10-708: Probabilistic Graphical Models, Spring 2015

5 : Generalized Linear Models

Lecturer: Eric P. Xing *Scribes: Uttara Ananthakrishnan, Mallory Nobles, Lujie(Karen) Chen*

1 Parameterizing Graphical Models

Let us consider the case of a Bayesian Network. We can represent the overall probability as

$$P(X) = \prod_{i=1:d} P(X_i|\mathbf{X}_{\pi_i})$$

We can consider all different combinations between A,B & C. This gives raise to three types of models - discrete, continuous or hybrid.Discrete model is represented as follows- Here A, B & C are discrete.Continuous model is similar where A, B & C are continuous and are drawn from normal distributions.Hybrid families can be of two types. In the first case, C is continuous while A & B are discrete. We can use $P(C|A,B) = f(\delta)\delta(A)\delta(B)$.Here $f(\delta)$ denotes the family of functions.In the second case, A & B are continuous while C is discrete. We can use logistic or linear regression to compute this.

1.1 Linear Regression

In Linear Regression problems the target variables and the inputs are related by the equation:

$$y_i = \theta^T x_i + \epsilon_i$$

Linear regression is a linear combination of inputs and noise. ϵ is an error term that denotes the unmodeled effects or random noise. ϵ is assumed to follow a Gaussian distribution $\mathcal{N}(0, \sigma)$. Therefore, we have

$$p(y_i \mid x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \theta^T x_i)}{2\sigma^2}\right)$$

1.2 Logistic Regression

This can be used with the conditional distribution is a Bernoulli distribution.

$$p(y \mid x) = \mu(x)^y (1 - \mu(x))^{1-y}$$

where μ is a logistic function and

$$\mu(x) = \frac{1}{1 + exp^{-\theta^T x}}$$

We can used the brute-force gradient method as in LR. However, this might not be easy.

Instead we can also generic laws by observing the p(y|x) is an exponential family function, more specifically, a generalized linear model. This ensures that we don't need to reinvent the algorithm to estimate the parameters.

1.3 Markov Random Fields

We can parameterize the other graphical models such as Markov random fields based on the node potentials

$$p(x) = \frac{1}{Z} \exp\left\{ -\sum_{c \in C} \phi_c(x_c) \right\} = \frac{1}{Z} \exp\{-H(x)\}$$

This can be rewritten as

$$p(X) = \frac{1}{Z} \exp\left\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \right\}$$

If X is continuous, it is taken as Multivariate Gaussian. We do this because if X is Gaussian, Z is constant and is known. If X is discrete, the model can be easily computed but Z is expensive to compute.

1.4 Restricted Boltzman Machines

The other model that can be parameterized are Restricted Boltzman Machines (RBM). RBM has hidden and visible units. Hidden units are the latent features while visible units are features that contain text, images etc. These features can be discrete, continuous or hybrid.

This can also be expressed in terms of the exponential family using the following form.

$$p(x, h|\theta) = exp\left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\theta) \right\}$$

1.5 Conditional Random Fields

The above idea can also be extended to conditional random fields. This allows us to design local features. The features are assumed to be inter-dependent. When labeling X_i future observations are taken into account.

2 Exponential Family

A distribution over a random variable **X** is in the exponential family if it can be expressed in the following form:

$$\begin{aligned} p(x \mid \eta) &= h(x) exp\{\eta^T T(x) - A(\eta)\} \\ &= \frac{1}{Z(\eta)} h(x) exp\{\eta^T T(x)\} \end{aligned}$$

Here η is the natural or canonical parameter. T(x) is the sufficient statistic. Function A(η) = log Z(η) is called the log normalizer.η and T interact as dot products, i.e, in a linear form. This form can be used to express many other distributions as discussed in the sections below.

