Lectures on Machine Learning (Fall 2017)

Hyeong In Choi Seoul National University

Lecture 4: Exponential family of distributions and generalized linear model (GLM) (Draft: version 0.9.2)

Topics to be covered:

- Exponential family of distributions
- Mean and (canonical) link functions
- Convexity of log partition function
- Generalized linear model (GLM)
- Various GLM models

1 Exponential family of distributions

In this section, we study a family of probability distribution called the exponential family (of distributions). It is of a special form, but most, if not all, of the well known probability distributions belong to this class.

1.1 Definition

Definition 1. *A probability distribution (PDF or PMF) is said to belong to the* **exponential family of distributions (in natural** *or* **canonical form)** *if it is of the form.*

$$P_\theta(y) = \frac{1}{Z(\theta)} h(y) e^{\theta \cdot T(y)}, \tag{1}$$

where $y = (y_1, \cdots, y_m)$ *is a point in* $\mathbb{R}^m$*; and* $\theta = (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$ *is a parameter called the* ***canonical (natural) parameter****;* $T : \mathbb{R}^m \to \mathbb{R}^k$ *is a map* $T(y) = (T_1(y), \cdots, T_k(y))$*; and* $Z(\theta) = \int h(y) e^{\theta \cdot T(y)} dy$ *is called the* ***partition function****, while its logarithm,* $A(\theta) = \log Z(\theta)$*, is called the* ***log partition (cumulant) function****.*

Remark. *In this lecture and throughout this course, the "dot" notation as in* $\theta \cdot T(y)$ *always means the inner (dot) product of two vectors.*

Equivalently, $P_\theta(y)$ can be written in the form

$$P_\theta(y) = \exp[\theta \cdot T(y) - A(\theta) + C(y)], \tag{2}$$

where $C(y) = \log h(y)$. More generally, one sometimes introduces an extra parameter ϕ, called the dispersion parameter, to control the shape of $P_\theta(y)$ by

$$P_\theta(y) = \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y, \phi)\right].$$

1.2 Standing assumption

In our discussion on the exponential family of distribution, we always assume the following.

- In case $P_\theta(y)$ is a probability density function (PDF), it is assumed to be continuous as a function of y. It means that there is no singularity in the probability measure $P_\theta(y)$;
- In case $P_\theta(y)$ is a probability mass function (PMF), there exists a range of discrete value of $P_\theta(y)$ that is the same for all θ and for all y.

If $P_\theta(y)$ satisfies either condition, we say it is **regular**, which is always assumed throughout this course.

Remark. *Sometimes people use the more general form of $P_\theta(y)$ to write*

$$P_\theta(y) = \frac{1}{Z(\theta)} h(y) e^{\eta(\theta) \cdot T(y)}.$$

But most of the time the same result can be obtained without the use of the general form $\eta(\theta)$. So it is not much of a loss of generality to stick to our convention of using just θ.

1.3 Examples

Let us now look at a few illustrative examples.

(1) Bernoulli(μ)

The Bernoulli distribution is perhaps the simplest in the exponential family. Let Y be a random variable taking its binary value in $\{0, 1\}$. Let $\mu = P[Y = 1]$. Its distribution (in fact, the probability mass function, PMF) is then succinctly written as $P(y) = \mu^y (1-\mu)^{1-y}$. Then,

$$\begin{aligned} P(y) &= \mu^y (1-\mu)^{1-y} \\ &= \exp[y \log \mu + (1-y) \log(1-\mu)] \\ &= \exp\left[y \log\left(\frac{\mu}{1-\mu} \right) + \log(1-\mu) \right]. \end{aligned}$$

Letting $T(y) = y$ and $\theta = \log\left(\frac{\mu}{1-\mu} \right)$ and recalling the definitions of logit and σ functions given in Lecture 3, we have

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu} \right).$$

Thus

$$\theta = \text{logit}(\mu). \tag{3}$$

Its inverse function is the sigmoid function:

$$\mu = \sigma(\theta), \tag{4}$$

where

$$\sigma(\theta) = \frac{1}{1+e^{-\theta}}.$$

Therefore, we have

$$A(\theta) = -\log(1-\mu) = \log(1+e^{\theta}). \tag{5}$$

Thus $P_\theta(y)$, written in canonical form as in (2), becomes

$$P_\theta(y) = \exp[\theta \cdot y - \log(1+e^{\theta})].$$

(2) Exponential distribution

The exponential distribution is a distribution that models the independent arrival time. Its distribution (the probability density function, PDF) is given as

$$P_\theta(y) = \theta e^{-\theta y}\mathbb{I}(x \geq 0).$$

To put it in the exponential family form, we use the same θ as the canonical parameter and we let $T(y) = -y$ and $h(y) = \mathbb{I}(y \geq 0)$. Since

$$Z(\theta) = \frac{1}{\theta} = \int e^{-\theta y}\mathbb{I}(y \geq 0)dy,$$

it is already in the canonical form given as in (1).

