CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Conjugacy and the Exponential Family

Lecturer: Michael I. Jordan *Scribes: Brian Milch*

1 Conjugacy

In the previous lecture, we saw conjugate priors for the multivariate Gaussian distribution. In this lecture, we discuss conjugacy more generally. A family of probability distributions $\mathcal{P}$ is *conjugate* for a probability model p if the posterior lies in $\mathcal{P}$ whenever the prior lies in $\mathcal{P}$. Note that the family of all probability distributions is conjugate for any model p.

What if we have a probability model $p(y|\theta)$ for some practical problem, but the standard conjugate family for p does not seem to contain a realistic prior distribution? One thing we can do is use a mixture model as our prior:

$$p(\theta) = \sum_{i=1}^{k} \alpha_{ij} p_j(\theta)$$

where $0 \leq \alpha_i \leq 1$, $\sum_i \alpha_i = 1$, and each $p_j(\theta)$ is in a family $\mathcal{P}_j$ that is conjugate for $p(y|\theta)$. If we multiply this $p(\theta)$ by $p(y|\theta)$, then each $p_j(\theta)$ is multiplied by $p(y|\theta)$, yielding another distribution in $\mathcal{P}_j$. So the posterior $p(\theta|y)$ is a mixture of the same form. The same result holds for an infinite mixture indexed by a continuous parameter τ:

$$p(\theta) = \int g(\tau) p(\theta; \tau)\, d\tau$$

where $g(\tau)$ is an arbitrary density function.

We can also take a limit of increasingly flat conjugate priors, yielding an *improper prior* with density 1 everywhere. This is the ultimate uninformative prior, but it is improper in that it does not integrate to 1. If we multiply it by a $p(y|\theta)$ distribution, such as a Gaussian, the resulting posterior may integrate to 1 and thus be a true density. However, this is not guaranteed for all choices of $p(y|\theta)$.

If we use a mixture model, how can we set the α's? To be fully Bayesian, we should set them based on some background knowledge. However, a common technique is *empirical Bayes*: estimate the α's by maximum likelihood on the training data.

2 The Exponential Family

For a fully general treatment of conjugate priors, we turn to a very large family of distributions called the *exponential family*, which actually contains every parametric family of distributions. The general form of an exponential family distribution is:

$$p(x|\theta) = h(x) \exp\left(\phi(\theta)^\top T(x) - A(\theta)\right) \tag{1}$$

Here $\phi(\theta)$ is the *canonical parameter*, often denoted η. $T(x)$ is the *sufficient statistic*, and $A(\theta)$ is the *cumulant generating function* or *log partition function*. The x here can range over any set, as long as $T(x)$

maps each x to a vector of fixed, finite dimension. $p(x|\theta)$ is a density with respect to some underlying measure μ. The $h(x)$ function can just be thought of as an adjustment to this underlying measure, and is not usually important.

Since the expression in Eq. 1 is a density, it must integrate to 1:

$$\int h(x)e^{\eta^\top T(x)-A(\theta)}\,dx = 1$$

where $\eta = \phi(\theta)$. Therefore:

$$A(\theta) = \ln \int h(x)e^{\eta^\top T(x)}\,dx \tag{2}$$

So $A(\theta)$ is fully determined by η and $T(x)$.

It can be shown that the set of valid η vectors $\{\eta : \int h(x)e^{\eta^\top T(x)}\,dx < \infty\}$ is convex. So convex combinations of valid parameter vectors are also valid parameter vectors. Also, optimizing over the set of valid η's is not too difficult.

So far we have written the cumulant generating function as $A(\theta)$. We can also write it as $A(\eta)$, using a different function A. We will limit ourselves to cases where $\phi(\theta)$ is one-to-one. The second convexity property of the exponential family is that $A(\eta)$ is always a convex function of η. This follows from a general convex analysis result about log-sum-exp equalities such as Eq. 2.