2.1 Examples

2.1.1 Multivariate Gaussian Distribution

Consider a continuous vector random variable X $\in R^k$. We now have

$$\begin{aligned} p(x \mid \mu, \Sigma) &= \frac{1}{(2\pi)^{k/2} \mid \Sigma \mid^{1/2}} exp\left\{ \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\} \\ &= \frac{1}{(2\pi)^{k/2}} exp\left\{ \frac{1}{2} tr(\Sigma^{-1} x x^T) + \mu^T \Sigma^{-1}(x) - \frac{1}{2}\mu^T \Sigma^{-1}\mu - log \mid \Sigma \mid \right\} \end{aligned}$$

We can write the above as an exponential family distribution

$$\begin{aligned} \eta &= \left[\Sigma^{-1}\mu; -\frac{1}{2} vec(\Sigma^{-1}) \right] = \left[\eta_1, vec(\eta_2) \right], \\ \eta_1 &= \Sigma^{-1}\mu \text{ and } \eta_2 = -\frac{1}{2}\Sigma^{-1}. \end{aligned}$$

Here η_1 and η_2 are both natural parameters.

$$\begin{aligned} T(x) &= \left[x; vec(xx^T) \right] \\ A(\eta) &= \frac{1}{2}\mu^T \Sigma^{-1}\mu + log \mid \Sigma \mid = -\frac{1}{2} tr(\eta_2 \eta_1 \eta_1^T) - \frac{1}{2} log(-2\eta_2) \\ h(x) &= (2\pi)^{-k/2} \end{aligned}$$

Here, k-dimensional Gaussian is a (d+d^2)-parameter distribution with a (d+d^2) element vector of sufficient statistics. Because of symmetry and positivity,parameters are constrained and have lower degree of freedom.

2.1.2 Multinomial distribution

A random variable is multinomial distributed if x multi(x| π)

$$\begin{aligned} p(x \mid \pi) &= \pi_1^{x_1} \pi_2^{x_2} ... \pi_k^{x_k} = exp\left\{ \sum_k x_k ln \pi_k \right\} \\ &= exp\left\{ \sum_k^{k-1} x_k \ln \pi_k + \left(1 - \sum_k^{k-1} x_k \right) \ln \left(1 - \sum_k^{k-1} \pi_k \right) \right\} \\ &= exp\left\{ \sum_k^{k-1} x_k \ln \left(\frac{\pi_k}{1 - \sum_k^{k-1} \pi_k} \right) + \ln \left(1 - \sum_k^{k-1} \pi_k \right) \right\} \end{aligned}$$

We can write the above as an exponential family distribution

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

3 Exponential Family and properties

1. The exponential family has an interesting property called the moment generating property. This can be explained by taking the dth derivative of the log normalizer $A(\eta)$ and the result will be the d'th centered moment centered. This expands to the following results
   - The first derivative of the log normalizer function is the mean of T(X);
   - The second derivative of the log normalizer function is is its variance

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

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

2. The above result shows that the log normalizer function $A\eta$ is convex since the second derivative must always be positive,as the second derivative in this case is the variance which is always non-negative. Here η is the slope of A and η takes a particular value for $\mu = f(\eta)$ and $\eta = \psi(\eta)$
3. Therefore we can establish a 1:1 relationship between canonical parameters and the moment parameter. i.e, we can write the first derivative of the log normalizer function in form of the natural parameter or the canonical parameter. We can then equate this to mean, and then invert it to obtain the natural parameter in terms of the moment parameter. This is given as $\eta = \psi(\eta)$

4 MLE for the Exponential Family

For iid data, the log likelihood is

$$\ell(\eta; D) = \log \prod_n h(x_n) \exp\{\eta^T T(x_n) - A(\eta)\} = \sum_n \log\, h(x_n) + (\eta^T \sum_n T(x_n)) - NA(\eta)$$

The derivative is

$$\frac{\partial \ell}{\partial \eta} = \sum_n T(x_n) - N\frac{\partial A(\eta)}{\partial \eta}$$

When we set the derivative to zero, we see that

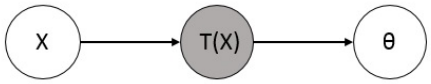
$$\frac{\partial A(\eta)}{\partial \eta} = \frac{1}{N}\sum_n T(x_n)$$

Since $\widehat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_n T(x_n)$, we can do moment matching and infer the canonical parameters using $\widehat{\eta}_{\text{MLE}} = \psi(\widehat{\mu}_{\text{MLE}})$.

It is worth noting that the MLE involves the data only through the sufficient statistic $T(X)$. In other words $T(X)$ contains all the essential information in X regarding θ. This is an informal definition of sufficiency.

