Exponential Families and Generalized Linear Models

Seungjin Choi

Department of Computer Science
POSTECH, Korea
seungjin@postech.ac.kr

Introduction

- The exponential family of distributions, a family that includes the Gaussian, binomial, multinomial, Poisson, gamma, Rayleigh and beta distributions, as well as many others.
- In the conditional setting, in which we have a directed model, $X \rightarrow Y$, with X and Y observed, and with Y having an exponential family distribution for each value of X. To parameterize this conditional distribution we introduce a class of models known as generalized linear models (GLIM's).
- GLIM's retain an important role for lineariy, while introducing appropriate nonlinearities so as to cope with the idiosyncracies of the particular exponential family distribution at hand.

Outline

- The exponential family
  - Moments
  - Sufficiency
  - IID sampling
  - Maximum likelihood estimates
  - Kullback-Leibler divergence
- Generalized linear models

The exponential family

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

- η : natural parameter
- $T(X)$: sufficient statistic
- $h(x)$ simply reflects the underlying measure with respect to which $p(x|\eta)$ is a density.
- The logarithm of a normalization factor

$$\begin{aligned} A(\eta) &= \log \int h(x) \exp\{\eta^T T(x)\} dx \\ \Longrightarrow p(x|\eta) &= \tfrac{1}{Z(\eta)} h(x) \exp\{\eta^T T(x)\} \quad \text{where } A(\eta) = \log Z(\eta) \end{aligned}$$

EXAMPLE: The Bernoulli distribution

$$\begin{aligned} p(x|\pi) &= \pi^x(1-\pi)^{1-x} \\ &= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \end{aligned}$$

The Bernoulli distribution is an exponential family distribution with :

$$\begin{aligned} \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ T(X) &= x \\ h(x) &= 1 \\ A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}). \end{aligned}$$

The relationship between η and π is invertible. We can have the logistic function,

$$\pi = \frac{1}{1+e^{-\eta}}.$$

EXAMPLE: The Poisson distribution

$$\begin{aligned} p(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{1}{x!}\exp\{x\log\lambda - \lambda\} \end{aligned}$$

The Poisson distribution is an exponential family distribution with :

$$\begin{aligned} \eta &= \log\lambda \\ T(X) &= x \\ h(x) &= \frac{1}{x!} \\ A(\eta) &= \lambda = e^{\eta}. \end{aligned}$$

We can obviously invert the relationship between η and λ:

$$\lambda = e^{\eta}.$$

EXAMPLE: The Gaussian distribution

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{\sigma^2}\mu^2 - \ln\sigma\right\} \end{aligned}$$

The Gaussian distribution is in the exponential family form, with :

$$\begin{aligned} \eta &= \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \\ T(X) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} \\ h(x) &= \frac{1}{\sqrt{2\pi}} \\ A(\eta) &= \frac{\mu}{\sigma^2} + \ln\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\ln\eta_2). \end{aligned}$$

EXAMPLE: The multinomial distribution

$$\begin{aligned} p(x|\pi) &= \frac{n!}{x_1!x_2!\cdots x_m!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_m^{x_m} \\ &= \frac{n!}{x_1!x_2!\cdots x_m!}\exp\left\{\sum_{i=1}^{m} x_i\ln\pi_i\right\} \end{aligned}$$

★ Problem : $A(\eta) = 0$ and $\sum_{i=1}^{m}\pi_i = 1$
⇒ To achieve a full rank representation for the multinomial, we parameterize the distribution using the first $m-1$ components of π, $\pi_m = 1 - \sum_{i=1}^{m-1}\pi_i$.

$$\begin{aligned} p(x|\pi) &= \frac{n!}{x_1!x_2!\cdots x_m!}\exp\left\{\sum_{i=1}^{m-1} x_i\ln\pi_i + \left(n-\sum_{i=1}^{m-1}x_i\right)\ln\left(1-\sum_{i=1}^{m-1}\pi_i\right)\right\} \\ &= \frac{n!}{x_1!x_2!\cdots x_m!}\exp\left\{\sum_{i=1}^{m-1}\ln\left(\frac{\pi_i}{1-\sum_{i=1}^{m-1}\pi_i}\right)x_i + n\ln\left(1-\sum_{i=1}^{m-1}\pi_i\right)\right\} \\ &= h(x)\exp\left\{\sum_{i=1}^{m-1}\eta_i x_i - A(\eta)\right\} \end{aligned}$$

We obtain the softmax function, $\pi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{m} e^{\eta_j}}$.

