# The Exponential Family

Elif Yılmaz

# Exponential family

The exponential family of distributions over x, given parameters $\eta$, is defined to be the set of distributions of the form

$$p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$$

where x may be scalar or vector, and may be discrete or continuous, $\eta$ are the natural parameters of the distribution, and u(x) is some function of x. The function $g(\eta)$ can be interpreted as the coefficient ensuring the normalized distribution and it satisfies

$$g(\eta)\int h(x)exp\{\eta^T u(x)\}dx = 1 \ for \ continuous \ variable \ x,$$

and

$$g(\eta)\sum h(x)exp\{\eta^T u(x)\}dx = 1 \ for \ discrete \ variable \ x.$$

Exponential families include many of the most common distributions. Among many others, exponential families includes normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart, inverse Wishart and geometric distributions.

A number of common distributions are exponential families, but only when certain parameters are fixed and known. For example:

- binomial (with fixed number of trials)
- multinomial (with fixed number of trials)
- negative binomial (with fixed number of failures)

Consider first the Bernoulli distribution

$$p(x|\mu) = Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

When we take the exponential of the logarithm for the right-hand side, we have

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

If we compare this with $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$ , we can identify

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

which we can solve for $\mu$ to give $\mu = \sigma(\eta)$, where

$$\sigma(\eta) = \frac{1}{1 + exp(-\eta)}$$

is called the logistic sigmoid function.

Thus we can write the Bernoulli distribution using the standard representation $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$ in the form

$$p(x|\eta) = \sigma(-\eta)exp(\eta x)$$

where we have used $1 - \sigma(\eta) = \sigma(-\eta)$.
Comparison with $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$ shows that

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = \sigma(-\eta).$$

For a single observation x, when considering the multinomial distribution taking the form

$$p(x|\mu) = \prod_{k=1}^{M} \mu_k^{x_k} = exp\Big\{\sum_{k=1}^{M} x_k \ln \mu_k\Big\}$$

where $x = (x_1, ..., x_N)^T$.. When we write it in the standard representation, we get

$$p(x|\eta) = exp(\eta^T x)$$

where $\eta_k = \ln \mu_k$, and we have defined $\eta = (\eta_1, ..., \eta_M)^T$.

By comparing with $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$ , we have

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = 1.$$

Note that the parameters $\eta_k$ are not independent because the parameters $\mu_k$ have the constraint

$$\sum_{k=1}^{M} \mu_k = 1.$$

Therefore, the value of the remaining parameter is fixed for given any M - 1 of the parameters $\mu_k$. Note that these remaining parameters are still subject to the constraints

$$0 \leq \mu_k \leq 1 \;\; \mathit{and} \;\; \sum_{k=1}^{M-1} \mu_k \leq 1.$$

By using the constraint $\sum_{k=1}^{M} \mu_k = 1$, the multinomial distribution in this representation becomes

$$\mathit{exp}\Big\{ \sum_{k=1}^{M} x_k \ln \mu_k \Big\} = \mathit{exp}\Big\{ \sum_{k=1}^{M-1} x_k \ln \mu_k + \Big(1 - \sum_{k=1}^{M-1} x_k\Big) \ln \Big(1 - \sum_{k=1}^{M-1} \mu_k\Big)\Big\}$$

$$= \mathit{exp}\Big\{ \sum_{k=1}^{M-1} x_k \ln \Big(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\Big) + \ln \Big(1 - \sum_{j=1}^{M-1} \mu_j\Big)\Big\}$$

Now, we assign

$$\ln\left(\frac{\mu_k}{1-\sum_{j=1}^{M-1}\mu_j}\right)=\eta_k$$

When we solve this for $\mu_k$, we get

$$\mu_k=\frac{exp(\eta_k)}{1+\sum_{j=1}^{M-1}\eta_j}$$

This is called the softmax function or the normalized exponential.

