Exponential Family Distribution And Its Posterior Calculation

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Match 1, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Perform Bayesian learning by taking examples of Gaussian random variables
- Today's agenda
  - Discuss exponential family distribution
- Reference
  - Probabilistic Machine Learning: Advanced Topics: Section 2.3, 2.4, 3.4.5

Exponential Family Distribution

- Exponential family distribution is a class of distributions, which is of the form

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

- $\mathbf{x} \in \mathcal{X}^m$ is the random variable being modeled ($\mathcal{X}$ denotes some space e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\boldsymbol{\theta} \in \mathbb{R}^d$ Natural or canonical parameters defining the distribution
- $\phi(\mathbf{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Knowing this quantity suffices to estimate parameter $\boldsymbol{\theta}$ from $\mathbf{x}$
- $Z(\boldsymbol{\theta}) = \int h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \phi(\mathbf{x})] d\mathbf{x}$: Partition Function
- $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$: Log-partition function (also called cumulant function)
  - $Z(\boldsymbol{\theta})$ and $A(\boldsymbol{\theta})$ are functions of only natural parameters θ
- $h(\mathbf{x})$: Constant which doesn't depend on $\boldsymbol{\theta}$

Expressing a Distribution in Exponential Family Form

- Recall the form of exponential family distribution

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

- To write any exp-fam dist $p()$ in the above form, write it as $\exp(\log p())$

$$\begin{aligned}
\exp(\log \text{Binomial}(x|\mu)) &= \exp\left(\log\binom{N}{\mu}\mu^x(1-\mu)^{N-x}\right) \\
&= \exp\left(\log\binom{N}{\mu} + x\log\mu + (N-x)\log(1-\mu)\right) \\
&= \binom{N}{\mu}\exp\left(x\log\frac{\mu}{1-\mu} + N\log(1-\mu)\right)
\end{aligned}$$

- Now compare the resulting expression with (1), we have
  - $\theta = \log\frac{\mu}{1-\mu}$; $\phi(x) = x$; Constant $h(x) = \binom{N}{\mu}$; Log partition function $A(\theta) = -N\log(1-\mu)$

Scalar Gaussian as Exponential Family

- Let's try to write a univariate Gaussian in the exponential family form

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

- Recall the PDF of a univariate Gaussian

$$\begin{aligned}\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] &= \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]\\ &= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}^T\begin{bmatrix}x\\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]\end{aligned}$$

- Here

$$\boldsymbol{\theta} = \begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \quad \phi(x) = \begin{bmatrix}x\\ x^2\end{bmatrix}, \quad \text{and} \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2}\\ -\frac{1}{2\theta_2}\end{bmatrix}$$

- And

$$h(x) = \frac{1}{\sqrt{2\pi}} \quad A(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$$

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (https://en.wikipedia.org/wiki/Exponential_family)
- Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)
  - Student-t distribution
  - Mixture distributions (e.g., mixture of Gaussians)

Log-Partition Function

- Recall our exponential family distribution

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

- Log-partition func. $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) = \log \int h(\mathbf{x})\exp[\boldsymbol{\theta}^T \phi(\mathbf{x})]d\mathbf{x}$, is also called cumulant function
    - Derivatives of $A(\boldsymbol{\theta})$ can be used to generate the cumulants of sufficient statistics $\phi(\mathbf{x})$
- Assume scalar θ (thus $\phi(x)$ is also scalar). Show that first and second derivatives of $A(\theta)$ are

$$\frac{dA(\theta)}{d\theta} = \mathbb{E}_{p(x|\theta)}\left[\phi(x)\right] \tag{2}$$

$$\frac{d^2A(\theta)}{d\theta^2} = \mathbb{E}_{p(\mathbf{x}|\theta)}\left[\phi^2(x)\right] - \left[\mathbb{E}_{p(x|\theta)}\left[\phi(x)\right]\right]^2 \tag{3}$$