(3) Normal distribution

The normal (Gaussian) distribution given by

$$P(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

is the single most well known distribution. As far as its relation with the exponential family is concerned there are two views. (Here, this σ is a number, not the sigmoid function.)

- **1st view** (σ^2 as a dispersion parameter)
  This is the case when the main focus is the mean. In this case, the variance σ^2 is regarded as known or as a parameter that can be fiddled with as if known. Writing $P(y)$ in the form

$$P_\theta(y) = \exp\left[\frac{-\frac{1}{2}y^2 + y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right],$$

One can see right away it is in the form of the exponential family if we set

$$\begin{aligned} \theta &= (\theta_1, \theta_2) = (\mu, 1) \\ T(y) &= (y, -\frac{1}{2}y^2) \\ \phi &= \sigma^2 \\ A(\theta) &= \frac{1}{2}\mu^2 = \frac{1}{2}\theta_1^2 \\ C(y, \phi) &= -\frac{1}{2}\log(2\pi\sigma^2). \end{aligned}$$

- **2nd view** ($\phi = 1$)
  When both μ and σ are parameters to be treated as unknown, we take this point of view. In here, we set the dispersion parameter $\phi = 1$. Writing out $P_\theta(y)$ we have

$$P_\theta(y) = \exp\left[-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}\mu^2 - \log\sigma - \frac{1}{2}\log 2\pi\right].$$

  Thus it is easy to see the following:

$$\begin{aligned} T(y) &= (y, \frac{1}{2}y^2) \\ \theta &= (\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2}) \\ A(\theta) &= \frac{1}{2\sigma^2}\mu^2 + \log\sigma = -\frac{1}{2}\frac{\theta_1^2}{\theta_2} - \frac{1}{2}\log(-\theta_2) \\ C(y) &= -\frac{1}{2}\log(2\pi). \end{aligned}$$

1.4 Properties of exponential family

The log partition function $A(\theta)$ plays a key role, so let us now look at it more carefully. First, since

$$\int P_\theta(y)dy = \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y, \phi)\right] dy = 1,$$

taking ∇_θ, we have

$$\int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] \frac{T(y) - \nabla_\theta A(\theta)}{\phi} dy = 0.$$

Thus

$$\begin{aligned}\int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] T(y) dy &= \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] \nabla_\theta A(\theta) dy \\ &= \nabla_\theta A(\theta) \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] dy.\end{aligned}$$

Since

$$\int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] dy = 1,$$

we have the following:

Proposition 1.

$$\nabla_\theta A(\theta) = \int P_\theta(y) T(y) dy = E[T(Y)],$$

where Y is the random variable with distribution $P_\theta(y)$.

Writing the component, we have

$$\frac{\partial A}{\partial \theta_i} = \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] T_i(y) dy.$$

Taking the second partial derivative, we have

$$\begin{aligned}\frac{\partial^2 A}{\partial \theta_i \partial \theta_j} &= \frac{1}{\phi} \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] \left\{T_j(y) - \frac{\partial A(\theta)}{\partial \theta_j}\right\} T_i(y) dy \\ &= \frac{1}{\phi} \left(\int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] T_i(y) T_j(y) dy \right. \\ &\quad \left. - \frac{\partial A(\theta)}{\partial \theta_j} \int \exp\left[\frac{\theta \cdot T(y) - A(\theta)}{\phi} + C(y,\phi)\right] T_i(y) dy \right).\end{aligned}$$

Using Proposition 1 once more, we have

$$\begin{aligned}\frac{\partial^2 A}{\partial\theta_i\partial\theta_j} &= \frac{1}{\phi}\Big\{E[T_i(Y)T_j(Y)] - E[T_i(Y)]E[T_j(Y)]\Big\} \\ &= \frac{1}{\phi}E\Big[\big(T_i(Y) - E[T_i(Y)]\big)\big(T_j(Y) - E[T_j(Y)]\big)\Big] \\ &= \frac{1}{\phi}\mathrm{Cov}(T_i(Y), T_j(Y)).\end{aligned}$$

Therefore we have the following result on the Hessian matrix of A.

