CS281A/Stat241A: Statistical Learning Theory

The Exponential Family (cont'd) and Generalized Linear Models

Lecturer: Michael I. Jordan *Scribes: Doron Tal*

Overview

There are several good reasons for writing our densities (when possible) in the form of the exponential family. Often, certain analyses of a distribution become easier if it is in the form of the exponential family. These include deriving moments, the sufficient statistic and conjugate priors. The first two were discussed in the last lecture and are only briefly reviewed here. The last one, deriving a conjugate prior for Bayesian estimation, is discussed in more detail below.

A note on notation: in this discussion vectors and vector-valued functions are written in boldface type while scalars and scalar functions are not.

The exponential family

In this discussion we write the general form of the exponential family in a slightly different manner. In the previous lecture we parametrized the exponential family by its natural parameter $\boldsymbol{\eta}$, but here we parametrize it by $\boldsymbol{\theta}$, substituting $\boldsymbol{\eta}(\boldsymbol{\theta})$ for $\boldsymbol{\eta}$. $\boldsymbol{\theta}$ is the parameter vector of some distribution that can be written in the form of the exponential family. Thus, we write a probability density in the form of the exponential family as

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(x) \exp\left\{ \boldsymbol{\eta}(\boldsymbol{\theta})^T T(x) - A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\}, \tag{1}$$

where $\boldsymbol{\eta}(\boldsymbol{\theta})$ is the *canonical parameter* or *natural parameter*, $h(x)$ is a positive function of the data called the *reference density*, $T(x)$ is a vector-valued function of the data and is a sufficient statistic for $\boldsymbol{\theta}$ and $A(\boldsymbol{\eta}(\boldsymbol{\theta}))$ is the log normalization factor of the distribution.

Moment generating properties of $A(\boldsymbol{\eta})$

One benefit of the exponential family representation is that we can obtain moments of the distribution by taking the derivatives of the log normalization function $A(\boldsymbol{\eta})$. Last lecture we showed that the ith derivative of $A(\boldsymbol{\eta})$ with respect to η is the ith "raw moment" (i.e. moment centered at zero) of $T(x)$. The textbook method of defining and computing moments involves integration, but the integral may not have an analytical solution. If our density is in the form of the exponential family, it is easy to obtain analytical solutions for the moments by differentiating $A(\boldsymbol{\eta})$.

A related benefit is that a distribution in the exponential family can be parametrized not only by $\boldsymbol{\eta}$ or $\boldsymbol{\eta}(\boldsymbol{\theta})$, but also by its mean $\boldsymbol{\mu}$. This parametrization is useful for the discussion of generalized linear models below. Here, we briefly discuss why a $\boldsymbol{\mu}$ parametrization is possible.

To parametrize the distribution by its mean $\boldsymbol{\mu}$, we need to find an expression for η in terms of $\boldsymbol{\mu}$ and substitute this function of $\boldsymbol{\mu}$ into Eqn. 1. We already showed that $\boldsymbol{\mu}$ can be written as a function of η, the derivative of $A(\eta)$), but we need to show that the mapping $\eta \rightarrow \boldsymbol{\mu}$ is invertible.

If a function is convex, the mapping between its argument and its first derivative with respect to its argument is a bijection, thus the mapping is invertible. Last time we showed that $A(\eta)$ is a convex function (because its second derivative, being a variance, is always positive). This implies that the relationship

$$\boldsymbol{\mu} = \frac{dA}{d\eta} \tag{2}$$

is invertible and therefore we can parametrize a distribution in the exponential family in terms of its mean $\boldsymbol{\mu}$. It can be shown that the set The set $\mathcal{S}$ over which $A(\eta)$ is a convex function is the convex set [1]

$$\left\{ \eta : \int p(x|\eta) d\eta < \infty \right\}. \tag{3}$$

Sufficient statistics

Another benefit of writing a density in the form of the exponential family is that $T(x)$, the sufficient statistic for θ, can be read off by inspection from the exponential family form. We now show that $T(x)$ is the sufficient statistic for a parameter θ in Eqn. 1 and then briefly discuss why the sufficient statistic of a distribution is a useful quantity.

