24 Learning Exponential Family Models

So far, in our discussion of learning, we have focused on discrete variables from finite alphabets. We derived a convenient form for the MAP estimate in the case of DAGs:

$$\theta_{i,\pi_i} = p_{x_i|\mathbf{x}_{\pi_i}}(\cdot|\cdot) = \hat{p}_{x_i|\mathbf{x}_{\pi_{x_i}}}(\cdot|\cdot). \tag{1}$$

In other words, we chose the CPD entries θ such that the model distribution p matches the empirical distribution $\hat{p}$. We will see that this is a special case of a more general result which holds for many families of distributions which take a particular form. However, we can't apply (1) directly to continuous random variables, because the empirical distribution $\hat{p}$ is a mixture of delta functions, whereas we ultimately want a continuous distribution p. Instead, we must choose p so as to match certain statistics of the empirical distribution.

24.1 Gaussian Parameter Estimation

To build an intuition, we begin with the special case of Gaussian distributions. Suppose we are given an i.i.d. sample $x_1, \ldots, x_K$ from a Gaussian distribution $p(\cdot; \mu, \sigma^2)$. Then, by definition,

$$\begin{aligned} \mu &= \mathbb{E}_p[x] \\ \sigma^2 &= \mathbb{E}_p[(x-\mu)^2]. \end{aligned}$$

This suggests setting the parameters equal to the corresponding empirical statistics. In fact, it is straightforward to check, by setting the gradient to zero, that the maximum likelihood parameter estimates are[1]

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{1}{K}\sum_{k=1}^{K} x_k \\ \hat{\sigma}^2_{ML} &= \frac{1}{K}\sum_{k=1}^{K}(x_k - \hat{\mu})^2. \end{aligned}$$

[1] In introductory statistics classes, we are often told to divide by $K-1$ rather than K when estimating the variance of a multivariate Gaussian. This is in order that the resulting estimate be unbiased. The maximum likelihood estimate, as it turns out, is biased. This is not necessarily a bad thing, because the biased estimate also has lower variance, i.e. there is a tradeoff between bias and variance.

The corresponding ML estimates for a multivariate Gaussian are given by:

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k$$
$$\hat{\boldsymbol{\Lambda}}_{ML} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{ML})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{ML})^T.$$

24.1.1 ML parameter estimation in Gaussian DAGs

We saw that in a Gaussian DAG, the CPDs take the form

$$p(x_i|\mathbf{x}_{\pi_i}) = \mathtt{N}(\beta_0 + \beta_1 u_1 + \cdots + \beta_L u_L; \sigma^2),$$

where $\mathbf{x}_{\pi_i} = (u_1, \ldots, u_L)$. We can rewrite this in innovation form:

$$x_i = \beta_0 + \beta_1 u_1 + \cdots + \beta_L u_L + w,$$

where $w \sim \mathtt{N}(0, \sigma^2)$ is independent of the u_l's. This representation highlights the relationship with linear least squares estimation, and shows that $\hat{\boldsymbol{\theta}}_{ML}$ is the solution to a set of $L+1$ linear equations.

The parameters of this CPD are $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_L, \sigma^2)$. When we set the gradient of the log-likelihood to zero, we get:

$$\hat{\beta}_0 = \hat{\mu}_x - \hat{\boldsymbol{\Lambda}}_{xu}\boldsymbol{\Lambda}_{uu}^{-1}\hat{\boldsymbol{\mu}}_u$$
$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}_{uu}^{-1}\hat{\boldsymbol{\Lambda}}_{ux}$$
$$\hat{\Lambda}_0 = \hat{\Lambda}_{xx} - \hat{\boldsymbol{\Lambda}}_{xu}\boldsymbol{\Lambda}_{uu}^{-1}\hat{\boldsymbol{\Lambda}}_{ux},$$

where

$$\hat{\mu}_x = \frac{1}{K}\sum_{k=1}^{K} x_i$$
$$\hat{\Lambda}_{xx} = \frac{1}{K}\sum_{k=1}^{K}(x_i - \hat{\mu}_x)^2$$
$$\hat{\boldsymbol{\Lambda}}_{xu} = \frac{1}{K}\sum_{k=1}^{K}(x_i - \hat{\mu}_x)(\mathbf{u}_k - \hat{\boldsymbol{\mu}}_u)^T$$
$$\hat{\boldsymbol{\Lambda}}_{uu} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{u}_i - \hat{\boldsymbol{\mu}}_u)(\mathbf{u}_i - \hat{\boldsymbol{\mu}}_u)^T.$$

Recall, however, that if $p(x|\mathbf{u}) = \mathtt{N}(\beta_0 + \boldsymbol{\beta}^T\mathbf{u}, \Lambda_0)$, then

$$\begin{aligned}\beta_0 &= \mu_x - \boldsymbol{\Lambda}_{xu}\boldsymbol{\Lambda}_{uu}^{-1}\boldsymbol{\mu}_u \\ \boldsymbol{\beta} &= \boldsymbol{\Lambda}_{uu}^{-1}\boldsymbol{\Lambda}_{ux} \\ \boldsymbol{\Lambda}_0 &= \Lambda_{xx} - \boldsymbol{\Lambda}_{xu}\boldsymbol{\Lambda}_{uu}^{-1}\boldsymbol{\Lambda}_{ux}.\end{aligned}$$

In other words, ML in Gaussian DAGs is another example of moment matching.