Proposition 2. $D^2_\theta A = \Big(\dfrac{\partial^2 A}{\partial\theta_i\partial\theta_j}\Big) = \dfrac{1}{\phi}\mathrm{Cov}(T(Y), T(Y)) = \dfrac{1}{\phi}\mathrm{Cov}(T(Y))$, *where Y is the random variable with distribution $P_\theta(y)$.*

Here, $\mathrm{Cov}(T(Y), T(Y))$ denotes the covariance matrix of $T(Y)$ with itself, which is sometimes called the variance matrix. Since the covariance matrix is always positive semi-definite, we have

Corollary 1. *$A(\theta)$ is a convex function of θ.*

Remark. *In most exponential family models, $\mathrm{Cov}(T(Y))$ is positive definite, in which case $A(\theta)$ is strictly convex. In this lecture and throughout the course, we always assume that $\mathrm{Cov}(T(Y))$ is positive definite and therefore that $A(\theta)$ is* strictly convex.

1.5 Maximum likelihood estimation

Let $\mathfrak{D} = \{y^{(i)}\}_{i=1}^N$ be a given data of IID samples, where $y^{(i)} \in \mathbb{R}^m$. Then its likelihood function $L(\theta)$ and the log likelihood function $l(\theta)$ are given by

$$\begin{aligned}L(\theta) &= \prod_{i=1}^N \exp\left[\frac{\theta\cdot T(y^{(i)}) - A(\theta)}{\phi} + C(y^{(i)}, \phi)\right] \\ l(\theta) &= \frac{1}{\phi}\big\{\theta\cdot\sum_{i=1}^N T(y^{(i)}) - NA(\theta)\big\} + \sum_{i=1}^N C(y^{(i)}, \phi).\end{aligned}$$

Since $A(\theta)$ is strictly convex, $l(\theta)$ has a unique maximum at $\widehat{\theta}$ that is the unique solution of $\nabla_\theta l(\theta) = 0$. Note that

$$\nabla_\theta l(\theta) \;=\; \frac{1}{\phi}\{\sum_{i=1}^{N} T(y^{(i)}) - N\nabla_\theta A(\theta)\}.$$

So we have the following:

Proposition 3. *There is a unique $\widehat{\theta}$ that maximizes $l(\theta)$ at which*

$$\nabla_\theta A(\theta)|_{\theta=\widehat{\theta}} = \frac{1}{N}\sum_{i=1}^{N} T(y^{(i)}).$$

1.5.1 Example: Bernoulli Distribution

Let us apply the above argument to the Bernoulli distribution Bernoulli(μ) and see what it leads to. First, differentiating $A(\theta)$ as in (5), we get

$$A'(\theta) = \frac{1}{1+e^{-\theta}} = \sigma(\theta).$$

Thus by Proposition 3 we have a unique $\widehat{\theta}$ satisfying

$$\sigma(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} T(y^{(i)}).$$

Since $T(y) = y$, and y takes on 0 or 1 as its value, we have $T(y) = \mathbb{I}(y = 1)$. Defining $\widehat{\mu} = \sigma(\widehat{\theta})$ by (4), we then have

$$\begin{aligned}
\widehat{\mu} &= \sigma(\widehat{\theta}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y^{(i)} = 1) \\
&= \frac{\text{Number of 1's in the sample } \{y^{(i)}\}}{N}.
\end{aligned}$$

This confirms the usual estimate of $\widehat{\mu}$.

Remark. ***T is a sufficient statistic.***
It is well known that the statistic $S(y_i, \cdots, y_n) = \sum_{i=1}^{n} T(y^{(i)})$ is sufficient, which means that θ is a constant on the level set

$$\{(y_1, \cdots, y_n) \,|\, S(y_1, y_2, \cdots, y_n) = \text{const}\}.$$

This provides many advantages in estimating θ, as one needs not to look inside the level sets.

1.6 Maximum Entropy Distribution

Exponential family is useful in that it is the most random distribution under some constraints. Let Y be a random variable with an unknown distribution $P(y)$. Suppose the expected values of certain features $f_1(y), \cdots, f_k(y)$ are nonetheless known. Namely,

$$\int f_i(y)P(y)dy = C_i \tag{6}$$

is known for some given constant C_i for $i = 1, \cdots, k$. The question is how to fix $P(y)$ so that it is as random as possible while satisfying the above constraints. One approach is to construct a maximum entropy distribution satisfying the constraints (6). The entropy is the measure of randomness and in general, the higher it is, the more random it is deemed. The definition of entropy is

$$H = \sum_y P(y) \log P(y),$$

for the discrete case; for the continuous case, the sum becomes an integral so that

$$H = \int P(y) \log P(y) dy.$$

We will use the integral notation here without the loss of generality. To solve the problem, we use the Lagrange multiplier method to set up the following variational problem of minimizing $J(P, \lambda)$:

$$J(P, \lambda) = -\int P(y) \log P(y) dy + \lambda_0 (1 - \int P(y) dy) + \sum_{i=1}^{k} \lambda_i \Big\{ C_i - \int f_i(y) P(y) dy \Big\}.$$

