The AI-ML and Data Science Club of IITH

Exponential Family

Contents

Compiled by : Divyanshu Bhatt

1. Introduction

The Exponential Family is a class of probability distributions that exhibit a specific mathematical structure and share common properties. It is defined by its characteristic form, which allows for efficient computation and bayseian inference by giving a closed-loop solution.

A distribution belongs to the Exponential Family if its probability density function (PDF) or probability mass function (PMF) can be expressed as:

$$p(x \mid \eta) = h(x) \exp\left(\langle \eta, \varphi(x) \rangle - A(\eta)\right)$$

where η is called as the natural parameter or canonical parameter, $\varphi(x)$ is the sufficient statistics and $A(\eta)$ is called log partition function. The above form is derived from the Fisher Neyman Theorem which is explained below.

Exponential family can also be written in the form

$$p(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp\left(\langle \eta, \varphi(x) \rangle\right)$$

where $Z(\eta) = \exp(A(\eta))$. We can think of $Z(\eta)$ as the normalising constant for the probability distribution as it is independent of the random variable i.e.

$$Z(\eta) = \int_x h(x) \exp\left(\eta^T \varphi(x)\right) dx$$

2. Fisher Neyman Theorem

Let a random variable X have the probability density function $p(x \mid \eta)$ that can be written in the form

$$p(x \mid \eta) = h(x) g_\eta(\varphi(x))$$

Here, $h(x)$ is a distribution independent of η. The function g_η captures all the dependence of the probability distribution on η via a function $\varphi(x)$ which is called as sufficient statistics.

Sufficient statistics carry all the information about η which is needed to make inference. They condense the data into a reduced form while retaining the essential information needed for making inferences about the parameter.

Exponential Family as defined above is a special case of Fisher Neyman's equation where

$$g_\eta(\varphi(x)) = \exp\left(\langle \eta, \varphi(x) \rangle - A(\eta)\right)$$

3. Properties

The two main properties of the exponential family that helps in bayesian inference are

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

$$\frac{d^2A}{d\eta^2} = \mathbb{V}_{p(x|\eta)}\left[\varphi(x)\right]$$

where $\mathbb{E}$ and $\mathbb{V}$ represents the expectation and the variance. As variance is always non-negative, we get the double differentiation of the log partition function to be always non-negative i.e.

$$\frac{d^2A}{d\eta^2} \geq 0$$

Thus, $A(\eta)$ is a convex function.

For multivariable case the first formula remains the same but the second formula changes because we can differentiate with respect to some other dimension also i.e.

$$\frac{\partial A}{\partial \eta_i} = \mathbb{E}[\varphi_i(x)] \Rightarrow \nabla A = \mathbb{E}[\Phi(x)]$$

$$\frac{\partial A}{\partial \eta_i \partial \eta_j} = \mathbb{E}[\varphi_i(x)\varphi_j(x)] - \mathbb{E}[\varphi_i(x)]\mathbb{E}[\varphi_j(x)] \Rightarrow \nabla^2 A = \mathrm{Cov}(\Phi(x))$$

3.1. **Proofs.**

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

Proof. Writing the function A in the log partition form and diffentiating

$$\begin{aligned} A(\eta) &= \log Z(\eta) \\ \Rightarrow \frac{dA}{d\eta} &= \frac{1}{Z(\eta)}\frac{dZ}{d\eta} \end{aligned}$$

Expanding $Z(\eta)$ as the integral and differentiating the terms inside the integral with respect to η

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{1}{Z(\eta)}\frac{d}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)dx\right) \\ &= \frac{1}{Z(\eta)}\int h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)dx \\ &= \int \varphi(x)\frac{1}{Z(\eta)}h(x)\exp\left(\eta^T\varphi(x)\right) \\ &= \int \varphi(x)p(x \mid \eta)dx \\ &= \mathbb{E}_{p(x|\eta)}\left[\varphi(x)\right] \end{aligned}$$

□

$$\frac{d^2A}{d\eta^2} = \mathbb{V}_{p(x|\eta)}\left[\varphi(x)\right]$$

Proof. Double differentiating the function

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{1}{Z(\eta)}\frac{d}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)dx\right) \\ \frac{d^2A}{d\eta^2} &= \frac{d}{d\eta}\left(\frac{1}{Z(\eta)}\int h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)dx\right) \\ &= \frac{1}{Z(\eta)^2}\left(Z(\eta)\frac{d}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)\right) - \frac{dZ}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)\right)\right) \\ &= \frac{1}{Z(\eta)}\frac{d}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)dx\right) - \frac{1}{Z(\eta)^2}\frac{dZ}{d\eta}\left(\int h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)dx\right) \\ &= \int \frac{1}{Z(\eta)}h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)^2dx - \frac{1}{Z(\eta)}\frac{dZ}{d\eta}\left(\int \frac{1}{Z(\eta)}h(x)\exp\left(\eta^T\varphi(x)\right)\varphi(x)dx\right) \\ &= \int \varphi(x)^2p(x \mid \eta)dx - \underbrace{\frac{1}{Z(\eta)}\frac{dZ}{d\eta}}_{\texttt{I}}\int \varphi(x)p(x \mid \eta)dx \end{aligned}$$