In the last lecture we defined the term 'sufficient statistic' from both Bayesian and frequentist points of view using directed graphical models for each point of view and then showed that the undirected graphical model is equivalent to either of the two graphical models. This graphical model implies a factorization

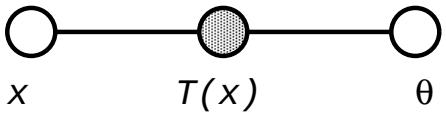

Figure 1: Graphical model demonstrating the relationship between the parameters θ, data x and the sufficient statistic $T(x)$.

$$p(x|\theta) = \psi_1(T(x), \theta)\psi_2(x, T(x)), \tag{4}$$

which 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. 4, with $\psi_1(T(x), \theta) = \eta(\theta)^T T(x) - A(\boldsymbol{\eta}(\boldsymbol{\theta}))$ and $\psi_2(x, T(x), \theta) = h(x)$. 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 have the sufficient statistic. A particularly good example of this benefit is the case of IID sampling.

[1] Some alternative definitions (from http://mathworld.wolfram.com): A set in n-dimensional space is *convex* if any line between a pair of its points lies entirely in the set. A *convex function* is one whose value at the midpoint of any interval in its domain does not exceed the average of its values at the ends of the interval. In other words, a function $f(x)$ is convex on an interval $[a, b]$ if for any two points x_1 and x_2 in $[a, b]$, $f[\frac{1}{2}(x_1 + x_2)] \leq \frac{1}{2}[f(x_1) + f(x_2)]$.

IID sampling

Consider replicates $\mathbf{x} = X_1, \ldots, X_N$ of IID random variables (data), all characterized by the same density which is in the exponential family. The joint density (also the likelihood) of the set is

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{n=1}^{N} p(X_n|\theta) \\ &= \prod_{n=1}^{N} h(X_n) \exp\left\{\eta(\theta)^T T(X_n) - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\right\} \\ &= \left(\prod_{n=1}^{N} h(X_n)\right) \exp\left\{\eta(\theta)^T \sum_{n=1}^{N} T(X_n) - NA(\boldsymbol{\eta}(\boldsymbol{\theta}))\right\}. \end{aligned} \tag{5}$$

The joint density of $\mathbf{x}$ is thus also in the exponential family, with the sufficient statistic

$$\sum_{n=1}^{N} T(X_n), \tag{6}$$

i.e. the sufficient statistic of $\mathbf{x}$ for θ is the sum of sufficient statistics for θ of the component variables in the set $\mathbf{x}$.

For many of the densities that can be put in the exponential family form (e.g. the Gaussian or the Poisson), $T(X)$ is just X. When we use such a density for a set IID random variables, all we need for running inference algorithms on θ is to keep a running sum of the variables X_n. Each variable X_n need only be encountered once, incorporated into a running sum, and then discarded (this reduction in complexity is one reason why the database community has recently become interested in sufficient statistics).

Maximum likelihood estimate

We have already derived the maximum likelihood (ML) estimate for a couple of instances of the exponential family – the Gaussian and the multinomial. In this section we derive the ML estimate for the exponential family in general.

Another benefit of knowing the sufficient statistic is that it eases computing the maximum-likelihood estimate. We show this for the above example – a collection of IID random variables that all have the same density, in the exponential family.

ML estiamte of the mean $\hat{\boldsymbol{\mu}}_{ML}$

To compute $\hat{\boldsymbol{\mu}}_{ML}$, we first write the log-likelihood

$$\begin{aligned} l(\eta; \mathbf{x}) &= \log h(\mathbf{x}) \exp\left\{\eta^T \sum_{n=1}^{N} T(X_n) - NA(\eta)\right\} \\ &= \log h(\mathbf{x}) + \eta^T \sum_{n=1}^{N} T(X_n) - NA(\eta) \end{aligned} \tag{7}$$

