COS513, LECTURE 13

SAM GERSHMAN, RICHARD SOCHER

1. EXPONENTIAL FAMILY DISTRIBUTIONS: MAXIMUM LIKELIHOOD ESTIMATION

Let $X_{1:N}$ denote our observed data (consisting of N datapoints) that we assume is drawn from a distribution in the exponential family (EF). Recall that an EF distribution is parameterized by a *natural parameter* η, a function $t(X)$ referred to as the *sufficient statistic* (SS), a function $a(\eta)$ referred to as the *log-normalizer*, and a function $h(X)$ that enforces the underlying measure with respect to which $P(X|\eta)$ is a density.[1] The *likelihood* is defined as

$$P(X_{1:N}|\eta) = \prod_{n=1}^{N} h(x_n) \exp\left\{\eta^{\mathrm{T}} t(X_{1:N}) - a(\eta)\right\} \tag{1}$$

$$= \left[\prod_{n=1}^{N} h(x_n)\right] \exp\left\{\eta^{\mathrm{T}} \sum_{n=1}^{N} t(x_n) - N a(\eta)\right\}. \tag{2}$$

A valuable aspect of working with the EF form is that all the information provided by the data is encapsulated by the SS. This means that it is not generally necessary to store all the datapoints in memory. For example, the SS of the Gaussian distribution is $t(X) = \langle \sum_n x_n, \sum_n x_n^2 \rangle$. Intuitively, the SS correspond to the sample mean and variance, respectively. As another example, the SS of the Bernoulli distribution is $t(X) = \sum_n x_n$. In this case, the SS can be interpreted intuitively as the number of "heads" in a series of coin flips.

Returning to the likelihood, we are interested in finding the value of the natural parameter that maximizes the likelihood function. This is the *maximum likelihood estimate* $\hat{\eta}_{ML}$. For numerical stability, it is often advantageous to work with the log-likelihood:

$$l(\eta; X_{1:N}) = \log P(X_{1:N}|\eta) = \sum_{n=1}^{N} \log h(x_n) + \eta^{\mathrm{T}} \sum_{n=1}^{N} t(x_n) - N a(\eta). \tag{3}$$

[1] Note that we use uppercase P generically to refer to either a density or a distribution function

Taking the gradient of the log-likelihood with respect to the natural parameter yields

$$\nabla_\eta l = \sum_{n=1}^{N} t(x_n) - N \nabla_\eta a(\eta). \tag{4}$$

Setting the gradient to zero gives:

$$\nabla_\eta A(\hat{\eta}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} t(x_n) \tag{5}$$

$$= \mathbb{E}\left[t(X_{1:N})\right]. \tag{6}$$

To state this result in words: *at the ML solution, the expectation of the SS is simply equal to the sample mean of the SS*. This result explains how one fits *all* EF distributions. For example, fitting a Gaussian entails computing the sample mean and variance, which is the expected SS. Exploiting the structure provided by the EF form obviates the need to directly optimize the likelihood function (e.g., with Lagrange multipliers).

2. BAYESIAN INFERENCE FOR EXPONENTIAL FAMILY DISTRIBUTIONS

In the Bayesian setting, we place a *prior* distribution $P(\eta)$ on the natural parameter and attempt to infer the *posterior* distribution $P(\eta|X_{1:N})$. Significant mathematical convenience is obtained when the prior and posterior have the same functional form. We say that the prior is *conjugate* to the data generating distribution.[2] We emphasize that conjugacy is useful for *mathematical convenience*, and many reasonable models will be non-conjugate. As an example, a Gaussian prior is conjugate to a Gaussian likelihood with fixed variance. Another example is the Beta distribution, which is conjugate to the Bernoulli distribution. To show this explicitly, we now work through this example in detail.

2.1. **Bayesian inference for the Beta-Bernoulli model.**

In its mean parameterization (with mean π), the Bernoulli distribution is written as:

$$P(x|\pi) = \pi^x (1-\pi)^{1-x}, \tag{7}$$

and the Beta distribution with parameters α and β is written as:

$$P(\pi|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}, \tag{8}$$

where Γ is the Gamma function (an extension of the factorial function to real and complex numbers). The first factor in Eq. 8 acts as a normalizing

[2] Note that no distribution is *universally* conjugate; conjugacy is relative to the data generating distribution.

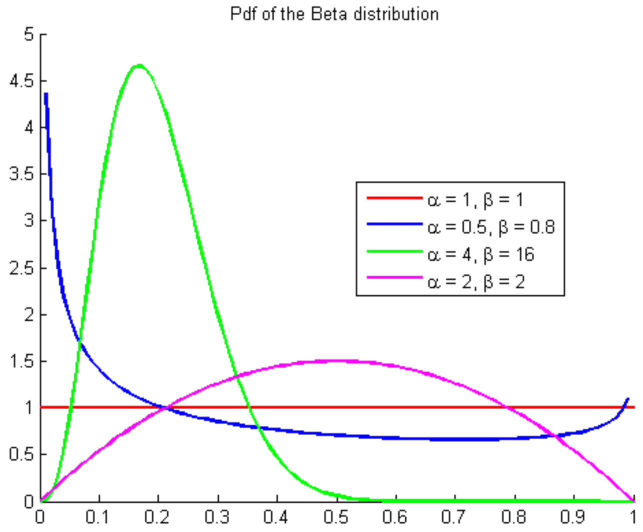

FIGURE 1. Probability density function of the Beta distribution with different parameters.

constant. Figure 1 shows the probability density function of several parameter values.

