CS281A/Stat241A: Statistical Learning Theory

The Exponential Family and Generalized Linear Models (10/7/04)

Lecturer: Michael I. Jordan *Scribes: Shariq Rizvi and Xia Jiang*

1 The Exponential Family of Distributions

(A large part of this lecture reviewed material that was covered in the previous one)

1.1 The Basics

The exponential family covers a large number (and well-known classes) of distributions:

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

Here x and T(x) are vectors (in general, of different dimensions)
η is the *canonical parameter*, a vector of parameters
A(η) is the *cumulative generating function*
h(x) is an arbitrary function of x (not a core part), called the *base measure*
A(η) is equal to $\log \int \exp\{\eta^T T(x)\} h(x) dx$.

When parameter θ enters exponential family as $\eta(\theta)$, we write a probability density in the form of the exponential family as

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

where $\eta(\theta)$ is the *canonical parameter or natural parameter*, θ is the parameter vector of some distribution that can be written in the form of the exponential family.

1.2 A Very Rich Formalism

As we saw in the previous lecture, several well-known distributions (such as Gaussian, Multinomial and Bernoulli) lie in the exponential family. In fact, any joint probability distribution on discrete random variables lies in the exponential family. We show this now.

Consider an arbitrary undirected graphical model on n random variables:

$$p(x \mid \Psi) = \frac{1}{Z(\Psi)} \prod_{C \in \mathcal{C}} \Psi_C(x_C)$$

Where the symbols have their usual meaning. Clearly, $Z(\Psi) \neq 0$ for any valid distribution. For x such that $\Psi(x) > 0$, we can write this as:

$$p(x \mid \Psi) = \exp\left(\sum_{C \in \mathcal{C}} \log(\Psi_C(x_C)) - \log Z(\Psi)\right)$$

Note that Ψ is not really a parameter that can be used directly to express this as an exponential family distribution. Instead, Ψ is an arbitrary function vector (satisfying some constraints) and not an arbitrary parameter vector. However, we now show that for any given Ψ, under the assumption of the random variables being discrete, this can be made to look like an exponential family distribution.

Let's look at a specific $C \in \mathcal{C}$. Then, $\Psi_C(X_C)$ is a function of the k random variables $\{X_1, X_2, ..., X_k\}$ that are a part of the maximal clique C. As these random variables are discrete, the function can be seen as a table of function values, where the values of the k random variables are indexes for the rows of this table. For e.g.

X_1	X_2	...	X_k	Ψ
v_1^1	v_2^1	...	v_k^1	0.08
v_1^2	v_2^1	...	v_k^1	0.11
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Here, v_i^j are the different values (as we range across j) that the random variable X_i takes. Given that the random variables are discrete, this table will be of finite size. Given such a table, we can write $\Psi_C(x_C)$ as a sum of products of delta functions:

$$\Psi_c(x_c) = \prod_{v_1^{i_1},..,v_k^{i_k}} \Psi_c(v_1^{i_1}, v_2^{i_2}, ..., v_k^{i_k})^{\delta(x_1,v_1^{i_1})\delta(x_2,v_2^{i_2})....\delta(x_k,v_k^{i_k})}$$

Where $\delta(x_i, v_i^j) = 1$ if $x_i = v_i^j$ and 0 otherwise. The summation is over all possible values of $v_1^{i_1}, .., v_k^{i_k}$

Plugging this form of Ψ_c into the earlier equation:

$$p(x \mid \Psi) = \exp\left\{\sum_{C \in \mathcal{C}}\left(\sum_{v_1^{i_1},..,v_k^{i_k}} (\delta(x_1, v_1^{i_1})\delta(x_2, v_2^{i_2})....\delta(x_k, v_k^{i_k}) \log \Psi_C(v_1^{i_1}, v_2^{i_2}, ..., v_k^{i_k}))\right) - \log Z(\Psi)\right\}$$

Which is the same as:

$$p(x \mid \Psi) = \exp\left\{\left(\sum_{v_1^{i_1},..,v_n^{i_n}} \sum_{C \in \mathcal{C}} (\delta(x_1, v_1^{i_1})\delta(x_2, v_2^{i_2})....\delta(x_k, v_k^{i_k}) \log \Psi_C(v_1^{i_1}, v_2^{i_2}, ..., v_k^{i_k}))\right) - \log Z(\Psi)\right\}$$

Now this representation of the joint probability density of the n random variables is easily seen to be in the exponential family. The parameter vector η consists of components that are the $\log \Psi_c(v_1^{i_1}, v_2^{i_2}, ..., v_k^{i_k})$ terms and the sufficient statistic function T has components that are products of the delta functions above.