I term is the same term as the preivous property. Thus, it can be written in terms of expectation. The other terms are expectations of $\varphi(x)^2$ and $\varphi(x)$ respectively with respect to the distribution $p(x \mid \eta)$

$$\begin{aligned} \frac{d^2A}{d\eta^2} &= \mathbb{E}_{p(x|\eta)}[\varphi(x)^2] - \mathbb{E}_{p(x|\eta)}[\varphi(x)]\mathbb{E}_{p(x|\eta)}[\varphi(x)] \\ \Rightarrow &= \mathbb{V}[\varphi(x)] \end{aligned}$$

□

4. MLE Estimation

With the help of the above defined properties, we can derive a closed-form solution for the MLE estimate for the parameters η. Assume $\mathcal{D} = \{x_1, \ldots, x_N\}$ be independent and identically distributed samples from an exponential family. The MLE problem will be formulated as

$$\max_{\eta} \ \log p(\mathcal{D} \mid \eta)$$

Doing some mathematical manipulations for getting a closed form solution

$$\begin{aligned}
p(\mathcal{D} \mid \eta) &= \prod_{n=1}^{N} p(x_n \mid \eta) \\
&= \prod_{n=1}^{N} h(x_n) \exp\left(\eta^T \varphi(x_n) - A(\eta)\right) \\
&= \left(\prod_{n=1}^{N} h(x_n)\right) \exp\left(\eta^T \sum_{n=1}^{N} \varphi(x_n) - NA(\eta)\right) \\
\log p(\mathcal{D} \mid \eta) &= \underbrace{\sum_{n=1}^{N} \log h(x_n)}_{\text{constant wrt } \eta} + \eta^T \sum_{n=1}^{N} \varphi(x_n) - NA(\eta)
\end{aligned}$$

As the first term is constant with respect to η we can remove it from the optimisation problem.

$$\max_{\eta} \ \underbrace{\eta^T \sum_{n=1}^{N} \varphi(x_n)}_{\texttt{I}} - \underbrace{NA(\eta)}_{\texttt{II}}$$

`I` is linear in η. As $A(\eta)$ is a convex function, it makes `II` a concave function, resulting in the whole optimisation problem to be a concave optimisation problem. Thus, making the gradient 0 will give the global optimum of the MLE problem.

$$\begin{aligned}
\nabla_\eta \log p(\mathcal{D} \mid \eta) &= \nabla_\eta \left(\eta^T \sum_{n=1}^{N} \varphi(x) - NA(\eta)\right) \\
&= \sum_{n=1}^{N} \varphi(x_n) - N\nabla_\eta A(\eta)
\end{aligned}$$

From the first property we can write the gradient in the form of expectation.

$$\begin{aligned}
\nabla_\eta \log p(\mathcal{D} \mid \eta) &= \sum_{n=1}^{N} \varphi(x_n) - N\mathbb{E}_{p(x|\eta)}\left[\varphi(x)\right] \\
\Rightarrow \mathbb{E}_{p(x|\eta)}\left[\varphi(x)\right] &= \frac{1}{N} \sum_{n=1}^{N} \varphi(x_n)
\end{aligned}$$

As left side is a function of η we can calculate the optimum value by solving the above equation. We are able to calculate the MLE solution of the parameters by only using $\varphi(x)$. This is the reason we say that $\varphi(x)$ i.e. the sufficient statistics captures all the important information of the parameters.

