COS 513:FOUNDATIONS OF PROBABILISTIC MODELING LECTURE 12

YIYUE WU, YUEJIE CHI

1. GENERALIZED LINEAR MODELS

We have known that Generalized Linear Models (GLM's) are a general category of models that includes linear regression and linear classification models as special cases. In Fig. 1, the relationships between the variables in a GLM are illustrated where the GLM makes the following 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 $\beta^T x$.
- The observed output y is assumed to be characterized by an exponential family distribution with conditional mean μ.
- The conditional mean μ is represented as a function $f(\beta^T x)$ of the linear combination $\beta^T x$ of the observed input x. f is named as the response function.
- The natural parameter η can be mapped from the conditional mean μ as a function $\eta = \psi(\mu)$.

$$\begin{tikzcd}
\beta \arrow[dr] & & & \\
& \beta^T x \arrow[r, "f"] & \mu \arrow[r, "\psi"] & \eta \\
x \arrow[ur] & & &
\end{tikzcd}$$

FIGURE 1. Relationships between the variables in a GLM model.

1.1. Overdispersed GLM's.

Within the GLM framework, it is convenient to work with a slight variation on the exponential family theme. The overdispersed GLM is one variation based on the original GLM which is given by

$$p(y|\eta) = h(y,\delta)\exp\left\{\frac{\eta^T y - a(\eta)}{\delta}\right\} \tag{1}$$

The distribution in this form is called in the *overdispersed exponential family*.

Many exponential family distributions, including the Gaussian and the gamma, are naturally expressed in this form. As an example, we can fit the linear regression into the overdispersed setting. The linear regression model is expressed as

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y-\beta^T x)^2}{2\sigma^2} \right\} \tag{2}$$

We can rewrite equation (2) as

$$\begin{aligned} p(y|x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-y^2 + 2y(\beta^T x) - (\beta^T x)^2}{2\sigma^2} \right\} \\ &= \frac{\exp\left\{-\frac{y^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{2y(\beta^T x) - (\beta^T x)^2}{2\sigma^2} \right\} \\ &= \frac{\exp\left\{-\frac{y^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{y(\beta^T x) - \frac{1}{2}(\beta^T x)^2}{\sigma^2} \right\} \end{aligned} \tag{3}$$

Therefore, fitting linear regression model into the overdispersed GLM (assuming $\eta = \beta^T x$ here), we have

$$h(y,\delta) = \frac{\exp\left\{-\frac{y^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}}, \quad a(\eta) = \frac{\eta^2}{2}, \quad \delta = \sigma^2 \quad f, \psi : \text{Identity}.$$

1.2. **Two choices in a GLM.** There are two principal choices in the specification when setting up a GLM:

(1) The choice of the exponential family distribution of the observed output y.
(2) The choice of the response function f.

The choice of the distribution is strongly constrained by the nature of the observed data y. For example, we might use Poisson distribution to model y if it is discrete, and use gamma distribution if it is real positive. For multi-class classification, we might used multinomial/categorical distribution. Therefore, the choice of the response function is the principal degree of freedom in the specification of a GLM.

The natural parameter η can be expressed as $\eta = \psi(f(\beta^T x))$. Suppose $f = \psi^{-1}$, then we have,

$$f = \psi^{-1} \Rightarrow \eta = \beta^T x \tag{4}$$

In this case, f is called the *canonical response function*. It should be clear that the use of canonical response function passes a sanity check automatically with respect to the range constraints. To see this, note that

$$f(\eta) = \psi^{-1}(\eta) = a'(\eta) = \mathbb{E}[Y|\eta]. \tag{5}$$

Thus, f(η) is equal to the conditional mean of the exponential family distribution in which η is the natural parameter.

Also the canonical response function is uniquely associated with a given exponential family distribution. It also needs to be emphasized that the canonical response function is not necessarily the best choice in all situations. Indeed, different choices of the response function can be appropriate in different situations, reflecting different underlying assumptions about the way that the data generated.

