Generative Models II

Aaron Courville

University of Montreal

Deep Learning Summer School 2017
Presenter: Chao Jiang

Generative models II: Outline

- Autoregressive models
  - PixelCNN
- Latent variable models
  - Variational Autoencoders
  - Generative Adversarial Networks

Comparision

Generative Models

- PixelRNN and PixelCNN — Explicit density model, optimizes exact likelihood, good samples. But inefficient sequential generation.
- Variational Autoencoders (VAE) — Optimize variational lower bound on likelihood. Useful latent representation, inference queries. But current sample quality not the best.
- Generative Adversarial Networks (GANs) — Game-theoretic approach, best samples! But can be tricky and unstable to train, no inference queries.

Also recent work in combinations of these types of models! E.g. Adversarial Autoencoders (Makhanzi 2015) and PixelVAE (Gulrajani 2016)

What is exponential family

Exponential family comprises a set of flexible distribution ranging both continuous and discrete random variables. The members of this family have many important properties which merits discussing them in some general format. Many of the probability distributions that we have studied so far are specific members of this family:

- Gaussian: $\mathbb{R}^p$
- Multinomial: *categorical*
- Bernoulli: binary $\{0, 1\}$
- Binomial: counts of success/failure
- Von mises: sphere
- Gamma: $\mathbb{R}^+$
- Poisson: $\mathbb{N}^+$
- Laplace: $\mathbb{R}^+$
- Exponential: $\mathbb{R}^+$
- Beta: $(0, 1)$
- Dirichlet: Δ (Simplex)
- Weibull: $\mathbb{R}^+$
- Weishart: symmetric positive-definite matrices

What is exponential family

A pdf or pmf $p(\mathbf{x}|\boldsymbol{\theta})$, for $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, is said to be in the **exponential family** if it is of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \tag{9.1}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \tag{9.2}$$

where

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \tag{9.3}$$

$$A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) \tag{9.4}$$

Here $\boldsymbol{\theta}$ are called the **natural parameters** or **canonical parameters**, $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^d$ is called a vector of **sufficient statistics**, $Z(\boldsymbol{\theta})$ is called the **partition function**, $A(\boldsymbol{\theta})$ is called the **log partition function** or **cumulant function**, and $h(\mathbf{x})$ is the a scaling constant, often 1. If $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, we say it is a **natural exponential family**.

Why should we care about it

- It can be shown that, under certain regularity conditions, the exponential family is the only family of distributions with finite-sized sufficient statistics, meaning that we can compress the data into a fixed-sized summary without loss of information. This is particularly useful for online learning, as we will see later.
- The exponential family is the only family of distributions for which conjugate priors exist, which simplifies the computation of the posterior.
- The exponential family is at the core of generalized linear models and variational inference.

Scalar parameter

A single-parameter exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form

$$f_X(x \mid \theta) = h(x) \exp(\eta(\theta) \cdot T(x) - A(\theta))$$

where $T(x)$, $h(x)$, $\eta(\theta)$, and $A(\theta)$ are known functions.

An alternative, equivalent form often given is

$$f_X(x \mid \theta) = h(x) g(\theta) \exp(\eta(\theta) \cdot T(x))$$

Vector parameter

The definition in terms of one *real-number* parameter can be extended to one *real-vector* parameter

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)^T.$$

A family of distributions is said to belong to a vector exponential family if the probability density function (or probability mass function, for discrete distributions) can be written as

$$f_X(x \mid \boldsymbol{\theta}) = h(x) \exp\left(\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(x) - A(\boldsymbol{\theta})\right)$$

Vector parameter, vector variable

The vector-parameter form over a single scalar-valued random variable can be trivially expanded to cover a joint distribution over a vector of random variables. The resulting distribution is simply the same as the above distribution for a scalar-valued random variable with each occurrence of the scalar x replaced by the vector

$$\mathbf{x} = (x_1, x_2, \cdots, x_k)\,.$$

Note that the dimension k of the random variable need not match the dimension d of the parameter vector, nor (in the case of a curved exponential function) the dimension s of the natural parameter $\boldsymbol{\eta}$ and sufficient statistic $T(\mathbf{x})$.