In this representation, the multinomial distribution takes the form

$$p(x|\eta) = \Big(1 + \sum_{k=1}^{M-1} exp(\eta_k)\Big)^{-1} exp(\eta^T x)$$

This is the standard form of the exponential family, with parameter vector $\eta = (\eta_1, ..., \eta_{M-1})^T$ in which

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = \Big(1 + \sum_{k=1}^{M-1} exp(\eta_k)\Big)^{-1}$$

When we interested in univariate Gaussian distribution, we have

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} exp\Big\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\Big\}$$

which in the standard exponential family form

$$\eta = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$$

$$u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$h(x) = (2\pi)^{-1/2}$$

$$g(\eta) = (-2\eta_2)^{1/2} exp\Big(\frac{\eta_1^2}{4\eta_2}\Big)$$

# Maximum likelihood and sufficient statistics

Consider the problem of estimating the parameter vector $\eta$ in the general exponential family distribution $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$ by using the maximum likelihood method. By taking the gradient of both sides of $g(\eta)\int h(x)exp\{\eta^T u(x)\}dx = 1$ with respect to $\eta$, we have

$$\nabla g(\eta) \int h(x)exp(\eta^T u(x))dx$$

$$+\nabla g(\eta) \int h(x)exp(\eta^T u(x))u(x)dx = 0$$

Therefore, we get

$$-\frac{1}{g(\eta)}\nabla g(\eta) = g(\eta)\int h(x)exp(\eta^T u(x))u(x)dx = \mathbb{E}[u(x)]$$

$$\Rightarrow -\nabla \ln g(\eta) = \mathbb{E}[u(x)]$$

Now consider a set of independent identically distributed data denoted by $X = x_1, ..., x_n$, for which the likelihood function is given by

$$p(X|\eta) = \Big(\prod_{n=1}^{N} h(x_n)\Big) g(\eta)^N exp\Big\{\eta^T \sum_{n=1}^{N} u(x_n)\Big\}$$

By setting $\nabla \ln p(X|\eta) = 0$ when taking gradient with respect to $\eta$, we get condition to be satisfied by the maximum likelihood estimator $\eta_{ML}$

$$\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} u(x_n)$$

which can in principle be solved to obtain $\eta_{ML}$.
** The solution for the maximum likelihood estimator depends on the data only through $\sum_n u(x_n)$, which is therefore called the sufficient statistic of the distribution $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$.
** If we consider the limit $N \rightarrow \infty$, then the right-hand side of above equation becomes $\mathbb{E}[u(x)]$, and so by comparing with $-\nabla \ln g(\eta) = \mathbb{E}[u(x)]$, we see that $\eta_{ML}$ will equal the true value $\eta$.

# Conjugate priors

* Let's go back to 2.2(Multinominal variables) to remember conjugate priors.

Let x be a K-dimensional vector in which one of the elements $x_k$ equals 1, and all remaining elements equal 0. If we denote the probability of $x_k = 1$ by the parameter $\mu_k$, then the distribution of x is given

$$p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

where $\mu = (\mu_1, ..., \mu_K)^T$, and the parameters $\mu_k$ are constrained to satisfy $\mu_k \geq 0$ and $\sum_{k=1}^{K} \mu_k = 1$ since they represent probabilities.

The distribution of $p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$ can be regarded as a generalization of the Bernoulli distribution to more than two outcomes. The distribution is normalized

$$\sum_x p(x|\mu) = p(x|\mu) = \prod_{k=1}^{K} \mu_k = 1$$

and

$$\mathbb{E}[x|\mu] = \sum_x p(x|\mu)x = (\mu_1, ..., \mu_K)^T = \mu$$

Now consider a data set $\mathcal{D}$ of N independent observations $x_1, ..., x_N$. The corresponding likelihood function takes the form

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k}$$

where $m_k = \sum_n x_{nk}$ which represent the number of observations of $x_k = 1$. These are called the *sufficient statistics* for this distribution.