Moments

- We can obtain moments of the distribution by taking derivatives of the log normalization function $A(\eta)$.

$$\begin{aligned}
\frac{dA}{d\eta} &= \frac{d}{d\eta}\left\{\log \int \exp\{\eta T(x)\}h(x)dx\right\} = \frac{\int T(x)\exp\{\eta T(x)\}h(x)dx}{\int \exp\{\eta T(x)\}h(x)dx} \\
&= \int T(x)\exp\{\eta^T T(x) - A(\eta)\}h(x)dx = ET(X) \triangleq \mu \\
\frac{d^2a}{d\eta^2} &= \int T(x)\exp\{\eta T(x) - A(\eta)\}(T(x) - a'(\eta)[=ET(X)])h(x)dx \\
&= \int T^2(x)\exp\{\eta T(x) - A(\eta)\}h(x)dx - ET(x)\int T(x)\exp\{\eta T(x) - A(\eta)\}h(x)dx \\
&= ET^2(x) - (ET(x))^2 = \mathrm{Var}[T(X)] > 0
\end{aligned}$$

- For a convex function there is necessarily a one-to-one relationship between the argument to the function and the first derivative of the function. Hence the mapping from η to μ is invertible.
- A distribution in the exponential family can be parameterized not only by η-the canonical parameterization-but also by μ-the moment parameterization.

Moments : Example

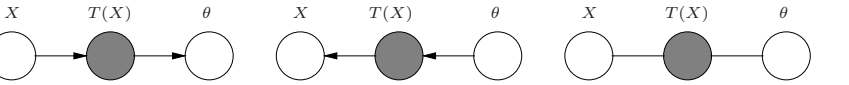
- the Bernoulli distribution : $A(\eta) = \log(1 + e^{\eta})$

$$\begin{aligned}
\frac{dA}{d\eta} &= \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} = \mu \\
\frac{d^2A}{d\eta^2} &= \frac{d\mu}{d\eta} = \mu(1-\mu)
\end{aligned}$$

- the univariate Gaussian distribution : $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$ where $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/2\sigma^2$

$$\begin{aligned}
\frac{dA}{d\eta_1} &= -\frac{\eta_1}{2\eta_2} = \frac{\mu/\sigma^2}{1/\sigma^2} = \mu \\
\frac{d^2A}{d\eta_1^2} &= -\frac{1}{2\eta_2} = \sigma^2
\end{aligned}$$

Sufficiency



(a) Bayesian approach $p(\theta|T(x),x)=p(\theta|T(x))$

(b) Frequentist approach $p(x|T(x),\theta)=p(x|T(x))$

(c) same semantics as (a) and (b) $p(x,T(x),\theta)=\psi_1(T(x),\theta)\psi_2(x,T(x))$

- Sufficiency characterizes what is essential in a data set, or, alternatively, what is inessential and can therefore be thrown away.
- A statistic is a function of a random variable.
- Let X be a random variable and let $T(X)$ be a statistic. Suppose that the distribution of X depends on a parameter θ.
  The intuitive notion of sufficiency is that $T(X)$ is sufficient for θ if there is no information in X regarding θ beyond that in $T(X)$. That is, having observed $T(X)$, we can throw away X for the purposes of inference with respect to θ.
- $T(x)$ is a deterministic function of x. $\Rightarrow p(x|\theta) = g(T(x),\theta)h(x,T(x))$
  The exponential family $\Rightarrow p(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$