The solution is found as the critical point. Namely, setting the variational (functional or Fréchet) derivative equal to 0,

$$\frac{\delta J}{\delta P} = -1 - \log P(y) - \lambda_0 - \sum_{i=1}^{k} \lambda_i f_i(y) = 0. \tag{7}$$

To see where this comes from, let η be any smooth function with compact support and define

$$J(P+\epsilon\eta, \lambda) = -\int (P+\epsilon\eta)\log(P+\epsilon\eta)dy + \lambda_0\big\{1 - \int (P+\epsilon\eta)dy)\big\} + \sum \lambda_i\big\{C_i - \int f_i(P+\epsilon\eta)dy\big\}.$$

Taking the derivative with respect to ϵ at $\epsilon = 0$ and setting it equal to 0, we get

$$\begin{aligned}
\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} J(P+\epsilon\eta, \lambda) &= -\int (\eta + \eta\log P)dy - \lambda_0 \int \eta dy + \sum \lambda_i\{-\int f_i \eta dy\} \\
&= \int \big(-1 - \log P - \lambda_0 - \sum_{i=1}^{k} \lambda_i f_i\big)\eta dy \\
&= 0.
\end{aligned}$$

Since the above formula is true for any η, we have

$$-1 - \log P - \lambda_0 - \sum_{i=1}^{k} \lambda_i f_i = 0,$$

which is (7). Solving (7) for $P(y)$, we have

$$P(y) = \frac{1}{Z}\exp(-\sum_{i=1}^{k} \lambda_i f_i(y)),$$

where $Z = e^{1+\lambda_0}$. Now

$$1 = \int P(y)dy = \frac{1}{Z}\int \exp(-\sum_{i=1}^{k} \lambda_i f_i(y))dy.$$

Thus

$$Z = \int \exp(-\sum_{i=1}^{k} \lambda_i f_i(y))dy.$$

Therefore $P(y)$ belongs to the exponential family of distribution. This kind of distribution is in general called the **Gibbs distribution**.

2 Generalized liner model (GLM)

2.1 Mean parameter and canonical link function

Before we embark on the generalized linear model, we need the following elementary fact.

Lemma 1. *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a strictly convex C^1 function. Then $\nabla_x f : \mathbb{R}^k \to \mathbb{R}^k$ is an invertible function.*

Proof. Let $p \in \mathbb{R}^k$ be any given vector. By the strict convexity and the absence of corners (the C^1 condition), there is a unique hyperplane H in $\mathbb{R}^{k+1}$ that is tangent to the graph of $y = f(x)$, having the normal vector of the form $(p, -1)$. Let $(\overline{x}, \overline{y})$ be the point of contact. Since the graph is the zero set of $F(x, y) = f(x) - y$, its normal vector at $(\overline{x}, \overline{y})$ is $(\nabla f(\overline{x}), -1)$. Therefore we must have $p = \nabla f(\overline{x})$. □

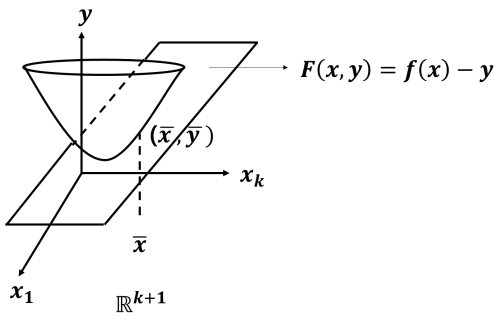

Figure 1: Graph of a strictly convex function with a tangent plane

Let us now look at the exponential family

$$P_\theta(y) = \exp\big[\theta \cdot T(y) - A(\theta) + C(y)\big], \tag{8}$$

where the dispersion parameter ϕ is set to be 1 for the sake of simplicity of the presentation. (The argument does not change much even with the presence of the dispersion parameter.) Recall that Proposition 1 says that $\nabla_\theta A(\theta) = E[T(Y)]$. We always assume that the log partition function $A : \mathbb{R}^k \to \mathbb{R}$ is strictly convex and C^1. Thus by the above Lemma, $\nabla_\theta A : \mathbb{R}^k \to \mathbb{R}^k$ is invertible. We now set up a few terminologies.