5. Bayesian Inference

For performing Bayesian inference, we need to consider a prior on the parameter. We are choosing it such that it is conjugate with the exponential family i.e.

$$p(\eta \mid \gamma, \tau) = h(\eta) \exp\left(\eta^T \tau - \gamma A(\eta) - A_c(\gamma, \tau)\right)$$

where γ and τ are new introduced parameter, $A(\eta)$ is the same function as defined in the exponential family, $A_c(\gamma, \tau)$ is introduced for normalising the above distribution i.e.

$$\exp(A_c(\gamma, \tau)) = \int_\eta h(\eta) \exp\left(\eta^T \tau - \gamma A(\eta)\right)$$

The posterior after observing the data $\mathcal{D}$ can be written as

$$\begin{aligned} p(\eta \mid \mathcal{D}) &\propto p(\eta) p(\mathcal{D} \mid \eta) \\ &\propto h(\eta) \exp\left(\eta^T \tau - \gamma A(\eta)\right) \exp\left(\eta^T \sum_{n=1}^{N} \varphi(x_n) - N A(\eta)\right) \\ &\propto h(\eta) \exp\left(\eta^T \left(\tau + \sum_{n=1}^{N} \varphi(x_n)\right) - (\gamma + N) A(\eta)\right) \end{aligned}$$

On comparing the above equation with the prior equation, we can see that the prior distribution hyperparmaters got updated as given below resulting in the posterior distribution

$$\begin{aligned} \tau &\leftarrow \tau + \sum_{n=1}^{N} \varphi(x_n) \\ \gamma &\leftarrow \gamma + N \end{aligned}$$

So, we got a closed-form solution for updating the prior of the parameters. Now, calculating the closed-form solution for the posterior-predictive distribution i.e.

$$\begin{aligned} p(\mathcal{D}' \mid \mathcal{D}) &= \int p(\mathcal{D}' \mid \eta) p(\eta \mid \mathcal{D}) d\eta \\ &= \int \left(\prod_{n=1}^{N'} h(x'_n)\right) \exp\left(\eta^T \sum_{n=1}^{N'} \varphi(x'_n) - N' A(\eta)\right) \\ &\qquad h(\eta) \exp\left(\eta^T \left(\tau + \sum_{n=1}^{N} \varphi(x_n)\right) - (\nu + N) A(\eta) - A_c\left(\nu + N, \tau + \sum_{n=1}^{N} \varphi(x_n)\right)\right) d\eta \end{aligned}$$

Abbreviating $\sum_{n=1}^{N} \varphi(x_n)$ as $\varphi(\mathcal{D})$ and similarly, $\varphi(\mathcal{D}') = \sum_{n=1}^{N'} \varphi(x'_n)$

$$\begin{aligned} p(\mathcal{D}' \mid \mathcal{D}) &= \frac{\prod_{n=1}^{N'} h(x'_n)}{\exp\left(A_c\left(\gamma + N, \tau + \varphi(\mathcal{D})\right)\right)} \int h(\eta) \exp\left(\eta^T \left(\tau + \varphi(\mathcal{D}) + \varphi(\mathcal{D}')\right) - \left(\gamma + N + N'\right) A(\eta)\right) \\ &= \frac{\prod_{n=1}^{N'} h(x'_n)}{\exp\left(A_c\left(\gamma + N, \tau + \varphi(\mathcal{D})\right)\right)} \exp\left(A_c(\tau + \varphi(\mathcal{D}) + \varphi(\mathcal{D}'), \gamma + N + N')\right) \\ &= \prod_{n=1}^{N'} h(x'_n) \frac{\exp\left(A_c(\tau + \varphi(\mathcal{D}) + \varphi(\mathcal{D}'), \gamma + N + N')\right)}{\exp\left(A_c\left(\gamma + N, \tau + \varphi(\mathcal{D})\right)\right)} \end{aligned}$$

Thus, we obtained a closed form solution for the predictive probability distribution which is the ratio of the exponential of log partition functions one with $N + N'$ datapoints and other with N datapoints

6. Exponential Dispersion Family

Exponential Dispersion Family is an extenstion to Exponential Family in which we introduce a new parameter σ called dispersion parameter.

$$p(x \mid \eta, \sigma^2) = h(y, \sigma^2) \exp\left(\frac{\eta^T x - A(\eta)}{\sigma^2}\right)$$

If the value of σ is fixed then it is the standard exponential family. Adding this parameter doesn't change the properties much

$$\frac{dA}{d\eta} = \mathbb{E}_{p(x \mid \eta)}\left[\varphi(x)\right]$$

$$\frac{d^2A}{d\eta^2} = \frac{1}{\sigma^2}\mathbb{V}_{p(x|\eta)}\left[\varphi(x)\right]$$

7. Generalised Linear Models

The linear models such as Linear Regression and Logistic Regression models have constraints on their output i.e. $\mathbb{R}^2$ and binary values respectively. For having some other constraints in the output, example while predicting the weight of a person given other features, we don't want output to be negative. Applying Linear Regression directly won't be that effective as it can produce negative values also.