1.3 Some Properties of the Exponential Family

The cumulative generative function $A(\eta)$ has some very important properties:

$$\bigtriangledown_\eta A(\eta) = E(T(X))$$

That is, the first derivative of the function $A(\eta)$ at the particular value of η is the same as the expected value of $T(X)$ for that given η. This means that we can find this expected value (which normally needs an integration) by differentiating the function $A(\eta)$.

Similarly,

$$\bigtriangledown_\eta^2 A(\eta) = Var(T(X))$$

Where the right-hand side is the covariance matrix of T(X). As a specific case, consider the Poisson distribution:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Which can be written explicitly in the exponential family form:

$$p(x) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$

Where $\eta = \log \lambda$, or $\lambda = \exp\{\eta\}$. So, $A(\eta) = \exp\{\eta\}$. Taking the first and second derivatives of this function and noting that here $T(x) = x$:

$$E(X) = \bigtriangledown_\eta e^\eta = e^\eta = \lambda$$

$$Var(X) = \bigtriangledown_\eta^2 e^\eta = e^\eta = \lambda$$

As expected, the mean and variance of the Poisson distribution turn out to be the parameter λ.

1.4 Sufficiency

We will show that T(x) is the sufficient statistic for a parameter θ and briefly discuss why the sufficient statistic of a distribution is a useful quantity.

$$p(x|\theta) = \Psi_1(T(x), \theta)\Psi_2(x, T(x)) \tag{1}$$

Eqn 1 is a sufficient and necessary condition for $T(x)$ to be a sufficient statistic for θ, by the Neymann factorization theorem. The exponential family of probability distributions can be factorized as in Eqn 1, with $\Psi_1(T(x), \theta) = \exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}$ and $\Psi_2(x, T(x)) = h(x)$. This proves that T(x) is a sufficient statistic for θ. The sufficient statistic is a useful quantity for inference on θ because it allows for data-free inference – we can throw the data x away when doing inference on θ if we keep just the sufficient statistic. A good example is in the case of IID sampling.

Let us consider an IID sampling of a random variable $\{x_1, ..., x_n\}$ that has a density function that falls in the exponential family. Then, the joint probability distribution looks like:

$$p(x_1, x_2, ..., x_n \mid \eta) = \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left\{\eta^T \sum_{i=1}^{n} T(x_i) - nA(\eta)\right\}$$

This joint probability of the n samples of the random variable is again in the exponential family form. Also, the canonical parameter η is the same as in the initial distribution.

This expression also shows us that $\sum_{i=1}^{n}$ T(x_i) is the sufficient statistic function for the joint probability distribution. For purposes of reasoning about η, we need to retain only the values of $\sum_{i=1}^{n}$ T(x_i) and can throw away the raw x_i values. Note that this will be helpful if n is greater than the dimension of $T(x)$ (which is almost always the case).

Let's find the maximum likelihood estimator of η given the data from n IID samplings. We have

$$\underset{\eta}{argmax}(p(x_1, x_2, ..., x_n \mid \eta)) = \text{ value of } \eta \text{ where } \bigtriangledown_\eta \, p(x_1, x_2, ..., x_n \mid \eta) = 0$$

This happens when:

$$\sum_{i=1}^{n} T(x_i) - n \bigtriangledown_\eta A(\eta) = 0$$

Or

$$\bigtriangledown_\eta A(\eta) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$$

Which means that E(T(X)) = $\frac{1}{n}\sum_{i=1}^{n}$ T(x_i). This implies that for all exponential family distributions, the expected value of the sufficient statistics can be computed directly from the data, without having to estimate η. For many distributions, $T(x) = x$, so this equation can be used to calculate the expected value of the random variable itself (in those cases).

1.5 Convexity of $\bigtriangledown_\eta A(\eta)$

Given that the second derivative of $A(\eta)$ is a variance, it is always positive. This means that $\bigtriangledown_\eta A(\eta)$ is a convex function (example Figure 1).

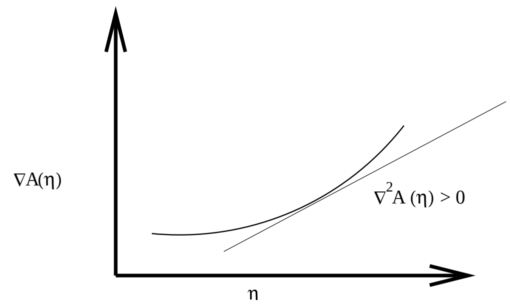

Figure 1: An alternate parameterization

This observation implies that $\bigtriangledown_\eta A(\eta)$ is a one-to-one function of η. In other words, we can consider $\bigtriangledown_\eta A(\eta)$ or $E(T(X))$ as the parameter of an exponential distribution. In cases where $T(X) = x$, this means that the expected value of the random variable (the mean) can be used as a parameter.