Next, we take the derivative of the log-likelihood with respect to the parameter η. The term with $h(\mathbf{x})$ becomes zero because it's a constant factor, since it isn't a function of η:

$$\begin{aligned} \bigtriangledown_\eta l(\eta; \mathbf{x}) &= \sum_{n=1}^{N} T(X_n) - N \bigtriangledown_\eta A(\boldsymbol{\eta}(\boldsymbol{\theta})) \\ &= \sum_{n=1}^{N} T(X_n) - NE(T(X_n)). \end{aligned} \tag{8}$$

Setting the derivative of the log-likelihood to zero and defining $\boldsymbol{\mu} = E(T(X_n))$ we get

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N} \sum_{n=1}^{N} T(X_n). \tag{9}$$

The ML estimate of the mean is therefore a sample mean of the sufficient statistics of the individual random variables in the IID set.

In general, the ML estimate of the mean is a sample mean of some quantity. In the case of the exponential family, it is the sample mean of the sufficient statistic. For a Gaussian, for example, the sufficient statistic is X, so the ML estimate is just the sample mean of the data. The above ML estimate of the mean of a distribution is another quantity made easy to compute by the exponential family when dealing with IID data, because one only has to read off the sufficient statistic and divide by N.

ML estimate of $\hat{\theta}_{ML}$

What about $\hat{\theta}_{ML}$? In general, we can reparametrize a distribution as long as the transformation between the new parameter and the old parameter is invertible, i.e. as long as there is a one-to-one relationship between the two parameters. We already proved that the exponential family can be reparametrized by $\boldsymbol{\mu}$, by showing that $\boldsymbol{\mu}$ is in one-to-one correspondence with η due to the convexity of $A(\eta)$. Since θ is a parameter of the original density that we are writing in the exponential family form, by definition, it is in one-to-one correspondence with η.

In many distributions that we can write in the form of the exponential family, the parameter of the distribution θ is $\boldsymbol{\mu}$. When θ is not $\boldsymbol{\mu}$, we can compute $\hat{\theta}_{ML}$ simply by solving for it in terms of $\hat{\boldsymbol{\mu}}_{ML}$. We can do this because both θ and $\boldsymbol{\mu}$ are *parameters* and thus there is a one-to-one relationship between them that allows us to solve for one in terms of the other. Maximum likelihood estimation, therefore, is invariant to one-to-one transformations of the parameter.

Maximum likelihood and Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is a measure of distance or discrepancy between two probability distributions p and $\tilde{p}$, defined as

$$\begin{aligned} \mathcal{D}(\tilde{p}(x) \| p(x)) &\triangleq \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x)} \\ &= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x), \end{aligned} \tag{10}$$

where the first term is called *negative entropy* and the second *cross entropy*. An important fact about the KL divergence is that it is zero if and only if $\tilde{p}(x) = p(x)$.

The KL divergence is related to *minimum length code*. Minimum length code is only briefly mentioned here but will be treated later on: given a set of variables each of whose probability density is $\tilde{p}(x)$, we wish to transmit over a channel using a code that is built on a different probability density, $p(x)$. The KL divergence would tell us the "extra amount we pay" when we use p instead of $\tilde{p}$ to build our code.

Relationship of KL divergence to likelihood

To relate the KL divergence to the likelihood, we begin by calling $\tilde{p}(x)$ the *empirical distribution*, defined as

$$\tilde{p}(x) \stackrel{\triangle}{=} \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n), \tag{11}$$

recalling that $\delta(x, x_n)$ is the Kronecker delta function (for continuous x values) or, in the discrete case, a function that is 1 when its arguments are equal and 0 otherwise. The graph of the empirical distrubution for generic data x_n is given in Figure 2. Whenever we make the empirical distribution multiply some function

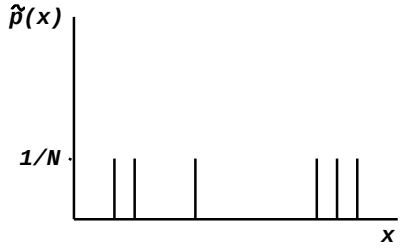

Figure 2: A plot of some empirical distribution. The empirical distribution $\tilde{p}(x)$ places a point mass at each data point x_n.