The distribution in this case is written as

$$f_X(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x}) \exp\left(\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(\mathbf{x}) - A(\boldsymbol{\theta})\right)$$

Bernoulli distribution

As an example of a discrete exponential family, consider the binomial distribution with *known* number of trials n. The probability mass function for this distribution is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, 2, \ldots, n\}.$$

This can equivalently be written as

$$f(x) = \binom{n}{x} \exp\left(x \log\left(\frac{p}{1-p}\right) + n \log(1-p)\right),$$

which shows that the binomial distribution is an exponential family, whose natural parameter is

$$\eta = \log \frac{p}{1-p}.$$

This function of p is known as logit.

Normal distribution: unknown mean, known variance

As a first example, consider a random variable distributed normally with unknown mean μ and *known* variance σ^2. The probability density function is then

$$f_\sigma(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This is a single-parameter exponential family, as can be seen by setting

$$h_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$
$$T_\sigma(x) = \frac{x}{\sigma}$$
$$A_\sigma(\mu) = \frac{\mu^2}{2\sigma^2}$$
$$\eta_\sigma(\mu) = \frac{\mu}{\sigma}.$$

If σ = 1 this is in canonical form, as then $\eta(\mu) = \mu$.

Normal distribution: unknown mean and unknown variance

Next, consider the case of a normal distribution with unknown mean and unknown variance. The probability density function is then

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

This is an exponential family which can be written in canonical form by defining

$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)^{\mathrm{T}}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$T(x) = \left(x, x^2\right)^{\mathrm{T}}$$

$$A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \ln|\sigma| = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\ln\left|\frac{1}{2\eta_2}\right|$$

Conjugate prior

- A conjugate prior is one which, when combined with the likelihood and normalised, produces a posterior distribution which is of the same type as the prior.
- For example, if one is estimating the success probability of a binomial distribution, then if one chooses to use a beta distribution as one's prior, the posterior is another beta distribution.
- An arbitrary likelihood will not belong to the exponential family, and thus in general no conjugate prior exists. The posterior will then have to be computed by numerical methods.

Conjugate prior for exponential family

- In the case of a likelihood which belongs to the exponential family there exists a conjugate prior, which is often also in the exponential family.

Conjugate prior for exponential family

First, assume that the probability of a single observation follows an exponential family, parameterized using its natural parameter:

$$p_F(x \mid \boldsymbol{\eta}) = h(x)g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{T}(x)\right)$$

Then, for data $\mathbf{X} = (x_1, \ldots, x_n)$, the likelihood is computed as follows:

$$p(\mathbf{X} \mid \boldsymbol{\eta}) = \left(\prod_{i=1}^{n} h(x_i)\right) g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i)\right)$$

Then, for the above conjugate prior:

$$p_\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \propto g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi})$$

Conjugate prior for exponential family

where s is the dimension of $\boldsymbol{\eta}$ and $\nu > 0$ and $\boldsymbol{\chi}$ are hyperparameters (parameters controlling parameters). ν corresponds to the effective number of observations that the prior distribution contributes, and $\boldsymbol{\chi}$ corresponds to the total amount that these pseudo-observations contribute to the sufficient statistic over all observations and pseudo-observations. $f(\boldsymbol{\chi}, \nu)$ is a normalization constant that is automatically determined by the remaining functions and serves to ensure that the given function is a probability density function (i.e. it is normalized). $A(\boldsymbol{\eta})$ and equivalently $g(\boldsymbol{\eta})$ are the same functions as in the definition of the distribution over which π is the conjugate prior.

Conjugate prior for exponential family

We can then compute the posterior as follows:

$$
\begin{aligned}
p(\boldsymbol{\eta} \mid \mathbf{X}, \boldsymbol{\chi}, \nu) &\propto p(\mathbf{X} \mid \boldsymbol{\eta}) p_{\pi}(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) \\
&= \left(\prod_{i=1}^{n} h(x_i) \right) g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i) \right) f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \\
&\propto g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i) \right) g(\boldsymbol{\eta})^{\nu} \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \\
&\propto g(\boldsymbol{\eta})^{\nu+n} \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \left(\boldsymbol{\chi} + \sum_{i=1}^{n} \mathbf{T}(x_i) \right) \right)
\end{aligned}
$$

Conjugate prior for exponential family

- This shows that the update equations can be written simply in terms of the number of data points and the sufficient statistic of the data.

The update equations are as follows:

$$\begin{aligned}\boldsymbol{\chi}' &= \boldsymbol{\chi} + \mathbf{T}(\mathbf{X}) \\ &= \boldsymbol{\chi} + \sum_{i=1}^{n} \mathbf{T}(x_i) \\ \boldsymbol{\nu}' &= \boldsymbol{\nu} + n\end{aligned}$$