$A(\eta)$ is called the cumulant generating function because its derivatives with respect to η are the cumulants (central moments, i.e., mean, variance, kurtosis, etc.) of the sufficient statistic $T(X)$. For the first cumulant, we take the first derivative:

$$\nabla A(\eta) = \frac{\int h(x)e^{\eta^\top T(x)}T(x)\,dx}{\int h(x)e^{\eta^\top T(x)}\,dx}$$

By Eq. 2, the denominator is $e^{A(\eta)}$, so:

$$\begin{aligned}\nabla A(\eta) &= \int h(x)e^{\eta^\top T(x)-A(\eta)}T(x)\,dx \\ &= \mathbb{E}T(x)\end{aligned}$$

Similarly, the Hessian matrix $\nabla^2 A(\eta)$ is the covariance matrix $\mathrm{Var}(T(x))$. Thus, to find the cumulants of an exponential family distribution with a given $A(\eta)$, we don't have to do messy integrals; we just have to take derivatives.

For more information on the exponential family, see the recent technical report by Wainwright and Jordan, "Graphical models, exponential families, and variational inference" (available from Prof. Jordan's web page). A good book on the subject is by Lawrence Brown, *Fundamentals of Statistical Exponential Families*, published in the IMS Lecture Notes series in 1986. The exponential family is quite powerful: it includes all the standard distributions such as the Bernoulli, Gaussian, gamma, Poisson, Raleigh, etc. However, exponential family distributions are parametric: the parameter vector η has a fixed dimension.

2.1 Conjugacy and the exponential family

Consider the setup in Fig. 1, where we take n samples y_i from an exponential family distribution:

$$p(y_i|\theta) = h(y_i)\exp\left(\phi(\theta)^\top T(y_i) - A(\theta)\right) \tag{3}$$

Let y be the random vector formed by concatenating all the samples y_i. Then:

$$p(y|\theta) = \left(\prod_i h(y_i)\right)\exp\left(\phi(\theta)^\top \sum_i T(y_i) - nA(\theta)\right) \tag{4}$$

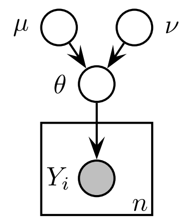

Figure 1: Graphical model for an exponential family distribution and its conjugate prior.

Thus, y also has an exponential family distribution: it has the same canonical parameter $\phi(\theta)$; its sufficient statistic is $\sum_i T(y_i)$; and its cumulant generating function is $nA(\theta)$. We can construct a conjugate family of prior distributions as follows, with two parameters μ and ν:

$$p(\theta|\mu,\nu) \propto \exp\left(\phi(\theta)^\top \mu - \nu A(\theta)\right) \tag{5}$$

Then the posterior distribution is:

$$p(\theta|y,\mu,\nu) \propto \exp\left(\phi(\theta)^\top \left(\mu + \sum_i T(y_i)\right) - (n+\nu)A(\theta)\right) \tag{6}$$

We can also take a hierarchical Bayesian approach, putting priors on μ and ν as well. It is worth noting that this is just the *minimal* conjugate family for $p(y|\theta)$, in the sense that is has a minimal number of parameters. Of course there are other conjugate families, such as the family of mixtures of these distributions, and the family of all distributions.

2.2 ML estimation of mean parameters

Putting aside the Bayesian approach for the moment, suppose we want a maximum likelihood (ML) estimate of the mean parameter $\mu = \mathbb{E}T(X)$ for some exponential family distribution. This μ is a function of η: specifically $\mu = \nabla_\eta A(\eta)$. So it suffices to maximize the likelihood with respect to η. Based on Eq. 4, the log likelihood is:

$$\ell(\eta; y) = \eta^\top \sum_{i=1}^n T(y_i) - nA(\eta) + c(y) \tag{7}$$

where $c(y)$ is some function that does not depend on η. Differentiating with respect to η, we get:

$$\begin{aligned} \nabla_\eta \ell &= \sum_{i=1}^n T(y_i) - n\nabla_\eta A(\eta) \\ &= \sum_{i=1}^n T(y_i) - n\mathbb{E}T(X) \end{aligned}$$

Setting this to zero, we get:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^n T(y_i) \tag{8}$$

In other words, the maximum likelihood estimate for the expectation of the sufficient statistic is just the empirical mean of the sufficient statistic. We already knew this fact for common distributions such as the Gaussian, Poisson, etc.; now we have a general proof.

2.3 Exercises with the exponential family

Here are some exercises for the reader involving the exponential family. Consider the Poisson distribution:

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

and the binomial distribution (with a fixed n):

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Express each of these distributions in exponential family form. For example, the canonical parameter η for the binomial distribution is $\ln\frac{\theta}{1-\theta}$. Compute $A(\theta)$, and then differentiate it to obtain the mean μ for each distribution.