- Above result also holds when $\boldsymbol{\theta}$ and $\phi(\mathbf{x})$ are vector-valued (the “var” will be “covar”)

Proof of (2)

- We need to show

$$\frac{dA(\theta)}{d\theta} = \mathbb{E}_{p(x|\theta)}\left[\phi(x)\right]$$

- We begin as

$$\begin{aligned}
\frac{dA(\theta)}{d\theta} &\overset{(a)}{=} \frac{d}{d\theta}\log Z(\theta) = \frac{1}{Z(\theta)}\frac{d}{d\theta}Z(\theta) \overset{(b)}{=} \frac{1}{Z(\theta)}\frac{d}{d\theta}\left(\int h(x)\exp[\theta\phi(x)]dx\right) \\
&= \frac{1}{Z(\theta)}\int h(x)\frac{d}{d\theta}\left(\exp[\theta\phi(x)]\right)dx = \frac{1}{Z(\theta)}\int h(x)\phi(x)\exp[\theta\phi(x)]dx \\
&\overset{(c)}{=} \int \phi(x)p(x|\theta)dx = \mathbb{E}_{p(x|\theta)}\left[\phi(x)\right] \qquad (4)
\end{aligned}$$

- Equality (a) uses $A(\theta) = \log Z(\theta)$
- Equality (b) uses definition of $Z(\theta) = \int h(x)\exp[\theta\phi(x)]dx$
- Equality (c) is because $p(x|\theta) = \frac{1}{Z(\theta)}h(x)\exp[\theta\phi(x)]$

Maximal likelihood estimate for Exponential Family Distributions

- Assume data $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ drawn i.i.d. from an exponential family distribution with parameter $\boldsymbol{\theta}$
$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})\exp[\boldsymbol{\theta}^T\phi(\mathbf{x}) - A(\boldsymbol{\theta})]$$
- We want to calculate maximal likelihood estimate of $\boldsymbol{\theta}$
- Overall likelihood is a product of individual likelihoods

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right]\exp\left[\boldsymbol{\theta}^T\sum_{i=1}^{N}\phi(\mathbf{x}_i) - NA(\boldsymbol{\theta})\right] = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right]\exp\left[\boldsymbol{\theta}^T\phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right]$$

- To estimate $\boldsymbol{\theta}$ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N}\phi(\mathbf{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^{N}\phi(\mathbf{x}_i)$ does not grow with N, (same as the size of each $\phi(\mathbf{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply update the sufficient statistics as data comes
  - Useful in probabilistic inference with large-scale data sets and "online" parameter estimation

Maximal likelihood parameter estimation for exponential family

- Likelihood is of the form $p(\mathcal{D}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\theta}^T \phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right]$
- Log-likelihood is (ignoring constant w.r.t. $\boldsymbol{\theta}$)

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \phi(\mathcal{D}) - NA(\boldsymbol{\theta})$$

- Maximal likelihood estimation for exp-fam distributions can seen as doing moment-matching

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^T \phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right] \overset{(a)}{=} \phi(\mathcal{D}) - N\nabla_{\boldsymbol{\theta}}[A(\boldsymbol{\theta})] &= \phi(\mathcal{D}) - N\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\phi(\mathbf{x})\right] \\
&= \sum_{i=1}^{N} \phi(\mathbf{x}_i) - N\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\phi(\mathbf{x})\right]
\end{aligned}$$

  Equality (a) uses (2) in previous slide
- For maximal likelihood estimate of $\hat{\boldsymbol{\theta}}$, we must have

$$\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[\phi(\mathbf{x})\right] = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i) \tag{5}$$

- LHS in (5) – Expected moment; RHS in (5) – Empirical moment (computed using data)

Moment matching: an example

- Given data drawn $\mathcal{D} = \{x_1, \cdots, x_N\}$ i.i.d. from a scalar Gaussian $p(x) = \mathcal{N}(x|\mu, \sigma^2)$