Maximum likelihood

- **IID sampling**
  Suppose that we have a collection of N independent random variables, $\mathcal{D} = (X_1, X_2, \ldots, X_N)$, characterized by the same exponential family density.

$$\begin{aligned}
p(\mathcal{D}|\eta) &= \prod_{n=1}^{N} h(x_n)\exp\{\eta^T T(x_n) - A(\eta)\} \\
&= \left(\prod_{n=1}^{N} h(x_n)\right)\exp\left\{\eta^T \sum_{n=1}^{N} T(x_n) - NA(\eta)\right\}
\end{aligned}$$

- **Maximum likelihood estimates**

$$l(\eta|\mathcal{D}) = \log p(\mathcal{D}|\eta) = \log\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^T\left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta)$$

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta) \Rightarrow \nabla_\eta A(\eta) = \frac{1}{N}\sum_{n=1}^{N} T(x_n) = \hat{\mu}_{ML}$$

Maximum likelihood and the KL divergence

Define the empirical distribution, $\tilde{p}(x) \triangleq \frac{1}{N}\sum_{n=1}^{N}\delta(x, x_n)$.
Obtain the log likelihood by computing a cross-entropy between the empirical distribution and the model (discrete case) :

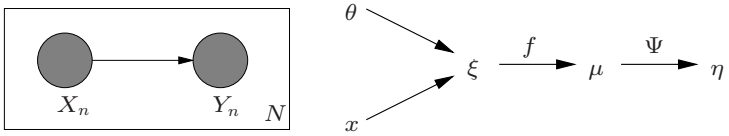
$$\begin{aligned}\sum_x \tilde{p}(x)\log p(x|\theta) &= \sum_x \frac{1}{N}\sum_{n=1}^{N}\delta(x,x_n)\log p(x|\theta)\\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_x \delta(x,x_n)\log p(x|\theta) = \frac{1}{N}\sum_{n=1}^{N}\log p(x_n|\theta)\\ &= \frac{1}{N}l(\theta|\mathcal{D})\end{aligned}$$

Calculate the KL divergence between the empirical distribution and the model :

$$\begin{aligned}D(\tilde{p}(x)\|p(x|\theta)) &= \sum_x \tilde{p}(x)\log\frac{\tilde{p}(x)}{p(x|\theta)}\\ &= \sum_x \tilde{p}(x)\log\tilde{p}(x) - \sum_x \tilde{p}(x)\log p(x|\theta)\\ &= \sum_x \tilde{p}(x)\log\tilde{p}(x) - \frac{1}{N}l(\theta|\mathcal{D})\end{aligned}$$

Minimizing the KL divergence to the empirical distribution is equivalent to maximizing the likelihood.

Generalized linear models



A GLIM makes three assumptions regarding the form of the conditional probability distribution $p(y|x)$.

- The observed input x is assumed to enter into the model via a linear combination $\xi = \theta^T x$,
- The conditional expectation μ is represented as a function $f(\xi) = f(\theta^T x)$ of the linear combination ξ, where f is known as the response function,
  (linear regression: the identity function, linear classification: the logistic function or the cumulative Gaussian)
- The observed output y is assumed to be characterized by an exponential family distribution with conditional expectation μ.
  (linear regression: Gaussian, linear classification: Bernoulli or multinomial)

The specification of a GLIM

- **The choice of exponential family distribution** is generally rather strongly constrained by the nature of the data Y : class labels by Bernoulli or multinomial distributions, counts by the Poisson distribution, intervals by the exponential or gamma distributions, etc.
- **The choice of the response function** is constrained by the conditional expectation : $(0, 1)$ for Bernoulli and multinomial distributions, $(0, \infty)$ for gamma distribution, etc. In general for any given distribution there are many possible choices of response function.
- However, there is a particular response function - the canonical response function - that is uniquely associated with a given exponential family distribution and has some appealing mathematical properties. In particular, if we assume that $\xi = \eta$, or equivalently that $f(\cdot) = \psi^{-1}(\cdot)$, we obtain the canonical response function.
- The modeled values $\mu = f(\eta)$ are guaranteed to be possible values of the conditional expectation. ($\because f(\eta) = \psi^{-1}(\eta) = a'(\eta) = E[Y|\eta]$)
- Some examples

