10-708: Probabilistic Graphical Models 10-708, Spring 2018

7 : Exponential Families: Learning the Parameters of a Fully Observed BN

Lecturer: Kayhan Batmanghelich *Scribes: Anusha Kamath*

1 Introduction

Graphical models can be represented and solved as exponential families. We can decompose a bayesian network into conditional distribution factors and marginals. The exponential family and Generalized linear models are used to model the marginals and conditionals respectively for parameter estimation.

2 The Exponential Family

The exponential family includes many familiar distributions like the Gaussian distribution, exponential distribution, Bernoulli, multinomial,Poisson and gamma distributions. It is the only family of distributions where the size of the sufficient stastics (more on this below) does not grow with the size of the data. The family has conjugate priors and is the distribution that is closest to uniform and hence the distribution that maximizes entropy.
A distribution for a numeric random variable X is an exponential family distribution if you can write it as

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

Basically we write the function of the density by seperating all the terms that only belong to X and call it base measure h(x). The only term that has interaction between the parameters and the function is show as a linear inner product. Here, η is the parameter and T is the sufficient statistic. A(η) is the normalizer.

3 Examples

3.1 The Exponential distribution

The density of the exponential distribution is given by

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0$$
$$= 0 \qquad otherwise$$

We can write this equation in the exponential family format as follows.

$$f(x) = e^{-\lambda x + log(\lambda)} I(x > 0)$$

$$\text{where} \quad h(x) = I(x > 0) \quad T(x) = x \quad A(\eta) = log(\frac{-1}{\eta}) \quad and \quad \eta = -\lambda$$

3.2 The Multinomial distribution

$$p(x \mid \pi) = \frac{M!}{x_1!x_2!\cdots x_K!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_K^{x_K},$$

$$p(x \mid \pi) = \frac{M!}{x_1!x_2!\cdots x_m!}\exp\left\{\sum_{k=1}^{K} x_k \log \pi_k\right\}.$$

This natural parameter space for the density function is Rk-1 because of the linear constraint on the components.

We can parametrize the distribution using the first K-1 components as below

$$\begin{aligned} p(x \mid \pi) &= \exp\left\{\sum_{k=1}^{K} x_k \log \pi_k\right\} \\ &= \exp\left\{\sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right)\log\left(1 - \sum_{k=1}^{K-1}\pi_k\right)\right\} \\ &= \exp\left\{\sum_{k=1}^{K-1}\log\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right)x_k + \log\left(1-\sum_{k=1}^{K-1}\pi_k\right)\right\}. \end{aligned}$$

where we have used the fact that $\pi_K = 1 - \sum_{k=1}^{K-1}\pi_k$.

From this representation we obtain:

$$\eta_k = \log\left(\frac{\pi_k}{1-\sum_{k=1}^{K-1}\pi_k}\right) = \log\left(\frac{\pi_k}{\pi_K}\right)$$

$$\pi_k = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}},$$

$$A(\eta) = -\log\left(1 - \sum_{k=1}^{K-1}\pi_k\right) = \log\left(\sum_{k=1}^{K} e^{\eta_k}\right)$$

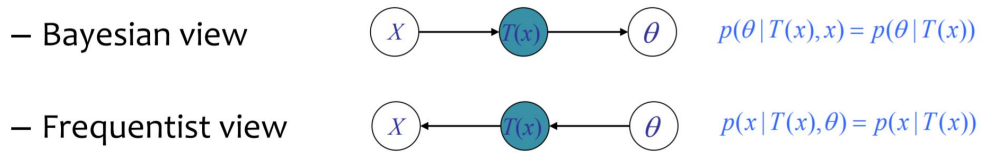
4 Moments and the partition function

1. The exponential family has the following property called the moment generating property : the *dth* derivative of the log partition equals the *dth* centered moment of the sufficient statistic.
   E.g., the first derivative of the log partition function is the mean of T(X); the 2nd is its variance.