1.3. **Example: logistic regression.** If the observed output $y \in \{0, 1\}$, then linear regression does not make sense at all! In this case, we assume that y follows Bernoulli distribution, *i.e.*

$$p(y|\pi) = \pi^y (1-\pi)^{1-y} \tag{6}$$

where π is the mean. Equation (6) can be rewritten as

$$p(y|\pi) = \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) y + \log(1-\pi) \right\} \tag{7}$$

Fitting equation (7) into the exponential family distribution, we have

$$\eta = \psi(\pi) = \log\left(\frac{\pi}{1-\pi}\right). \tag{8}$$

By inverting the relationship between η and π, we have

$$\pi = \psi^{-1}(\eta) = \frac{1}{1+e^{-\eta}} \tag{9}$$

which is the *logistic function*. Fig. 2 illustrates the curve of the logistic function.

The response function is also called link function. Again, noncanonical links are also possible. For example, let Φ be the probit function (inverse cumulative density function associated with the standard normal distribution), we can use

$$y = \Phi^{-1}(\beta^T x)$$

as our response function.

Motivation: In logistic regression, data points far away from the boundary don't matter or matter very little. This is similar to the Supporting Vector Machine (SVM), where only data points near the margin matter. Hence, you can get similar predictions from SVM and logistic regression. The logistic model can also be motivated as an approximation of the step function.

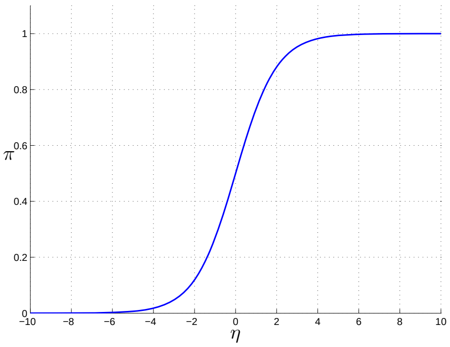

FIGURE 2. Logistic function.

Note that, in the multinomial distribution, β is a matrix, x ia a vector, $\beta^T x$, μ and η are vectors of the same size, f and ψ are functions mapping from a vector into another vector.

1.4. **MLE of β.** Consider an IID data set, $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$. The MLE of β is

$$\beta = \arg\max_\beta \log p(y_{1:N}|x_{1:N}, \beta) = \arg\max_\beta l(\beta, \mathcal{D}) \tag{10}$$

where $l(\beta, \mathcal{D})$ is the log likelihood function given the data set $\mathcal{D}$ and β. We restricted ourselves to scalar y in order to simplify the representation. It is straightforward to generalize to nonscalar case. But still, we keep each element in x as a vector.

Define η_n to be the per-observation natural parameter, i.e.

$$\eta_n = \psi(f(\beta^T x_n)). \tag{11}$$

We obtain the following log likelihood:

$$\begin{aligned} l(\beta, \mathcal{D}) &= \log \prod_{n=1}^N h(y_n) \exp(\eta_n y_n - a(\eta_n)) \\ &= \sum_{n=1}^N \log h(y_n) + \sum_{n=1}^N (\eta_n y_n - a(\eta_n)). \end{aligned}$$

Taking its gradient with respect to β yields:

$$\begin{aligned}\nabla_\beta l(\beta, \mathcal{D}) &= \sum_{n=1}^{N} \frac{dl}{d\eta_n} \nabla_\beta \eta_n \\ &= \sum_{n=1}^{N} (y_n - a'(\eta_n)) \nabla_\beta \eta_n.\end{aligned}$$

According to Eq. (11), the gradient of η_n with respect to β is:

$$\nabla_\beta \eta_n = \psi'(f(\beta^T x_n)) f'(\beta^T x_n) x_n. \tag{12}$$

Define μ_n be the mean response given x_n, i.e.

$$\mu_n = \mathbb{E}[Y|X_n] = f(\beta^T x_n). \tag{13}$$

Note that, the mean response μ_n can also be computed as

$$\mu_n = a'(\eta_n).$$

Assume $f = \psi^{-1}$ is the canonical response function, then $\eta_n = \beta^T x_n$. The log likelihood function now becomes

$$l(\beta, \mathcal{D}) = \sum_{n=1}^{N} \log h(y_n) + \sum_{n=1}^{N} \beta^T x_n y_n - \sum_{n=1}^{N} a(\beta^T x_n). \tag{14}$$