1.6 Conjugacy of the Exponential Family

Consider the directed graphical model given in Figure 2:

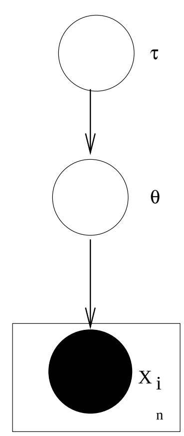

Figure 2: Conjugacy of the Exponential Family

We have already derived the conjugate prior distribution for the Gaussian, Multinomial and Dirichlet, but here we derive it in general, for any distribution in the exponential family. Consider the graphical model in Figure 2. In this model, θ is the set of parameters of the likelihood, τ is a set of hyperparameters and the nodes x_n in the plate are the N replicates IID variables having the same distribution, which we assume is in the exponential family.

We write the prior distribution as p($\theta|\tau$) instead of just p(θ), where τ is the hyperparameter. For p($\theta|\tau$) to be a conjugate prior, it needs to be in form that, when multiplied by the likelihood p($x|\theta$), yields a posterior distribution that is also in the exponential family, that is, we seek a p($\theta|\tau$) such that

$$p(\theta|\tau)p(x|\theta)$$

is in the exponential family (since the posterior density p($\theta|x$) is proportional to the above quantity).

We first write the likelihood, p($x|\theta$), for the graphical model in Figure 2 of N IID variables whose likelihood is in the exponential family:

$$p(x \mid \theta) = (\prod_{n=1}^{N} h(x_n)) \exp\{\eta^T(\theta) \sum_{n=1}^{N} T(x_n) - NA(\eta(\theta))\} \tag{2}$$

To be in the form of the exponential family, the prior p($\theta|\tau$) must be an exponential of two terms, one multiplying $\eta(\theta)$ and another multiplying A($\eta(\theta)$). We let $\tau = \{\tau_1, ..., \tau_k\}$ be the set of parameters multiplying $\eta(\theta)$ and single parameter τ_0 to multiply $A(\eta(\theta))$, obtaining:

$$p(\theta \mid \tau) \propto \exp\left\{\tau^T \eta(\theta) - \tau_0 A(\eta(\theta)\right\}$$

To make this a probability distribution, we divide by normalizing factor Z(τ,τ_0), which is the integral of the above expression with respect to θ:

$$Z(\tau, \tau_0) = \int \exp\{\tau^T \eta(\theta) - \tau_0 A(\eta(\theta))\} d\theta$$

Therefore, the conjugate prior is

$$p(\theta \mid \tau) = \frac{1}{Z(\tau, \tau_0)} \exp\{\tau^T \eta(\theta) - \tau_0 A(\eta(\theta))\} \tag{3}$$

To verify that this is indeed a conjugate prior, we multiply it by the likelihood (Eqn 2), obtaining the posterior.

$$p(\theta \mid x) \propto p(x \mid \theta) p(\theta \mid \tau) = \exp\left\{ \left(\tau + \sum_{n}^{N} T(x_n) \right)^T \eta(\theta) - (\tau_0 + N) A(\eta(\theta)) \right\}$$

The conjugate prior (Eqn 3) is therefore in the exponential family. Note that to compute the posterior from the likelihood and the conjugate prior, we simply add the sum of sufficient statistics to the original parameters of the conjugate prior. We can use the exponential family conjugate prior for Bayesian prediction.

1.7 Prediction

Consider the directed graphical model given in Figure 3:

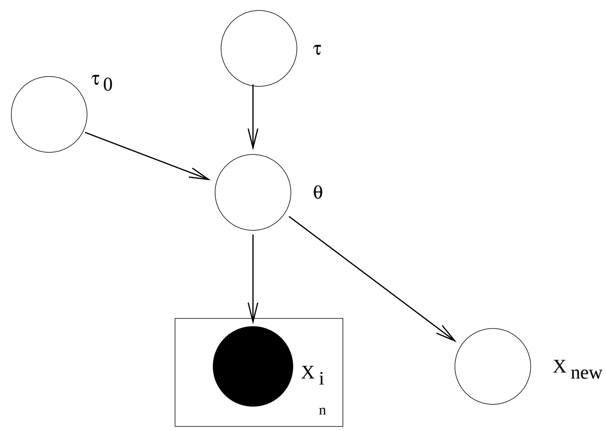

Figure 3: Prediction

We want to predict x_{new} based on the observations of x_i's. We have:

$$p(x_{new} \mid x_1, ..., x_n, \tau, \tau_0) = \int p(x_{new} \mid \theta, \tau, \tau_0) p(\theta \mid x_1, ..., x_n, \tau, \tau_0) d\theta$$

Given that both the conditional probability distributions inside the integral are of the exponential family form, this can be shown to be equal to a ratio of normalization factors. This will be a homework question.