Now suppose that the observed data $X_{1:N}$ was generated according to the following generative model:

$$\pi \sim Beta(\alpha, \beta) \tag{9}$$

$$x_n \sim Bern(\pi), n = 1 \dots N. \tag{10}$$

We then have following expression for the posterior:

$$\begin{aligned} P(\pi|X_{1:N}, \alpha, \beta) &\propto P(\pi, X_{1:N}|\alpha, \beta) \quad (11)\\ &= P(\pi|\alpha, \beta) \prod_{n=1}^{N} P(x_n|\pi) \quad (12)\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \prod_{n=1}^{N} \pi^{x_n}(1-\pi)^{1-x_n} \quad (13)\\ &\propto \pi^{\alpha+\sum_n x_n - 1}(1-\pi)^{\beta+\sum_n(1-x_n)-1}. \quad (14) \end{aligned}$$

Thus we see that the posterior distribution is a Beta distribution with parameters $\alpha' = \alpha + \sum_n x_n$ and $\beta' + \sum_n(1 - x_n)$. Intuitively, α and β function as "fictional" datapoints prior to observing $X_{1:N}$. With more and more data, the prior will lose its importance.

The expected value of the mean parameter under a $Beta(\alpha, \beta)$ distribution is

$$\mathbb{E}\left[\pi\right] = \frac{\alpha}{\alpha + \beta}. \tag{15}$$

2.2. Bayesian inference for the general exponential family model.

After having looked at the specific example of the Beta distribution, we can now derive MLE for an arbitrary distribution in the exponential family. Let us assume the following abstract generative process

$$\begin{aligned} \eta &\sim Conj(\lambda) && (16) \\ X_n &\sim Exp - fam(\eta). && (17) \end{aligned}$$

The likelihood of the data and the posterior of the parameters become

$$\begin{aligned} p(x_n|\eta) &= h_\eta(x) \exp\{\eta^T t(x_n) - a(\eta)\} && (18) \\ p(\eta|\lambda) &= h_\lambda(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a_x(\eta)) - a_c(\lambda)\}, && (19) \end{aligned}$$

where $\langle \lambda_1, \lambda_2 \rangle$ are the natural parameters with dimension $\dim(\eta) + 1$. The sufficient statistics are $\langle \eta, -a_x(\eta) \rangle$ and $a_c(\lambda)$ is the log normalizer of the conjugate prior. We can now find a close form solution for the posterior

$$\begin{aligned} & p(\eta|x_{1:N}, \lambda) && (20) \\ \propto\ & p(\eta|\lambda) \prod_n p(x_n|\eta) && (21) \\ \propto\ & h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a_x(\eta)) - a_c(\lambda)\} && (22) \\ \cdot\ & \exp\{\eta^T \sum_n t(x_n) - N a_x(\eta)\} && (23) \\ \propto\ & h(\eta) \exp\{(\lambda_1 + \sum_n t(x_n))^T \eta + (\lambda_2 + N)(-a_x(\eta))\} && (24) \end{aligned}$$

Because the prior is conjugate with respect to the likelihood function, the posterior is in the same functional family as the prior with the parameters $\hat{\lambda}_1 = \lambda_1 + \sum_n t(x_n)$ and $\hat{\lambda}_2 = \lambda_2 + N$.

3. Latent Variable Models (chapter 10)

In a graphical model, latent variables are random variables which are not observed. They are very useful in settings where we assume that our data are generated by a hidden cause. Furthermore, they provide structure to the underlying distribution of observations. We will first investigate the simplest such model, a mixture of Gaussians.

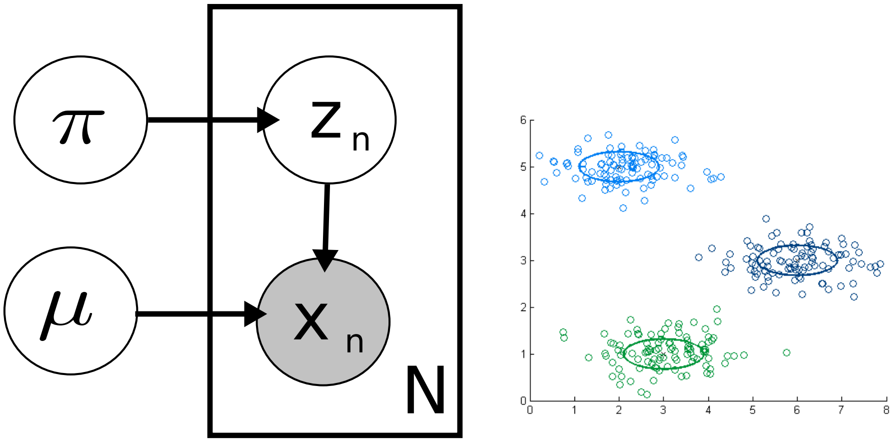

FIGURE 2. (Left) Graphical model representation of the Gaussian mixture model. (Right) Synthetic data generated from a manually set common covariance matrix and fixed means.

3.1. **Gaussian mixture model.** Fig 2 (left) shows the graphical model. The mixture components are Gaussian distributions. π are the mixing proportions and μ are the means of the mixture components. The assumed underlying generating process is simply

For n = 1..N

$$z_n \sim Mult(\pi) \tag{25}$$

$$x_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}), \tag{26}$$

where we assume that the covariance matrices Σ are given. Fig. 2 (right) shows data generated from a mixture of Gaussians.