Definition 2. *The* **mean parameter** μ *is defined as*

$$\mu = E[T(Y)] = \nabla_\theta A(\theta),$$

and the function $\nabla_\theta A : \mathbb{R}^k \to \mathbb{R}^k$ *is called the* **mean function**.

Since $\nabla_\theta A$ is invertible, we define

Definition 3. *The inverse function* $(\nabla_\theta A)^{-1}$ *of* $\nabla_\theta A$ *is called the* **canonical link function** *and is denoted by* ψ. *Thus*

$$\theta = \psi(\mu) = (\nabla_\theta A)^{-1}(\mu).$$

Therefore, combining the above definitions, the mean function is also written as

$$\mu = \psi^{-1}(\theta) = \nabla_\theta A(\theta).$$

Example: Bernoulli(μ)

Recall, for Bernoulli(μ), $P(y)$ is given by

$$\begin{aligned} P(y) = \mu^y(1-\mu)^{1-y} &= \exp\left[y \log \frac{\mu}{1-\mu} + \log(1-\mu)\right] \\ &= \exp\left[\,\theta y - \log(1+e^\theta)\,\right] \\ &= \exp\left[\,\theta y - A(\theta)\,\right], \end{aligned}$$

where $\theta = \log\left(\dfrac{\mu}{1-\mu}\right)$ and $A(\theta) = \log(1+e^\theta)$. Thus

$$\begin{aligned} \mu &= \text{sigmoid}(\theta) = \mathrm{A}'(\theta) \\ \theta &= \text{logit}(\mu) = (\mathrm{A}')^{-1}(\theta). \end{aligned}$$

Therefore for the Bernoulli distribution the sigmoid function is the mean function and the logit function the canonical link function.

2.2 Conditional probability in GLM

From now on in this lecture, to simplify notation, we use ω to stand for both w and b and extend $x \in \mathbb{R}^d$ to $x \in \mathbb{R}^{d+1}$ by adding $x_0 = 1$. (In the previous lectures, we used θ as a generic term to represent both w and b. But in the exponential family notation, θ is reserved to denote the canonical parameter. So we are forced to use ω. One more note: although the typography is difficult to discern, this ω is the Greek lower case 'omega,' not English w.)

2.2.1 GLM recipe

With this notation, the expression $\omega = (b, w_1, \cdots, w_d)$ as (1) of Lecture 3 is now written as $\omega \cdot x$, and thus the conditional probability in Lecture 3 is written as

$$\begin{aligned} P(y=1 \,|\, x) &= \frac{e^{\omega \cdot x}}{1+e^{\omega \cdot x}} \\ P(y=0 \,|\, x) &= \frac{1}{1+e^{\omega \cdot x}}, \end{aligned}$$

which can be simplified as

$$P(y \,|\, x) = \frac{e^{(\omega \cdot x)y}}{1+e^{\omega \cdot x}},$$

for $y \in \{0, 1\}$. Using (5), this is easily seen to be equivalent to the following conditional probability model

$$\begin{aligned} P(y \,|\, x) &= \exp\left[(\omega \cdot x)y - A(\omega \cdot x)\right] \\ &= \exp\left[\theta y - A(\theta)\right]. \end{aligned}$$

Summarizing this process, we have the following:

- **GLM Recipe:**
  $P(y \,|\, x)$ is gotten from $P(x)$ by replacing the canonical parameter θ in (8) with a linear expression of x.

In the binary case, the linear expression is $\omega \cdot x$. The multiclass case mimics it in exactly the same way. To describe it correctly, first define the parameter

$$W = \begin{pmatrix} (\omega_1)^T \\ \vdots \\ (\omega_i)^T \\ \vdots \\ (\omega_k)^T \end{pmatrix} = \begin{pmatrix} \omega_{10} & \cdots & \omega_{1j} & \cdots & \omega_{1d} \\ \vdots & & \vdots & & \vdots \\ \omega_{i0} & \cdots & \omega_{ij} & \cdots & \omega_{id} \\ \vdots & & \vdots & & \vdots \\ \omega_{k0} & \cdots & \omega_{kj} & \cdots & \omega_{kd} \end{pmatrix}. \tag{9}$$

Then replace θ in (8) with Wx to get the conditional probability as

$$P(y \,|\, x) = \exp[(Wx) \cdot T(y) - A(Wx) + C(y)]. \tag{10}$$

Written this way, we can see that

$$(Wx) \cdot T(y) = \sum_{\ell=1}^{k} \sum_{j=0}^{d} \omega_{\ell j} x_j T_\ell(y).$$