24.1.2 Bayesian Parameter Estimation in Gaussian DAGs

In our discussion of discrete DAGs, we derived the ML estimates, and then derived Bayesian parameter estimates in order to reduce variance. Now we do the equivalent for Gaussian DAGs. Observe that the likelihood function for univariate Gaussians takes the form

$$\begin{aligned}p(\mathcal{D}; \mu, \sigma^2) \propto &\frac{1}{\sigma^K}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{K}(x_k - \mu)^2\right) \\ &= \frac{1}{\sigma^K}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{K}x_k^2 - 2\sum_{k=1}^{K}x_k\mu + K\mu^2\right)\right) \qquad (2)\end{aligned}$$

As in the discrete case, we use this functional form to derive conjugate priors. First, if σ^2 is known, we see the conjugate prior takes the form

$$p(\mu) \propto \exp\left(a\mu - b\mu^2\right).$$

In other words, the conjugate prior for the mean parameter of a Gaussian with known variance is a Gaussian.

What if the mean is known and we want a prior over the variance? We simply take the functional form of (2), but with respect to σ this time, to find a conjugate prior

$$p(\sigma) \propto \frac{1}{\sigma^a}\exp\left(-\frac{b}{\sigma^2}\right).$$

This distribution has a more convenient form when we rewrite it in terms of the precision $\tau = 1/\sigma$. Then, we see that the conjugate prior for τ is a gamma distribution:

$$p(\tau) \propto \tau^{\alpha-1}e^{-\beta\tau}.$$

The corresponding prior over σ is known as an *inverse Gamma distribution*. When both μ and τ are unknown, we get what is called an *Gaussian-Gamma* prior:

$$p(\mu, \tau) = \mathtt{N}(\mu; a, (b\lambda)^{-1})\,\Gamma(\lambda; \alpha, \beta).$$

Analogous results hold for the multivariate Gaussian case. The conjugate prior for $\boldsymbol{\mu}$ with known $\mathbf{J}$ is a multivariate Gaussian. The conjugate prior for $\mathbf{J}$ with known $\boldsymbol{\mu}$ is a matrix analogue of the gamma distribution called the *Wishart distribution*. When both $\boldsymbol{\mu}$ and $\mathbf{J}$ are unknown, the conjugate prior is a *Gaussian-Wishart distribution*.

24.2 Linear Exponential Families

We've seen three different examples of maximum likelihood estimation which led to similar-looking expectation matching criteria: CPDs of discrete DAGs, potentials in undirected graphs, and Gaussian distributions. These three examples are all special cases of a very general class of probabilistic models called *exponential families*. A family of distributions is an exponential family if it can be written in the form

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\theta)} \exp\left(\sum_{i=1}^{k} \theta_i f_i(\mathbf{x})\right),$$

where $\mathbf{x}$ is an N-dimensional vector and $\boldsymbol{\theta}$ is a k-dimensional vector. The functions f_i are called features, or *sufficient statistics*, because they are sufficient for estimating the parameters. When the family of distributions is written in this form, the parameters $\boldsymbol{\theta}$ are known as the *natural parameters*.

Let's consider some examples.

1. A multinomial distribution can be written as an exponential family with $f_{a^0}(x) = \mathbb{1}_{x=a^0}$ and the natural parameters are $\theta_{a^0} = \ln p(a^0)$.

2. In an undirected graphical model, the distribution can be written as:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \\ &= \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \ln \psi_C(\mathbf{x}_C)\right) \\ &= \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \sum_{\mathbf{x}'_C \in |\mathcal{X}|^{|C|}} \ln \psi_C(\mathbf{x}'_C) \mathbb{1}_{\mathbf{x}_C = \mathbf{x}'_C}\right). \end{aligned}$$

   This is an exponential family representation where the sufficient statistics correspond to indicator variables for each clique and each joint assignment to the variables in that clique:

$$f_{C,\mathbf{x}'_C}(\mathbf{x}) = \mathbb{1}_{\mathbf{x}_C = \mathbf{x}'_C},$$

   and the natural parameters $\boldsymbol{\theta}_{C,\mathbf{x}_C}$ correspond to the log potentials $\ln \psi_C(\mathbf{x}_C)$. Observe that this is the same parameterization of undirected graphical models which we used to derive the tree reweighted belief propagation algorithm in our discussion of variational inference.