From this, the gradient of the log likelihood function with respect to β becomes

$$\begin{aligned}\nabla_\beta l(\beta, \mathcal{D}) &= \sum_{n=1}^{N} y_n x_n - \sum_{n=1}^{N} \mu_n x_n \\ &= \sum_{n=1}^{N} (y_n - \mu_n) x_n. \qquad (15)\end{aligned}$$

The terms $(y_n - \mu_n)$ is called fitted residuals. In linear regression, $\mu_n = \beta^T x_n$. Eq. (15) has the appealing feature that the parameter vector and the fitted residuals are on the same scale.

2. Sufficiency

2.1. **Definition.** A statistic is a function of an observation. It is also described as a function of random variables. For the sake of discussion here, we let x be a random variable and $t(x)$ be a statistic.

Suppose the distribution of x depends on a parameter θ, then $t(x)$ is sufficient for θ if there's no information in the random variable x regarding θ beyond $t(x)$. If we are making inferences about θ, all we need to know is $t(x)$. Sufficiency characterzes what is essential in a data set, or alternatively, what is inessential and can be thrown away.

2.2. **Two approaches.** Sufficiency is defined in different ways in the Bayesian and frequentist frameworks.

In the *Bayesian* notion, we treat θ as a random variable, and it is natural to consider conditional independent relationships regarding θ. Here, $t(x)$ being a sufficient statistic for θ implies

$$\theta \perp x|t(x) \Longrightarrow p(\theta|t(x), x) = p(\theta|t(x)). \tag{16}$$

In the *frequentist* notion, θ is treated as a label rather than a random variable. $t(x)$ being a sufficient statistic implies that the conditional distribution of x given $t(x)$ does not depend on θ, i.e.

$$p(x|t(x), \theta) = p(x|t(x)). \tag{17}$$

This is equivalent to saying $x \perp \theta|t(x)$.

We can see that the Bayesian notion and the frequentist notion are really the same thing. Both the Bayesian and frequentist definitions of sufficiency imply a factorization of $p(x|\theta)$ as:

$$p(x|\theta) = g(t(x), \theta)h(x, t(x)). \tag{18}$$

For example, in the exponential family distribution,

$$p(x|\eta) = \underbrace{h(x)}_{h(x,t(x))} \underbrace{\exp(\eta^T t(x) - a(\eta))}_{g(t(x),\eta)}, \tag{19}$$

which has a one-to-one correspondence with (18). This is why we call $t(x)$ a sufficient statistic in the exponential family distribution.

2.3. **MLE of an exponential family.** Consider an IID data set $\mathcal{D} = \{x_n\}_{n=1}^N$, where x_n come from the same exponential family. The joint density is obtained by taking the product of the individual density:

$$\begin{aligned} p(x_{1:N}|\eta) &= \prod_{n=1}^{N} h(x_n) \exp(\eta^T t(x_n) - a(\eta)) \\ &= (\prod_{n=1}^{N} h(x_n)) \exp(\eta^T \sum_{n=1}^{N} t(x_n) - Na(\eta)) \end{aligned} \tag{20}$$

is itself in the exponential family with parameters:

$$\begin{aligned}\tilde{h}(x_{1:N}) &= \prod_{n=1}^{N} h(x_n),\\ \tilde{a}(\eta) &= Na(\eta),\\ \tilde{\eta} &= \eta,\\ \tilde{t}(x_{1:N}) &= \sum_{n=1}^{N} t(x_n).\end{aligned}$$

We can see that the sufficient statistic for the joint density is sum of the individual sufficient statistics. Therefore, we only need to keep track of the sum of the individual sufficient statistics $\sum_{n=1}^{N} t(x_n)$. The individual data points can be thrown away.

For the univariate normal distribution, the sufficient statistic is the pair $(\sum_{n=1}^{N} x_n, \sum_{n=1}^{N} x_n^2)$. For the Bernoulli distribution, the sufficient statistic is $\sum_{n=1}^{N} x_n$.