In order to find the maximum likelihood solution for $\mu$, we need to maximize $\ln p(\mathcal{D}|\mu)$ with respect to $\mu_k$ by taking account of $\sum_k \mu_k = 1$. This can be achieved using a Lagrange multiplier $\lambda$ and maximizing

$$\sum_{k=1}^{K} \mu_k \ln \mu_k + \lambda\Big(\sum_{k=1}^{K} \mu_k - 1\Big)$$

By setting the derivative with respect to $\mu_k$ to 0, we obtain

$$\mu_k = -m_k/\lambda$$

By solving for the Lagrange multiplier $\lambda$ and substituting into the constraint $\sum_{k=1}^{K} \mu_k = 1$, we get $\lambda = -N$. Therefore, the maximum likelihood solution is in the form

$$\mu_k^{ML} = \frac{m_k}{N}.$$

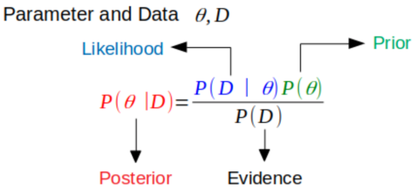
Consider the joint distribution of the quantities $m_1, ..., m_K$, conditioned on the parameters $\mu$ and on the total number N of observations. The *multinomial distribution* takes in the form

$$Mult(m_1, m_2, ..., m_K|\mu, N) = \binom{N}{m_1 m_2 ... m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1 m_2 ... m_K} = \frac{N!}{m_1! ... m_K!}$ and $\sum_{k=1}^{N} m_k = N$.
By introducing a family of prior distributions for the parameters $\{\mu_k\}$ of the multinomial distribution, we see that the *conjugate prior* is given by

$$p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

where $0 \leq \mu_k \leq 1$, $\sum_{k=1}^{K} \mu_k = 1$. and $\alpha_1, ..., \alpha_K$ are the parameters of the distribution which denotes $\alpha = (\alpha_1, ..., \alpha_K)^T$.



Parameter and Data $\theta, D$
Likelihood
Prior
$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$
Posterior
Evidence

In Bayesian probability theory, if the posterior distribution is in the same family of the prior distribution, then the prior and posterior are called conjugate distributions, and the prior is called the conjugate prior to the likelihood function.

For any of the exponential family $p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$, there exists a conjugate prior that can be written in the form

$$p(\eta|\mathcal{X}, v) = f(\mathcal{X}, v)g(\eta)\ v\ exp\{v\eta^T \mathcal{X}\}$$

where $f(\mathcal{X}, v)$ is a normalization coefficient, and $g(\eta)$ is the same function as appears in exponential family distribution.
By multiplying the prior of the above equation with the likelihood function $p(X|\eta) = (\prod_{n=1}^{N} h(x_n))g(\eta)^N exp\{\eta^T \sum_{n=1}^{N} u(x_n)\}$, we obtain the posterior distribution in the form

$$p(\eta|X, \mathcal{X}, v) \propto g(\eta)^{v+N}\ exp\Big\{\eta^T \Big(\sum_{n=1}^{N} u(x_n) + v\mathcal{X}\Big)\Big\}$$

# Noninformative priors

In many cases, we may have little idea of what form the distribution should take. Then, we may look for a form of prior distribution which have as little effect on the posterior distribution as possible. It is called a noninformative prior.

**Example:**

If a density takes the form

$$p(x|\mu) = f(x - \mu)$$

where the parameter $\mu$ is known as a location parameter. This family of densities exhibits translation invariance because if we shift x by a constant to give $\hat{x} = x + c$, then

$$p(\hat{x}|\hat{\mu}) = f(\hat{x} - \hat{\mu})$$

where $\hat{\mu} = \mu + c$.

Choose a prior that assigns equal probability mass to an interval $A \leq \mu \leq B$ as to the shifted interval $A - c \leq \mu \leq B - c$.

$$\Rightarrow \int_A^B p(\mu) d\mu = \int_{A-c}^{B-c} p(\mu) d\mu = \int_A^B p(\mu - c) d\mu$$

Since this must hold for all choices of A and B, then we have

$$p(\mu - c) = p(\mu)$$

which implies that $p(\mu)$ is constant.
* An example of a location parameter would be the mean $\mu$ of a Gaussian distribution. As we have seen, the conjugate prior distribution for $\mu$ in this case is a Gaussian $p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$, and we get noninformative prior by taking the limit when $\sigma_0^2 \rightarrow \infty$.