1.8 KL-Divergence and the Exponential Family

Let the distribution of X be in the exponential family. Suppose we are trying to infer η by IID sampling. Let $\widetilde{p}(X)$ be defined as the empirical distribution:

$$\widetilde{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x, x_i)$$

This function essentially puts a point mass at each of the observed values for the random variable X. Note that it also defines a valid probability density function. Let's calculate the Kullback-Leibler (KL) divergence between this function and the original distribution (assumed discrete):

$$\begin{aligned} D(\widetilde{p}(X) \mid\mid p(x)) &= \sum_x \widetilde{p}(x) \log \frac{\widetilde{p}(x)}{p(x)} \\ &= \sum_x \widetilde{p}(x) \log \widetilde{p}(x) - \sum_x \widetilde{p}(x) \log p(x) \\ &= \sum_x \widetilde{p}(x) \log \widetilde{p}(x) - \sum_x \left(\frac{1}{n} \sum_{i=1}^{n} \delta(x, x_i) \right) \log p(x) \end{aligned}$$

$$
\begin{aligned}
&= \sum_x \widetilde{p}(x)\log\widetilde{p}(x) - \sum_x \left(\frac{1}{n}\sum_{i=1}^{n}\delta(x,x_i)\log p(x)\right) \\
&= \sum_x \widetilde{p}(x)\log\widetilde{p}(x) - \frac{1}{n}\sum_{i=1}^{n}\sum_x \delta(x,x_i)\log p(x)) \\
&= \sum_x \widetilde{p}(x)\log\widetilde{p}(x) - \frac{1}{n}\sum_{i=1}^{n}\log p(x_i)
\end{aligned}
$$

The term $\sum_{i=1}^{n}\log p(x_i)$ is the log likelihood function for the IID sampling with n random variables. Hence, we can write the above equation as:

$$
D(\widetilde{p}(x) \mid\mid p(x \mid \eta)) = \sum_x \widetilde{p}(x)\log\widetilde{p}(x) - \frac{1}{n}l(x;\eta)
$$

This implies that the η that maximizes the log likelihood will be the same one that minimizes the KL-divergence on the left.

$$
\underset{\eta}{argmin}\, D(\widetilde{p}(x) \mid\mid p(x \mid \eta)) = \underset{\eta}{argmax}\, l(x;\eta)
$$

2 Generalized Linear Models (GLIMs)

In order to study the relationship between X and Y, we introduce *generalized linear model (GLIM)* in this section.
As we discussed in previous chapters, for linear regression, we choose a particular conditional expectation of Y. Letting μ denote the modeled value of conditional expectation,

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

For linear regression, this distribution is a Gaussian distribution. GLIM extends these ideas beyond the Gaussian, Bernoulli and multinomial setting to the more general exponential family.

As in Figure 4, X enters model linearly as $\theta^T X$; f is called a *response function*; while Ψ is a one-to-one function mapping μ to η. There are two principal choice points in the specification of a GLIM: (1) the choice of exponential family distribution, and (2) the choice of the response function $f(\cdot)$. While the choosing exponential family distribution is generally rather strongly constrained by the nature of the data Y, we focus on the selection of response function. Intuitively, the response function needs to be both monotonicity and differentiable. However, a particular response function — the *canonical response function* — is uniquely associated with a given exponential family distribution and has some appealing mathematical properties. Canonical response function:

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

or

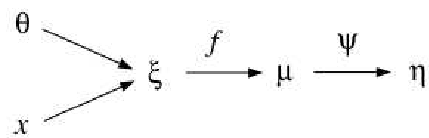

Figure 4: A diagram summarizing the relationships between the variables in a GLIM model

$$\xi = \eta$$

If we decide to use the canonical response function the choice of the exponential family density completely determines the GLIM. The maximum likelihood estimation in the text book shows an example of GLIM. To estimate parameters in this model, one can use, e.g., *iteratively reweighted least squares (IRLS) algorithm*.

An on-line algorithm

A general on-line estimation algorithm can be obtained by following the stochastic gradient of the log likelihood function.

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

Where $\mu_n^{(t)} = f(\theta^{(t)T} x_n)$ and where ρ is a step size. If we do not use the canonical response function, then the gradient also includes the derivatives of $f(\cdot)$ and $\psi(\cdot)$. These can be viewed as scaling coefficients that alter the step size ρ, but otherwise leave the general LMS form intact. The LMS algorithm is the generic stochastic gradient algorithm for models throughout the GLIM family. We will get a better understanding of the GLIM from Homework 3.