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- For Gaussian $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$. We have $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a scalar Gaussian, note that (we have, two equations, two unknowns (μ and σ^2))

$$\mathbb{E}[x] = \mu \text{ and } \mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$$

- Same solution that we get by directly doing maximal likelihood estimate of Gaussian

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ and } \sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Bayesian inference for exponential family distributions

- Already saw that the total likelihood given N i.i.d. observations $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\theta}) \propto \exp\left[\boldsymbol{\theta}^T \phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right] \quad \text{where } \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\mathbf{x}_i)$$

- Let's choose the following prior (note: looks similar in terms of $\boldsymbol{\theta}$ within exp)

$$\boxed{p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\tau}_0) = h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T \boldsymbol{\tau}_0 - \nu_0 A(\boldsymbol{\theta}) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

  - Its natural parameters and sufficient statistics are given as $\begin{bmatrix} \boldsymbol{\tau}_0 \\ \nu_0 \end{bmatrix}$, and $\begin{bmatrix} \boldsymbol{\theta} \\ -A(\boldsymbol{\theta}) \end{bmatrix}$, respectively.
  - Its log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_{\boldsymbol{\theta}} h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T \boldsymbol{\tau}_0 - \nu_0 A(\boldsymbol{\theta})\right] d\boldsymbol{\theta}$ is a function of natural parameters. Ignoring $A_c(\nu_0, \boldsymbol{\tau}_0)$, we have

$$\boxed{p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\tau}_0) \propto h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T \boldsymbol{\tau}_0 - \nu_0 A(\boldsymbol{\theta})\right]}$$

- Comparing the prior's form with the likelihood, note that
  - ν_0 is like the number of "pseudo-observations" coming from the prior
  - $\boldsymbol{\tau}_0$ is the sufficient statistics of the pseudo-observations

Posterior calculation of exponential family (1)

- Our likelihood and prior are

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &\propto \exp\left[\boldsymbol{\theta}^T\phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right] \quad \text{where } \phi(\mathcal{D}) = \sum_{i=1}^{N}\phi(\mathbf{x}_i) \\ p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\tau}_0) &\propto h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T\boldsymbol{\tau}_0 - \nu_0 A(\boldsymbol{\theta})\right], \end{aligned}$$

  with its log partition function being $A_c(\nu_0, \boldsymbol{\tau}_0)$
- Posterior is thus

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\nu_0, \boldsymbol{\tau}_0) \\ &\propto \exp\left[\boldsymbol{\theta}^T\phi(\mathcal{D}) - NA(\boldsymbol{\theta})\right] \times h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T\boldsymbol{\tau}_0 - \nu_0 A(\boldsymbol{\theta})\right] \\ &\propto h(\boldsymbol{\theta})\exp\left[\boldsymbol{\theta}^T(\phi(\mathcal{D}) + \boldsymbol{\tau}_0) - (N + \nu_0)A(\boldsymbol{\theta})\right], \end{aligned}$$

- Natural parameters of posterior are $\begin{bmatrix} \boldsymbol{\tau}_0 + \phi(\mathcal{D}) \\ \nu_0 + N \end{bmatrix}$
  - Log partition function of posterior is therefore $A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))$

Posterior calculation of exponential family (2)

- Every exponential family likelihood has a conjugate prior with the form above
- Posterior's hyperparameters $\tau_0^{'}, \nu_0^{'}$ obtained by adding "stuff" to prior's hyperparams

$$\nu_0^{'} \leftarrow \nu_0 + N$$
$$\boldsymbol{\tau}_0^{'} \leftarrow \boldsymbol{\tau}_0 + \phi(\mathcal{D})$$

- $\nu_0^{'}$: Number of hypothetical-observations plus number of actual observations
- $\boldsymbol{\tau}_0^{'}$ Sufficient -statistics of hypothetical observations plus sufficient-statistics of actual observations

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exponential family distributions make parameter updates very simple
- Other quantities such as posterior predictive distribution can be computed in closed form
- Useful in designing generative models for unsupervised learning