Gaussian	Bernoulli	multinomial	Poisson	gamma
$\mu = \eta$	$\mu = 1/(1+e^{-\eta})$	$\mu_i = \eta_i / \sum_j e^{\eta_i}$	$\mu = e^{\eta}$	$\mu = -\eta^{-1}$

Maximum likelihood estimation

- Consider an IID data set, $\mathcal{D} = \{(x_n, y_n); n = 1, \ldots, N\}$.
- The loglikelihood for GLIM models is :

$$\begin{aligned}l(\theta|\mathcal{D}) &= \log\left(\sum_{n=1}^{N} h(y_n)\exp\{\eta_n y_n - A(\eta_n)\}\right)\\ &= \sum_{n=1}^{N}\log h(y_n) + \sum_{n=1}^{N}(\eta_n y_n - A(\eta_n))\\ &\text{where} \quad \eta_n = \psi(\mu_n), \quad \mu_n = f(\xi_n) \quad \text{and} \quad \xi_n = \theta^T x_n\end{aligned}$$

- In the case of $\eta_n = \theta^T x_n$, the log likelihood simplifies:

$$\begin{aligned}l(\theta|\mathcal{D}) &= \sum_{n=1}^{N}\log h(y_n) + \sum_{n=1}^{N}(\theta^T x_n y_n - A(\eta_n))\\ &= \sum_{n=1}^{N}\log h(y_n) + \theta^T\sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} A(\eta_n))\end{aligned}$$

  $\Rightarrow$ The sum $\sum_{n=1}^{N} x_n y_n$ is a sufficient statistic for θ.
- Let us now calculate the gradient of the log likelihood :

$$\begin{aligned}\nabla_\theta l &= \sum_{n=1}^{N}\frac{dl}{d\eta_n}\nabla_\theta\eta_n = \sum_{n=1}^{N}(y_n - a'(\eta_n))\nabla_\theta\eta_n = \sum_{n=1}^{N}(y_n-\mu_n)\frac{d\eta_n}{d\mu_n}\frac{d\mu_n}{d\xi_n}x_n\\ &= \sum_{n=1}^{N}(y_n-\mu_n)x_n\end{aligned}$$

Parameter Estimation

- An on-line algorithm

$$\theta^{(t+1)} = \theta^{(t)} + \rho(y_n - \mu_n^{(t)})x_n \quad \text{where} \quad \mu_n^{(t)} = f({\theta^{(t)}}^T x_n)$$

- A batch algorithm : iteratively reweighted least squares (IRLS) algorithm

$$\begin{aligned}
\nabla_\theta l &= \textstyle\sum_n (y_n - \mu_n) x_n = X^T(y-\mu) \\
H &= -\textstyle\sum_n \frac{d\mu_n}{d\eta_n} x_n x_n^T = -X^T W X \quad \text{where} \quad W \triangleq diag\left\{\frac{d\mu_1}{d\eta_1}, \frac{d\mu_1}{d\eta_2}, \ldots, \frac{d\mu_1}{d\eta_N}\right\} \\
\Rightarrow \theta^{(t+1)} &= \theta^{(t)} - H^{-1}\nabla_\theta J \\
&= \theta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (y - \mu^{(t)}) \\
&= (X^T W^{(t)} X)^{-1} \left[X^T W^{(t)} X \theta^{(t)} + X^T (y - \mu^{(t)})\right] \\
&= (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)} \quad \text{where} \quad z^{(t)} = \eta + [W^{(t)}]^{-1}(y - \mu^{(t)})
\end{aligned}$$