Now, let $\mathfrak{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ be a given data, where the x-part of the i-th data is represented by the column vector

$$x^{(i)} = [x_0^{(i)}, \cdots, x_d^{(i)}]^T.$$

Using the conditional probability as in (10), one can write the likelihood function and the log likelihood function by

$$\begin{aligned} L(W) &= \prod_{i=1}^{N} \exp[(Wx^{(i)}) \cdot T(y^{(i)}) - A(Wx^{(i)}) + C(y^{(i)})] \\ l(W) &= \log L(W) = \sum_{i=1}^{N} [(Wx^{(i)}) \cdot T(y^{(i)}) - A(Wx^{(i)}) + C(y^{(i)})] \\ &= \sum_{i=1}^{N} \left[\sum_{\ell=1}^{k} \sum_{j=0}^{d} \omega_{\ell j} x_j^{(i)} T_\ell(y^{(i)}) - A(\omega_1^T x^{(i)}, \cdots, \omega_k^T x^{(i)}) + C(y^{(i)}) \right]. \end{aligned}$$

Thus taking the derivative, one gets

$$\begin{aligned} \frac{\partial l(W)}{\partial \omega_{rs}} &= \sum_{i=1}^{N} \left[\sum_{\ell=1}^{k} \sum_{j=0}^{d} x_j^{(i)} \delta_{\ell r} \delta_{js} T_\ell(y^{(i)}) - \sum_{\ell=1}^{k} \frac{\partial A}{\partial \omega_\ell} \sum_{j=0}^{d} x_j^{(i)} \delta_{\ell r} \delta_{js} \right] \\ &= \sum_{i=1}^{N} \left[x_s^{(i)} T_r(y^{(i)}) - \frac{\partial A}{\partial \omega_r} x_s^{(i)} \right] \\ &= \sum_{i=1}^{N} \left[T_r(y^{(i)}) - \frac{\partial A}{\partial \omega_r}(Wx^{(i)}) \right] x_s^{(i)}. \end{aligned} \tag{11}$$

This is a generalized form of what we got for the logistic regression (see (10) of Lecture 3).

2.3 Multiclass classification (softmax regression)

We now look at the multiclass classification problem. The formula (11) is the derivative formula with which the gradient descent algorithm can be used.

But for multiclass regression, the categorical distribution has the redundancy we talked about in Lecture 3.

As usual, let $y_i = \mathbb{I}(y = i)$ and let $\text{Prob}[Y = i] = \mu_i$. Then we must have $y_1 + \cdots + y_k = 1$ and $\mu_1 + \cdots + \mu_k = 1$. Thus the probability (PMF) can be written as

$$P(y) = \mu_1^{y_1} \mu_2^{y_2} \cdots \mu_k^{y_k}.$$

Rewrite $P(y)$ by

$$\begin{aligned} P(y) &= \mu_1^{y_1} \mu_2^{y_2} \cdots \mu_{k-1}^{y_{k-1}} \mu_k^{\left(1 - \sum_{i=1}^{k-1} y_i\right)} \\ &= \exp\left[y_1 \log \mu_1 + \cdots + y_{k-1} \log \mu_{k-1} + \left(1 - \sum_{i=1}^{k-1} y_i\right) \log \mu_k \right] \\ &= \exp\left[\sum_{i=1}^{k-1} y_i \log \frac{\mu_i}{\mu_k} + \log \mu_k \right]. \end{aligned}$$

Note that when $k = 2$, this is exactly the Bernoulli distribution. Define $\theta_i = \log(\mu_i/\mu_k)$ and $T_i(y) = y_i$ for $i = 1, \cdots, k-1$. Using the facts that $\mu_i = \mu_k e^{\theta_i}$ and $1 - \mu_k = \sum_{i=1}^{k-1} \mu_i$, and solving for μ_k and then for μ_j, we get

$$\begin{aligned} \mu_k &= \frac{1}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} \\ \mu_j &= \frac{e^{\theta_j}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}}, \end{aligned} \tag{12}$$

for $j = 1, \cdots, k-1$. The expression in the right hand side of (12) is called the **generalized sigmoid (softmax) function**. Therefore $P(y)$ can be written in the exponential family form as

$$P_\theta(y) = \exp[\theta \cdot T(y) - A(\theta)],$$

where

$$\theta = (\theta_1, \cdots, \theta_{k-1}),$$

$$T(y) = (y_1, \cdots, y_{k-1}),$$

and

$$A(\theta) = -\log \mu_k = \log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right).$$