4.1 Sufficiency

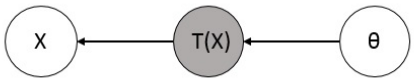
More formally, for a random variable X whose distribution depends on a parameter θ, a statistic $T(X)$ is sufficient for θ if there is no information in X regarding θ beyond that in $T(X)$.



(a)

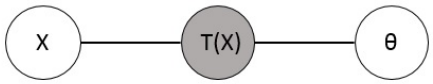
In the Bayesian approach, θ is an unknown random variable, and we say that $T(X)$ is sufficient for θ if $p(\theta|T(x), x) = p(\theta|T(x))$. The graphical representation is shown in Figure a.

In the frequentist view, θ is an unknown constant, and we say that $T(X)$ is sufficient for θ if $p(x|T(x), \theta) = p(x|T(x))$. The graphical representation is shown in Figure b.



(b)

The Neyman factorization theorem generalizes this relationship and states that $T(x)$ is sufficient for θ if $p(x, T(x), \theta) = \psi_1(T(x), \theta)\psi_2(x, T(x))$. The graphical representation is shown in Figure c.



(c)

Dividing by the distribution of θ, we see that $T(x)$ is sufficient for θ if $p(x|\theta) = g(T(x), \theta)h(x, T(x))$.

4.2 Examples: MLE for the Exponential Family

Recall that the multivariate Gaussian distribution is an exponential family distribution with $T(x) = [x; vec(xx^T)]$. Then $\widehat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_n T(x_n) = \frac{1}{N}\sum_n x_n$.

We also showed that the multinomial is an exponential family distribution. Here, the sufficient statistic $T(x)$ is $[x]$. Then $\widehat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_n T(x_n) = \frac{1}{N}\sum_n x_n$.

We know that the PMF of a Poisson random variable is $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\exp\{x\log\lambda - \lambda\}$. Then $\eta = \log x$, $T(x) = x$, $A(\eta) = \lambda$ and $h(x) = \frac{1}{x!}$. Then, $\widehat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_n T(x_n) = \frac{1}{N}\sum_n x_n$.

5 Bayesian View of Exponential Family Distributions

When taking the Bayesian approach, we assume that $p(x) \sim \exp\{\eta^T T(x) - A(\eta)\}$ and η is a random variable with prior distribution $p(\eta) \sim \exp\{\xi T'(\eta) + A'(\xi)\}$ where ξ is the hyper parameter. Bayesian inference involves inferring the posterior given the data and prior. We know that

$$p(\eta|x, \xi) \propto p(x|\eta)P(\eta|\xi) = \exp\{\eta^T T(x) + \xi^T T'(\eta) + A(\eta) + A(\xi)\}$$

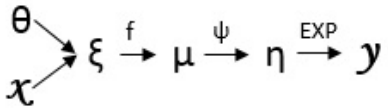
The exponential family formulation makes it easier to choose the prior so that you have conjugacy. We can make the prior and posterior take the same analytical form by assuming that $\eta = T'(\eta)$. Then,

$$p(\eta|x,\xi) \propto \exp\{(T(x)+\xi)T'(\eta) + A' + A\}.$$

6 Generalized Linear Models (GLIMs)

Linear regression and discriminative linear classification both address the problem of modeling the conditional relationship between a pair of random variables X and Y when both X and Y are observed. Both models take the form $E_p(Y) = \mu = f(\theta^T X)$ where $p()$ is the conditional distribution of Y and $f()$ is the response function. In linear regression, $f()$ is the identity function and $p()$ is Gaussian. For linear classification, $f()$ can take a variety of forms and $p()$ is Bernoulli in the binary case and multinomial for the multiway case. For logistic regression $f()$ takes the logistic function and $p()$ is binomial.

Generalized linear models (GLIMs) provide a framework for considering the more general case. GLIMs make three important assumptions. First, the observed input x is assumed to enter the model through a linear combination of its elements, $\xi = \theta^T x$. To model non-linear treatments of x, you can design features $f_1(x), \ldots, f_m(x)$ and treat these as additional input. Second, the conditional mean μ is assumed to be a function of ξ, $f(\xi)$. $f()$ is known as the response function. Third, the observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ. These assumptions are summarized in Figure d.