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

5 Sufficient statistic: What makes it sufficient?

- The sufficient here indicates that T(X) contains all the essential information in X.
- Looking at this from the Bayesian point of view : Sufficiency means that theta is independent of X when we condition on T(X). If we want to sample from theta, we only need the sufficient statistic and we don't need the data itself.
- Looking at this from the Frequentist point of view : if we have the information about the sufficiency of X, we can draw the distribution without knowing the parameters. We define T(X) as sufficient for theta if the conditional distribution of X given T(X) is not a function of theta.
- - All the information in X is compressed into a low dimensional representation T(X) which does not grow with our data.



6 Estimating parameters with Maximum Likelihood

We obtain the MLE parameters by maximizing the probability of the data given the parameters. This amounts to Moment matching.

Consider an i.i.d. data set, $\mathcal{D} = (x_1, x_2, \ldots, x_N)$.

$$l(\eta \,|\, \mathcal{D}) = \log\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^T \left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta).$$

Taking the gradient with respect to η yields:

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta),$$

and setting to zero gives:

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{n=1}^{N} T(x_n).$$

Finally, defining $\mu := E[T(X)]$, and recalling Eq. (**??**), we obtain:

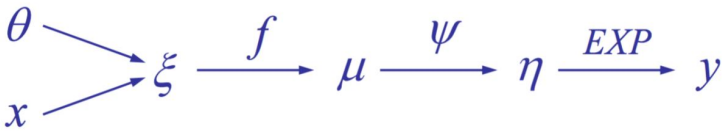
$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} T(x_n)$$

Because we are dealing with exponential families the log of the distribution opens up and we get an expression as in the example above. We get a term that is independent of parameters which we can get rid of when

maximizing with respect to the parameter. Then we have a term that is the inner product between the parameters and the sufficient statistic which does not grow with the size of the data The remaining factor is convex and the negative of that will be concave. Hence the overall function is $Concave{+}linear = Concave$and thus will have a unique maxima.

7 Generalized Linear Models

- The observed input X enters into the model as a linear combination of its elements
- The observed output y is characterized by an exponential family distribution with conditional mean μ
- We have the freedom of choosing ψ, the conditional distribution of Y, and choosing f, the response distribution
- The pipeline for the GLM is shown below



- We estimate the parameters θ on the X Y pairs wih Maximum Likelihood

8 Maximum Likelihood for GLMs with canonical response

When the link function makes the linear predictor η the same as the canonical parameter θ ,we say that we have a canonical link. The identity is the canonical link for the normal distribution. Some other pairs are shown below.

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

The Canonical response function $f = \psi^{-1}$, so that $\theta^T x$ directly corresponds to canonical parameter η.

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

For natural response the MLE can be obtained as below

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

$$\frac{\partial l}{\partial \theta} = \sum_n (x_n y_n - \frac{dA(\eta_n)}{\eta_n}\frac{d\eta_n}{d\theta}) = \sum_n (y_n - \mu_n)x_n$$

learning can be realized with Stochastic gradient descent as below

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

For batch learning in the case of cannonical GLMs we can use the Newton method.

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

The hessian matrix is derived below:

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

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 $A(\eta_n)$

9 Learning fully observed BNs

A bayesian network can be decomposed into conditionals and marginals. The conditionals can be modeled as GLMs and the marginals can be modeled with the exponential family. Thereby parameter learning is easy because of the decomposability of the bayesian network.

10 Summary

- Exponential family distributions are a candidate for marginal distributions P(Xi)
- MLE has closed form solution for the exponential family
- The benefit of distribution in exponential family is that the maximum likelihood estimation problem amounts to moment matching problem

- Generalized linear models are convenient for modeling conditionals $P(Xi|Yi)$
- MLE for GLIMs with canonical response can be solved by SGD
- The general algorithm used is Iteratively reweighted least squares
- Parameter learning in Fully observed BNs can be done with these distributions as the BN factors decompose into marginals and conditionals