Generalised Linear Models (GLMs) can be used to model having non-negative or integer outputs. GLMs model the response using an exponential dispersion family i.e.

$$p(y \mid \eta, \sigma^2) = h(y, \sigma^2)\exp\left(\frac{\eta y - A(\eta)}{\sigma^2}\right)$$

here, we considered $y \in \mathbb{R}$ for simiplicity which imples $\eta \in \mathbb{R}$. Moreover, here the sufficient statistics $\varphi(y) = y$ The natural parameter η is dependent on the input as

$$\eta = w^T x$$

7.1. Examples.

For Probablistic Linear Regression we have

$$\begin{aligned}
p(y \mid x, w, \sigma^2) &= \mathcal{N}(w^T x, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right) \\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{y^2}{2\sigma^2}\right)}_{h(y,\sigma^2)}\underbrace{\exp\left(\frac{w^T xy - \frac{(w^T x)^2}{2}}{\sigma^2}\right)}_{\exp\frac{\eta y - A(\eta)}{\sigma^2}}
\end{aligned}$$

Thus, linear regression is a type of GLM with $A(\eta) = \frac{\eta^2}{2}$ and $h(y, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{y^2}{2\sigma^2}$

For Bionomial Regression i.e. $y \in \{0, 1, \ldots, N\}$

$$\begin{aligned}
p(y \mid x, w) &= \binom{N}{y}\pi^y(1-\pi)^{N-y} \\
&= \binom{N}{y}\exp\left(y\log\pi + (N-y)\log(1-\pi)\right) \\
&= \underbrace{\binom{N}{y}}_{h(y)}\exp\left(\underbrace{y\log\left(\frac{\pi}{1-\pi}\right)}_{\eta y} + \underbrace{N\log(1-\pi)}_{A(\eta)}\right)
\end{aligned}$$

Thus, binomial regression is another type of GLM where $\pi = \sigma(w^T x)$, $\eta = \log\left(\frac{\pi}{1-\pi}\right) = w^T x$, $h(y) = \binom{N}{y}$ and $A(\eta) = N\log(1 + \exp\eta) = -N\log(1-\pi)$

Above, we used the standard exponential family and the $\sigma(\cdot)$ here represents the sigmoid function.

7.2. MLE Estimation.

$$\begin{aligned}
\mathcal{L} = \log p(Y \mid \eta) &= \log\prod_{n=1}^{N} h(y_N)\exp(y_n w^T x_n - A(\eta_n)) \\
&= \sum_{n=1}^{N}\log h(y_n) + w^T\sum_{n=1}^{N} y_n x_n - \sum_{n=1}^{N} A(\eta_n)
\end{aligned}$$

Differntiating with respect to the weights

$$\begin{aligned}\frac{d\mathcal{L}}{dw} &= \sum_{n=1}^{N} y_n x_n - \sum_{n=1}^{N} \frac{dA}{d\eta_n}\frac{d\eta_n}{dw} \\ &= \sum_{n=1}^{N} y_n x_n - \sum_{n=1}^{N} \mathbb{E}_{p(y|\eta)}[y] x_n \\ &= \sum_{n=1}^{N} \left(y_n - \mathbb{E}_{p(y|\eta)}\right) x_n\end{aligned}$$

Equating the above equation to 0 we will get the global optimum value because the function is concave in η which implies it is concave in w as there exists a linear relation between the two.

8. Limitations

1) **Distributional Assumption**: We need to specify a probability distribution for the output variable which is assumed to belong to the exponential family. Thus, it is not robust to capture data coming from any distribution.
2) **Model Misspecification**: If the chosen GLM does not accurately represent the true underlying relationship between the predictors and the response, the model may produce biased or misleading results.
3) **Linearity Assumption**: It assumes the presence of a linear relationship between the predictors and the linear prediction. However, achieving this linear relationship often requires preprocessing the features, which can be challenging in practice.
4) **Effect of Outliers**: GLMs gets highly affected by outliers and there predictions change drastically because of them.

9. Questions

9.1. Subjective.

1) What are the key assumptions underlying the use of GLMs?
2) Show that Bernoulli distribution belongs to the exponential family even though it doesn't have the function exp in it.
3) Under what assumption does Exponential Dispersion Family becomes the Standard Exponential Family.
4) Prove the properties of Exponential Dispersion Family and how are they different from the Standard Exponential Family results?

9.2. Objective.

1) What is the main difference between Generalized Linear Models (GLMs) and ordinary linear regression?
   - GLMs can handle non-normal response variables
   - GLMs require the assumption of linearity
   - GLMs are only applicable to binary data
   - GLMs use a different estimation method
2) Which of the following is NOT a commonly used distribution from the Exponential Family?
   - Beta distribution
   - Poisson distribution
   - Weibull distribution
   - Chi-squared distribution