(d)

In the GLIM framework, we focus on distributions where $T()$ is the identity function. This implies that the observed Y is itself a sufficient statistic. We also include a scale parameter ϕ. Therefore, the conditional distribution of Y has the form

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

When choosing the exponential family you must consider the nature of the data Y. For example, when y is a class label, it is natural to use the Bernoulli or multinomial distribution, and when y are counts, it is natural to use the Poisson distribution.

The main modeling decision when specifying a GLIM is the choice of the response function. Because $f()$ represents a conditional expectation, a few mild constraints are imposed. For example, in the case of Bernoulli random variables, the response function's range should be $[0, 1]$. We also want to keep our choice of $f()$ as simple as possible for computational purposes since we will need to find derivatives of f. For each exponential family distribution, there is a particular response function called the canonical response function that has nice mathematical properties. The canonical response function is defined as

$$f = \psi^{-1}(\cdot)$$

Note that this is equivalent to stating that $\theta^T x = \xi = \eta$. Furthermore, since ψ is determined by the choice of the exponential family distribution, when the canonical response function is used, the GLIM is completely

specified by the choice of the exponential family distribution. Table e gives the canonical response function for several exponential family distributions.

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	$\mu = -\eta^{-1}$

(e)

7 Online Learning of GLIMs with canonical response

When using the canonical response function, $\eta = \phi(\mu)$, $\mu = f(\xi)$ and $\xi = \theta^T x$. Then the log-likelihood takes the form $\ell = \sum_n \log h(y_n) \ + \ \sum_n (\theta^T x_n y_n - A(\eta_n))$. The derivative of this is

$$\frac{\partial \ell}{\partial \theta} = \sum_n (x_n y_n - \frac{\partial A(\eta_n)}{\partial \eta_n} \frac{\partial \eta_n}{\partial \theta}) = \sum_n (y_n - \mu_n) x_n = X^T (y - \mu)$$

This gives us the stochastic gradient ascent update function

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

where $\mu_n^t = (\theta^t)^T x_n$ and ρ is the step size. Depending on the learning rate, this can be slow to converge. This method also requires careful tuning of the step size. Thus, in some cases, batch learning may be preferred.

8 Batch Learning of GLIMs

Note that the Hessian matrix $H = \frac{\partial^2 \ell}{\partial \theta \partial \theta^T} = -X^T W X$ where $X = [x_n^T]$ is the design matrix and $W = \text{diag}(\frac{\partial \mu_1}{\partial \eta_1}, \ldots, \frac{\partial \mu_N}{\partial \eta_N})$. W can be computed by taking the second derivative of $A(\eta_n)$.

Also recall that in least mean squares, the cost function in matrix form is $J(\theta) = \frac{1}{2}(X\theta - \overrightarrow{y})^T (X\theta - \overrightarrow{y})$. To minimize this, we take the derivative and set it to zero, which gives us the normal equations $\theta^* = (X^T X)^{-1} X^T \overrightarrow{y}$.

8.1 Iteratively Reweighted Least Squares(IRLS) - general case

We now derive the general case updating rule using Iternatively Reweigted Least Square method (IRLS) to find MLE for GLIMs with natural response.

Recall that the update rule with Newton-Raphson method with cost function J (in our case, the J is the log-likelihood) is

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

from slide 23, we have

$$\nabla_\theta J = \frac{dl}{d\theta} = X^T (y - \mu)$$

from slide 24, we have

$$H = -X^T W X$$

now we have

$$\begin{aligned}
\theta^{t+1} &= \theta^t - H^{-1}\nabla_\theta J \\
&= \theta^t + (X^T W^t X)^{-1} X^T (y - \mu^t) \\
&= (X^T W^t X)^{-1}(X^T W^t X)\theta^t + (X^T W^t X)^{-1} X^T (y - \mu^t) \\
&= (X^T W^t X)^{-1}[X^T W^t X \theta^t + X^T (y - \mu^t)] \\
&= (X^T W^t X)^{-1}[X^T W^t X \theta^t + X^T W^t (W^t)^{-1} (y - \mu^t)] \\
&= (X^T W^t X)^{-1} X^T W^t [X\theta^t + (W^t)^{-1}(y - \mu^t)] \\
&= (X^T W^t X)^{-1} X^T W^t z^t
\end{aligned}$$