3. If the variables $\mathsf{x} = (x_1, \ldots, x_N)$ are jointly Gaussian, the joint PDF is given by

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2} \sum_{i,j} J_{ij} x_i x_j + \sum_i h_i x_j\right).$$

This is an exponential family with sufficient statistics

$$f(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_1 \\ \vdots \\ x_N \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_N^2 \end{pmatrix}.$$

and natural parameters

$$\boldsymbol{\theta} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \\ -\frac{1}{2}J_{11} \\ -\frac{1}{2}J_{12} \\ \vdots \\ -\frac{1}{2}J_{NN} \end{pmatrix}.$$

4. We might be tempted to conclude from this that every family of distributions is an exponential family. However, in fact families are not. As a simple example, even the family of Laplacian distributions with scale parameter 1

$$p(x;\theta) = \frac{1}{Z}e^{-|x-\theta|}$$

is not an exponential family.

24.2.1 ML Estimation in Linear Exponential Families

Suppose we are given i.i.d. samples $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_K$ from a discrete exponential family $p(x;\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(\boldsymbol{\theta}^T f(\mathbf{x}))$. As usual, we compute the gradient of the log likelihood:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\frac{1}{K}\ell(\theta;\mathcal{D}) &= \frac{1}{K}\sum_{k=1}^{K}\frac{\partial}{\partial\theta_i}\boldsymbol{\theta}^T f(\mathbf{x}) - \frac{\partial}{\partial\theta_i}\ln Z(\boldsymbol{\theta}) \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{\partial}{\partial\theta_i}\boldsymbol{\theta}^T f(\mathbf{x}) - \frac{\partial}{\partial\theta_i}\ln\sum_{\mathbf{x}}\exp\left(\boldsymbol{\theta}^T\mathbf{x}\right) \\
&= \mathbb{E}_{\mathcal{D}}[f_i(\mathbf{x})] - \mathbb{E}_{\boldsymbol{\theta}}[f_i(\mathbf{x})].
\end{aligned}$$

The derivation in the continuous case is identical, except that the partition function expands to an integral rather than a sum. This shows that in any exponential family,

the ML parameter estimates correspond to *moment matching*, i.e. they match the empirical expectations of the sufficient statistics:

$$\mathbb{E}_{\mathcal{D}}[f_i(\mathbf{x})] = \mathbb{E}_{\hat{\theta}}[f_i(\mathbf{x})].$$

Interestingly, in our discussion of Gaussians, we found the ML estimates by taking derivatives with respect to the information form parameters, and we wound up with ML solutions in terms of the covariance form. Our derivation here shows that this phenomenon applies to all exponential family models, not just Gaussians. We can summarize these analogous representations in a table:

	natural parameters	expected sufficient statistics
Gaussian distribution	information form	covariance form
multinomial distribution	log odds	probability table
undirected graphical model	log potentials	clique marginals

24.2.2 Maximum Entropy Interpretation

In the last section, we started with a parametric form of a distribution, maximized the data likelihood, and wound up with a constraint on the expected sufficient statistics. Interestingly, we can arrive at the same solution from the opposite direction: starting with constraints on the expected sufficient statistics, we choose a distribution which maximizes the entropy subject to those constraints, and it turns out to have that same parametric form. Intuitively, the entropy of a distribution is a measure of how spread out, or uncertain, it is. If all we know about some distribution is the expectations of certain statistics $f_i(\mathbf{x})$, it would make sense to choose the distribution with as little commitment as possible, i.e. the one that is most uncertain, subject to those constraints. This suggests maximizing the entropy subject to the moment constraints:

$$\begin{aligned} \max_{p} \quad & H(p) \\ \text{subject to} \quad & \mathbf{E}_p[f_i(\mathbf{x})] = \mathbf{E}_{\mathcal{D}}[f_i(\mathbf{x})] \\ & \sum_{\mathbf{x}} p(\mathbf{x}) = 1. \end{aligned}$$

For simplicity, assume p is a discrete distribution. To solve this optimization problem, we write out the Lagrangian:

$$\begin{aligned} \mathcal{L}(p) &= H(p) + \sum_{i=1}^{K} \lambda_i \left(\mathbb{E}_p[f_i(\mathbf{x})] - \mathbb{E}_{\mathcal{D}}[f_i(\mathbf{x})]\right) + \nu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1\right) \\ &= -\sum_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x}) + \sum_{i=1}^{K} \lambda_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) - \mathbf{E}_{\mathcal{D}}[f_i(\mathbf{x})]\right) + \nu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1\right). \end{aligned}$$

We take the derivative with respect to $p(\mathbf{x})$:

$$\frac{\partial \mathcal{L}}{\partial p(\mathbf{x})} = -\ln p(\mathbf{x}) - 1 + \sum_{i=1}^{K} \lambda_i f_i(\mathbf{x}) + \nu.$$

Setting this to zero, we find that

$$p(\mathbf{x}) \propto \exp\left(\sum_{i=1}^{K} \lambda_i f_i(\mathbf{x})\right).$$

This is simply the definition of an exponential family with sufficient statistics f_i. This shows a striking and philosophically interesting equivalence between exponential families and the principle of maximum entropy.

6.438 Algorithms for Inference

Fall 2014