Note that the mean parameter μ_i is given as

$$\mu_i = E[T_i(Y)] = E[Y_i] = \text{Prob}[Y_i = 1] = \text{Prob}[Y = i],$$

which again can be calculated by the fact $\mu = \nabla_\theta A$. Indeed one can verify that

$$\frac{\partial A}{\partial \theta_j} = \frac{e^{\theta_j}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} = \mu_j.$$

We have shown above that $\theta_j = \log(\mu_j/\mu_k)$. Thus for $j = 1, \cdots, k-1$, we have

$$\theta_j = \log\left(\frac{\mu_j}{1 - \sum_{i=1}^{k-1} \mu_i}\right).$$

The expression on the right side of the above equation is called the **generalized logit function.**

2.4 Probit regression

Recall that the essence of the GLM for the exponential family is the way it links the (outside) features $x_1, \cdots, x_d$ to the probability model. To be specific, given the probability model (8), the linking was done by setting $\theta = \psi(\mu) = \omega \cdot x$ so that the conditional probability becomes

$$P(y \,|\, x) = \exp[(\omega \cdot x)T(y) - A(\omega \cdot x) + C(y)].$$

But there is no a priori reason why μ has to be related to $\omega \cdot x$ via the canonical link function only. One may use any function as long as it invertible. So set $g(\mu) = \omega \cdot x$, where $g : \mathbb{R}^k \to \mathbb{R}^k$ is an invertible function. So we call g a **link function** as opposed to the *canonical* link function and its inverse g^{-1} the **mean function**. So we have

$$\begin{aligned} g(\mu) &= \omega \cdot x \quad &&: \text{link function} \\ \mu &= g^{-1}(\omega \cdot x) \quad &&: \text{mean function.} \end{aligned}$$

Let $\Phi(t)$ be the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0, 1)$, i.e.,

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-s^2/2}\, ds.$$

The probit regression deals with the following situation: the output y is binary taking on values in $\{0, 1\}$, and its probability model is Bernoulli. So $P(y) = \mu^y(1-\mu)^{1-y}$, where $\mu = P(y = 1)$. Its link to the outside variables (features) is via the link function $g = \Phi^{-1}$. Thus

$$\begin{aligned} g(\mu) &= \omega \cdot x = \omega_0 + \omega_1 x_1 + \cdots + \omega_d x_d \\ \mu &= g^{-1}(\omega \cdot x) = \Phi(\omega \cdot x). \end{aligned} \tag{13}$$

Therefore,

$$P(y \,|\, x) = \Phi(\omega \cdot x)^y (1 - \Phi(\omega \cdot x))^{1-y}.$$

Since this is still a Bernoulli hence an exponential family distribution, the general machinery of GLM can be employed. But, in the probit regression we change tack and take a different approach. First, we let $y \in \{-1, 1\}$. Then since

$$\begin{aligned} \mu = E[Y] &= 1 \cdot P(y = 1) + (-1) \cdot P(y = -1) \\ &= P(y = 1) - [1 - P(y = 1)], \end{aligned}$$

we have $\mu = 2P(y = 1) - 1$. Thus $P(y = 1) = \dfrac{1}{2}(1 + \mu)$ and $P(y = -1) = \dfrac{1}{2}(1 - \mu)$. Therefore,

$$P(y) = \frac{1}{2}(1 + y\mu).$$

Thus using (13), we have

$$P(y \,|\, x) = \frac{1}{2}[1 + y\Phi(\omega \cdot x)].$$

Now let the data $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^n$ be given. Then the log likelihood function is

$$\begin{aligned} l(\omega) &= \sum_{i=1}^n \log P(y_i \,|\, x_i) \\ &= \sum_{i=1}^n \log \left(\frac{1}{2}[1 + y_i \Phi(\omega \cdot x_i)] \right), \end{aligned}$$

Thus

$$\frac{\partial l(\omega)}{\partial \omega_k} = \sum_{i=1}^n \frac{y_i \Phi'(\omega \cdot x_i)}{1 + y_i \Phi(\omega \cdot x_i)} x_{ik},$$

which is in a neat form to apply the numerical methods introduced in Lecture 3, where $\Phi'(\omega \cdot x) = \dfrac{1}{\sqrt{2\pi}} e^{-(\omega \cdot x)^2}/2$.

Comments

(1). When one uses a link function which is not canonical, the probability model does not even have to belong to the exponential family. As long as one can get a hold of the conditional distribution $P(y \,|\, x)$, one is in business. Namely with it one can form the likelihood function and the rest of the maximum likelihood machinery.