of x inside an integral (or sum), the integral is nonzero only at the data points x_n. Using the empirical distribution, therefore, we can write the log likelihood (for the discrete case) as

$$\begin{aligned}
\sum_x \tilde{p}(x) \log p(x|\theta) &= \sum_x \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n) \log p(x|\theta) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_x \delta(x, x_n) \log p(x|\theta) \\
&= \frac{1}{N} \sum_{n=1}^{N} \log p(x_n|\theta) \\
&= \frac{1}{N} l(\theta|\mathcal{D}).
\end{aligned} \tag{12}$$

The above series of equations just shows how to write the log-likelihood of any distribution using the empirical distribution. The log-likelihood (scaled by N) is thus shown equal to the cross-entropy between the empirical distribution and the model $p(x|\theta)$.

We now calculate the KL divergence between the empirical distribution and the model $p(x|\theta)$:

$$\begin{aligned} \mathcal{D}(\tilde{p}(x))\|p(x) &= \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x|\theta)} \\ &= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x|\theta) \\ &= \sum_x \tilde{p}(x) \log \tilde{p}(x) - \frac{1}{N} l(\theta|\mathcal{D}). \end{aligned} \tag{13}$$

Since the term $\sum_x \tilde{p}(x) \log \tilde{p}(x)$ is not a function of θ, minimizing θ on the LHS is equal to maximizing θ on the RHS. We have just shown that *minimizing the KL divergence between the empirical distribution and the model distribution is equivalent to maximizing the likelihood of the model distribution.* Another way of saying this is that the KL divergence between the emprirical distribution and the model is equal to the maximum likelihood, up to a gain and an offset.

Finally, we note that the KL divergence, though not a metric, has distance-like properties that define a geometry, e.g., we will later prove a Pythagorean theorem for expressions involving the KL divergence.

Conjugacy of the exponential family and Bayesian estimates

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 3. In this model, $\boldsymbol{\theta}$ is the set of parameters of the likelihood, $\boldsymbol{\tau}$ is a set of hyperparameters, and the nodes X_n in the plate are N replicates IID variables having the same distribution, which we assume is in the exponential family.

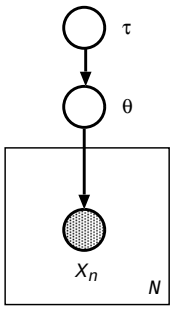

Figure 3: Graphical model of IID multivariate exponential family Bayesian estimation.

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

$$p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\tau}) \tag{14}$$

is in the exponential family (since the the posterior density $p(\boldsymbol{\theta}|\mathbf{x})$ is proportional to the above quantity, by Bayes rule).

We first write the likelihood, $p(\mathbf{x}|\boldsymbol{\theta})$, for the graphical model in Figure 3 of N IID variables whose likelihood

is in the exponential family:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{n=1}^{N} h(x_n) \exp\left\{ \boldsymbol{\eta}(\boldsymbol{\theta})^T \sum_{n=1}^{N} T(x_n) - NA(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\} \tag{15}$$

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

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) \propto \exp\left\{ \boldsymbol{\tau}^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \tau_0 A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\}. \tag{16}$$

To make this a probability distribution, we divide by a normalizing factor $Z(\boldsymbol{\tau}, \tau_0)$, which is the integral of the above expression

$$Z(\boldsymbol{\tau}, \tau_0) = \int \exp\left\{ \boldsymbol{\tau}^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \tau_0 A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\} \mathrm{d}\theta. \tag{17}$$

Therefore, the conjugate prior is

$$p(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{1}{Z(\boldsymbol{\tau}, \tau_0)} \exp\left\{ \boldsymbol{\tau}^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \tau_0 A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\}. \tag{18}$$

To verify that this is indeed a conjugate prior, we multiply it by the likelihood (Eqn. 15), obtaining the posterior

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\tau}) \\ &= \exp\left\{ (\boldsymbol{\tau} + \mathbf{T})^T \boldsymbol{\eta}(\boldsymbol{\theta}) - (\tau_0 + N) A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\}, \end{aligned} \tag{19}$$

where

$$\mathbf{T} = \sum_{n=1}^{N} T(x_n). \tag{20}$$

The conjugate prior (Eqn. 18) is therefore in the exponential family. Notice that to compute the posterior from the likelihood and the conjugate prior, we simply add the sum of sufficient statistics (of the likelihood) to the original parameters of the conjugate prior. This expression also implies that in order to compute posterior distributions for the graphical model depicted in Figure 3, we only need to keep track of the number of data points, N, and of the sum of sufficient statistics for the individual variables, but the data itself is not needed.