2.4 Parameterizations

We have seen several ways of parameterizing the exponential family, that is, indexing the set of exponential family distributions. There is the *canonical parameter* η, and also the *mean parameter* $\mu = \mathbb{E}T(X)$. Recall that $A(\eta)$ is a convex function, so $\mu = \nabla_\eta A(\eta)$ is nondecreasing in η. If $A(\eta)$ is *strictly* convex, then $\nabla_\eta A(\eta)$ is increasing in η, and therefore the mapping between η and μ is one-to-one. See Fig. 2 for an illustration of this relationship.

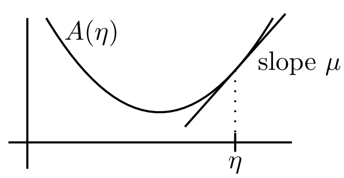

Figure 2: If η is one-dimensional, then the mean parameter μ corresponding to η is just the slope of a tangent line to the cumulant generating function A evaluated at η.

We have also used another parameter, denoted θ to index the set of exponential family distributions. This parameter does not have a special name; it's just some parameter that is convenient for defining the distribution.

3 The Exponential Family and Graphical Models

3.1 The Ising model

The previous course in this sequence dealt with probabilistic graphical models. We can also think of graphical models as defining exponential family distributions. As an example, consider the Ising model, illustrated in Fig. 3. This model comes from statistical physics, where each node represents the spin (up or down) of a particle. We represent each particle's spin with a variable X_i taking values $\{0, 1\}$ (we can formulate an equivalent model with $\{-1, 1\}$). The parameters are θ_i, representing the external field on particle i, and θ_{ij}, representing the attraction between particles i and j. If i and j are not adjacent in the graph, then $\theta_{ij} = 0$.

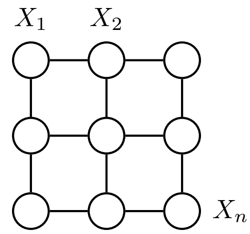

Figure 3: An Ising model with $n = 9$ nodes

The probability distribution is:

$$\begin{aligned} p(x|\theta) &= \exp\left(\sum_{i<j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i - A(\theta)\right) \\ &= \frac{1}{Z(\theta)} \exp\left(\sum_{i<j} \theta_{ij} x_i x_j + \sum_i \theta_i x_i\right) \end{aligned}$$

where $Z(\theta)$ is the *partition function*.

This is an exponential family distribution where the sufficient statistic $T(x)$ consists of all the values x_i and $x_i x_j$ (for $i < j$) concatenated together. So if $\mu \triangleq \mathbb{E}T(X)$, then $\mu_i = \mathbb{E}X_i$ and $\mu_{ij} = \mathbb{E}X_i X_j$. Thus, the μ vector contains the expectations and correlations of the particles' spins. It can be shown that $A(\eta)$ is strictly convex, so there is a one-to-one mapping between η and μ. However, actually computing μ from η is #P-hard.

In fact, we can think of the whole problem of probabilistic inference as computing the mean parameter μ from the canonical parameter η (or some other parameter θ). Of course, this problem is #P-hard in general.

3.2 Graphical models in general

In an undirected graphical model, we specify a potential ψ_C for each clique C of the graph. The joint probability distribution is then given by:

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C) \tag{9}$$

where Z is some normalization constant. We can write this in exponential family form as:

$$p(x) = \exp\left(\sum_C \ln \psi_C(x_C) - \ln Z\right) \tag{10}$$

If all the variables in the model have discrete values, then for each clique C, we can define an indicator vector with an entry for each configuration of the variables in C. The sufficient statistic $T(x)$ is formed by concatenating the indicator vectors for all the cliques. In general, if each ψ_C is an exponential family distribution, we can form $T(x)$ by concatenating the sufficient statistics for these distributions.

In a directed graphical model, we specify a conditional distribution for each variable given its parents in the graph. The joint distribution is:

$$p(x) = \prod_i p(x_i | x_{\pi(i)}) \tag{11}$$

The exponential family form of this distribution is simple:

$$p(x) = \exp\left(\sum_i \ln p(x_i|x_{\pi(i)})\right) \tag{12}$$

If each $p(x_i|x_{\pi(i)})$ is an exponential family distribution, then it is clear that the distribution in Eq. 12 is also in the exponential family.