(2). One may not even need the probability distribution as long as one can somehow estimate ω so that the decisions can be made.

(3). Historically the GLM was first developed for the exponential family but was later extended to the non-exponential family and even to the case where the distribution is not completely known.

2.5 GLM for other distributions

2.5.1 GLM for Poisson(μ)

Recall that the Poisson distribution is the probability distribution (PMF) on the non-negative integers given by

$$P(n) = \frac{e^{-\mu}\mu^n}{n!},$$

for $n = 0, 1, 2, \cdots$. It is an exponential family distribution as it can be written as

$$\begin{aligned} P(y) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp\{y \log \mu - \mu - \log(y!)\}, \end{aligned}$$

where $y = 0, 1, 2, \cdots$. It can be put in the exponential family form by setting

$$\begin{aligned} \theta &= \log \mu \quad : \text{canonical link function} \\ \mu &= e^{\theta} \quad : \text{mean function.} \end{aligned}$$

One can independently check that

$$E[Y] = \sum_{n=0}^{\infty} nP(n) = \mu.$$

2.5.2 GLM for $\Gamma(\alpha, \lambda)$

Recall that the gamma distribution is written in the following form:

$$\begin{aligned} P(y) &= \frac{\lambda e^{-\lambda y}(\lambda y)^{\alpha-1}}{\Gamma(\alpha)}, \qquad \text{for} \quad y \geq 0 \\ &= \exp\{\log\lambda - \lambda y + (\alpha-1)\log(\lambda y) - \log\Gamma(\alpha)\} \\ &= \exp\{\alpha\log\lambda - \lambda y + (\alpha-1)\log y - \log\Gamma(\alpha)\} \\ &= \exp\left\{\frac{-\lambda\phi y - (-\log\lambda)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma(1/\phi)\right\}. \end{aligned}$$

Set $\phi = \dfrac{1}{\alpha}$ as the dispersion parameter and let $\theta = -\lambda\phi$. Then $P(y)$ is an exponential family distribution written as:

$$P_\theta(y) = \exp\left\{\frac{\theta y - (-\log(-\theta))}{\phi} + C(y,\phi)\right\},$$

where

$$C(y,\phi) = \frac{1}{\phi}\log(1/\phi) + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma(1/\phi).$$

Therefore the log partition function is

$$A(\theta) = -\log(-\theta).$$

Differentiating this we get the mean parameter

$$\mu = A'(\theta) = -\frac{1}{\theta}.$$

Since $E[Y] = \dfrac{\alpha}{\lambda}$ for $Y \sim \Gamma(\alpha, \lambda)$, this fact is independently verified. To recap, the mean function and the canonical link function are given by

$$\begin{aligned} \mu &= \psi^{-1}(\theta) = -\frac{1}{\theta} \quad : \quad \text{mean function} \\ \theta &= \psi(\mu) = -\frac{1}{\mu} \quad : \quad \text{canonical link function.} \end{aligned}$$

In practice, this canonical link function is rarely used. Instead one uses one of the following three alternatives.

(i) the inverse link $g(\mu) = \dfrac{1}{\mu}$,

(ii) the log link $g(\mu) = \log \mu$,

(iii) the identity link $g(\mu) = \mu$.

2.5.3 GLM Summary

The GLM link and inverse link (mean) functions for various probability models are summarized in the following Table 1. In here, the range of j in the link and the inverse link of the categorical distribution is $j = 1, \cdots, k-1$, although $\mu = (\mu_1, \cdots, \mu_k)$.

	Range	Link	Inverse Link (mean)	Dispersion
$N(\mu, \sigma^2)$	$(-\infty, \infty)$	$\theta = \mu$	$\mu = \theta$	σ^2
Bernoulli(μ)	$\{0, 1\}$	$\theta = \text{logit}(\mu)$	$\mu = \sigma(\theta)$	1
Categorical(μ)	$\{1, \cdots, k\}$	$\theta_j = \log\left(\dfrac{\mu_j}{1 - \sum_{i=1}^{k-1} \mu_i}\right)$	$\mu_j = \dfrac{e^{\theta_j}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}}$	1
Probit	$\{0, 1\}$ or $\{-1, 1\}$	Φ^{-1}	Φ	1
Poisson(μ)	Z^+	$\theta = \log(\mu)$	$\mu = e^{\theta}$	1
Gamma(α, λ)	R^+	$\theta = \dfrac{1}{\mu}$	$\mu = \dfrac{1}{\theta}$	$\dfrac{1}{\alpha}$

Table 1: GLM-Summary