Example: using the exponential family conjugate prior for Bayesian prediction

We now show an example where the exponential family is used to derive a posterior distribution for an unobserved variable X_{new}, given N observed IID variables X_n, model parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\tau}$. The graphical model corresponding to this problem is shown in Figure 4. We are interested in $p(\mathbf{x}_{\text{new}}|\mathbf{x})$. More specifically, taking the model parameters into account, for a given (fixed) set of parameters $\boldsymbol{\theta}$, we have

$$\begin{aligned} p(\mathbf{x}_{\text{new}}|\mathbf{x}) &= p(\mathbf{x}_{\text{new}}, \boldsymbol{\theta}|\mathbf{x}) \\ &= \int p(\mathbf{x}_{\text{new}}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{x})\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{21}$$

The last step in Eqn. 21 follows directly from the Bayes ball algorithm. Now we simply plug in the two densities in Eqns 15 and 18: to obtain the posterior distribution for the predicted variable:

$$p(\mathbf{x}_{\text{new}}|\mathbf{x}) = h(\mathbf{x}_{\text{new}}) \frac{1}{Z(\boldsymbol{\tau} + \mathbf{T}, \tau_0 + N)} \int \exp\left\{ (\boldsymbol{\tau} + \mathbf{T} + \boldsymbol{\tau}_{\text{new}})^T - (\boldsymbol{\tau} + N + 1) + A(\boldsymbol{\eta}(\boldsymbol{\theta})) \right\} \mathrm{d}\boldsymbol{\theta}, \tag{22}$$

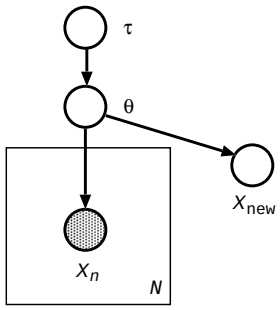

Figure 4: Graphical model of the Bayesian prediction of variable X_{new}, given N observed variables X_n, the model parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\tau}$.

where $\mathbf{T}_{\text{new}} = \mathbf{T}(X_{\text{new}})$. The reason we add one in the term $(\boldsymbol{\tau} + N + 1)$ above is that we have one more observation – that of the predicted variable. At first site this integral may seem intimidating, but it is actually easy to deal with because it is simply the normalization factor Z, as follows from Eqn. 17. Eqn. 17 implies

$$\begin{aligned} Z(\boldsymbol{\tau}, \tau_0) &= \int \exp\left\{\boldsymbol{\tau}^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \tau_0 A(\boldsymbol{\eta}(\boldsymbol{\theta}))\right\} \mathrm{d}\theta \\ &= h(\mathbf{x}_{\text{new}}) \frac{Z(\boldsymbol{\tau} + \mathbf{T} + \boldsymbol{\tau}_{\text{new}}, \tau_0 + N + 1)}{Z(\boldsymbol{\tau} + \mathbf{T}, \tau_0 + N)}, \end{aligned} \tag{23}$$

i.e. the original function $h()$ is absorbed in the new normalization factor.

In summary, we have derived the prior $p(\boldsymbol{\theta}|\boldsymbol{\tau})$ for exponential family IID variables, and we've showed that this prior is in the exponential family, which makes it a conjugate prior. In the conjugate prior, the function $\boldsymbol{\eta}(\boldsymbol{\theta})$ plays the role of the sufficient statistics and $A(\boldsymbol{\eta}(\boldsymbol{\theta}))$ has $\boldsymbol{\theta}$ in it, so it should go up front. It is important to note that in order toget the closure property we were seeking, we had to introduce as many hyperparameters as parameters.