ set to the identity.

2.1 Online learning of GLIMs

Online learning for $\boldsymbol{\theta}$ is done in the obvious way via stochastic gradient descent by (1) writing down the log likelihood $l(\cdot)$ of a single data point of a generic GLIM with the canonical response function (2) differentiating w.r.t. $\boldsymbol{\theta}$, setting equal to zero, and solving for $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ (3) doing this gives the update equations $\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} + \rho(y_i - \boldsymbol{\theta}_{\text{old}}^T \mathbf{x}_i)\mathbf{x}_i$, where ρ is a step size[3].

[3]Note that for the canonical response function, $\boldsymbol{\mu}_{\text{old}} = \boldsymbol{\theta}_{\text{old}}^T \mathbf{x}_i$

3 Batch Learning for canonical GLIMs

We first recall how to solve the minimization of the cost function through derivative directly. When we use least mean square training rule, we aim to minimize the cost function:

$$J(\theta) = 0.5 * \sum_{i=1}^{n} (x_i^T \theta - y_i)^2 = 0.5 * (X\theta - y)^T (X\theta - y)$$

where x_i represents the input of ith sample and y_i is the true output of the ith sample.

We set the derivative of $J(\theta)$ to zero to get the minimal solution:

$$\nabla J(\theta) = X^T X\theta - X^T y = 0 \Rightarrow \theta^* = (X^T X)^{-1} X^T y$$

Here we use Newton method to derive the batch learning algorithm to find the optimal solution iteratively. In Newton method, we have the update equation as below:

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

The hessian matrix is derived below:

$$\begin{aligned} H &= \frac{d^2 l}{d\theta d\theta^T} \\ &= \frac{d}{d\theta^T} \sum_n (y_n - u_n) x_n \\ &= \sum_n x_n \frac{du_n}{d\theta^T} \\ &= -\sum_n x_n \frac{u_n}{\eta_n} \frac{\eta_n}{d^T} \\ &= \sum_n x_n \frac{u_n}{\eta_n} x_n^T \\ &= -X^T W X \end{aligned} \tag{1}$$

where $X = [x_n^T]$, $W = diag\left[\frac{du_1}{d\eta_1}, \cdots, \frac{du_N}{d\eta_N}\right]$. Here, W can be computed by calculating the second derivative of $A(\eta_n)$

Substitute $\nabla J(\theta)$ and H we derived above, we can get

$$\theta^{t+1} = \left(X^T W^t X\right)^{-1} X^T W^t z^t$$

where $z^t = X\theta^t + \left(W^t\right)^{-1}(y - u^t)$. Since W is a diagonal matrix, the equation can be decoupled and the algorithm can be batched.

This can be viewed as Iteratively reweighted least squares problem:

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

In the rest of this section, we derived several examples.

3.1 Example: logistic regression

The condition distribution

$$p(y|x) = u(x)^y(q - u(x))^{1-y}$$

where $u(x) = \frac{1}{1+e^{-\eta(x)}}$

$P(y|x)$ is an exponential family function, with canonical response $\eta = \theta^T x$

Using IRLS method we discussed above, we calculate

$$\frac{du}{d\eta} = u(1-u)$$

$$W = \begin{pmatrix} u_1(1-u_1) & & \\ & \cdots & \\ & & u_N(1-u_N)) \end{pmatrix}$$

N is the number of training samples, d is the dimension of input x. IRLS method takes $O(Nd^3)$ per iteration. We can use Quasi-Newton method to approximate hessian matrix to reduce computational cost.

Conjugate gradient takes $O(Nd)$ per iteration. It usually works best in practice. Stochastic gradient decent method can also be used if N is large.

3.2 Example: linear regression

The condition distribution

$$p(y|x, \theta, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(y - u(x))^T\Sigma^{-1}(y - u(x))\right\}$$

where $u(x)$ is a linear function $u(x) = \theta^T x$.

Using IRLS, we calculate

$$\frac{du}{d\eta} = 1$$

$$W = 1$$

The update equation is:

$$\theta^{t+1} = \theta^t + (X^T X)^{-1} X^T (y - u^t)$$

4 GM as building block for complex Bayesian Networks

The three classical problems in machine learning can be represented by graphical model. For density estimation, it includes parametric and non-parametric methods. For regression, it includes linear regression, conditional mixture and non-parametric methods. For classification, it includes discriminative and generative approach.

Here we also discuss the relations among many GMs, like PCA, ICA, HMM, etc. All the models here are designed by human instead of learned from data. This is useful because we can embed human intelligence into the structure of the model and make it into good use.

For maximum likelihood estimation for general Bayesian network, the log likelihood function can be decomposed, assuming the parameters are globally independent and all nodes are fully observed. Therefore, the MLE method of GM reduces to MLE for each GLIM.

5 Parameter Prior

To solve the problem of defining appropriate parameter prior without missing information, Geiger and Hecherman introduced several assumptions and derive the distribution satisfy the simple assumptions.

Under the assumption of global parameter independence, they derive that for discrete DAG models, we can use Dirichlet prior

$$p(\theta) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$

For gaussian GAG models, we can use Normal-Wishart prior. The expression can refer to slide 36 for lecture 6.

6 Summary

1.The benefit of distribution in exponential family is that the maximum likelihood estimation problem amounts to moment matching problem
2.For generalized linear model:
-The general algorithm Iteratively reweighted least squares
-It is the building bolck of most GM model in practice use
3. We discuss how to appropriate priors and conclusion under simplified assumptions.