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

Comparing to the normal equation for LMS $\theta^* = (X^T X)^{-1} X^T \overrightarrow{y}$, the update equation can be thought of solving the following "iteratively reweighted least squares" problem

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

8.2 IRLS Example 1: logistic regression

We now apply this general update equation to some specific problems, the first example is logistic regression, in which case
The conditional distribution is a Bernoulli

$$p(y|x) = \mu(x)^y (1 - \mu(x))^{1-y}$$

where μ is a logistic function $\mu(x) = \frac{1}{1+e^{-\eta(x)}}$ and $p(Y|X)$ is an exponential family function with mean $E[y|x] = \mu = \frac{1}{1+e^{-\eta(x)}}$ and canonical response function $\eta = \xi = \theta^T x$

so we have

$$\frac{d\mu}{d\eta} = \mu(1 - \mu), W = \begin{bmatrix} \mu_1(1-\mu_1) & & \\ & \ddots & \\ & & \mu_n(1-\mu_n) \end{bmatrix}$$

We may use regularized MLE by introducing a prior. In this setup

$$\begin{aligned}
&p(y = \{+1, -1\}|x, \theta) = \frac{1}{1 + \exp(-y\theta^T x)} = \sigma(y\theta^T x) \\
&p(\theta) \sim Normal(0, \lambda^{-1} I) \\
&l(\theta) = \sum_n log(\sigma(y_n \theta^T x_n)) - \frac{\lambda}{2}\theta^T \theta \\
&\nabla_\theta l = (1 - \sigma(y_n \theta^T x_n)) y_n x_n - \lambda\theta
\end{aligned}$$

8.3 IRLS Example 2: Linear regression

The conditional distribution is a Gaussian

$$p(y|x,\theta,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(y-\mu(x))^T\Sigma^{-1}(y-\mu(x))}$$
$$\Rightarrow h(x)exp\{-\frac{1}{2}\Sigma^{-1}(\eta^T(x)y - A(\eta))\} \text{ (rescale)}$$

where μ is a linear function

$$\mu(x) = \theta^T x = \eta(x)$$

$p(y|x)$ is an exponential family function with mean: $E(y|x) = \mu = \theta^T x$ and with canonical response function $\eta = \xi = \theta^T x$

The IRLS update rules:

$$\begin{aligned} \frac{du}{d\eta} &= 1 \\ W &= I \\ \theta^{t+1} &= (X^T W^t X)^{-1} X^T W^t z^t \\ &= (X^T X)^{-1} X^T (X\theta^t + (y - \mu^t)) \\ &= \theta^t + (X^T X)^{-1} X^T (y - \mu^t) \end{aligned}$$

when $t \to \infty$, $\theta = (X^T X)^{-1} X^T Y$, that is converge to normal equation solution.

9 How to define parameter prior

When defining parameter priors, should we use a single prior to cover all variables or should we assign separate priors for each node and each configuration of parent nodes, or can we group sets of variables which share the priors? The first option will introduce "coupling" which will defy the purpose of graphical modeling, the second option will not introduce "coupling" however will increase the complexity due to the possibly large number of priors.

The (Geiger and Heckerman 97,99) paper gives principles we may follow in selecting priors so that achieve a good trade off between the following extreme options above. The idea is that we need to achieve those two types of independence in assigning priors:

- Global Parameter Independence: the nodes from nuclear family, i.e. parent and child, should have separate priors;
- Local Parameter Independence: For the same child node, the different configuration of parents should have separate prior distribution, for example, in the direct graph where "Alarm" is the parent of "Call", then $P(\theta_{Call|Alarm=YES})$ should be defined independently of $P(\theta_{Call|Alarm=No})$

For discrete DAG models, Dircichlet prior satisfies our assumptions, for Gaussian DAG models, normal prior and Normal-Wishart prior satisfy our assumptions.

10 Summary

To summarize, when we parametrize graphical model, we should do it in a way so that it displays nice behavior and make the algorithm tractable. Some of the good choices are using exponential family distribution and GLIM model and choosing priors so that do they not introduce additional dependence. For a large fully observed directed graphical model, you can always find ways to decompose into local independent structures.