Generalized linear models

Machine Learning Techniques

Dr. Ashish Tendulkar

IIT Madras

Generalized linear models

What will be covered in this module?

In this section, we will show that regression and classification models are special cases of a broader family of models, called Generalized Linear Models (GLMs).

We will also show how other models in the GLM family can be derived and applied to other classification and regression problems.

Exponential Families

We say that a class of distributions is in the exponential family if it can be written in the form

sufficient statistic

log partition function

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

natural or canonical parameter of distribution

We say that a class of distributions is in the exponential family if it can be written in the form

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta))$$

- η is called the natural parameter (also called the canonical parameter) of the distribution.
- $T(y)$ is the sufficient statistic. In most of the cases $T(y) = y$.
- $a(\eta)$ is the log partition function.
- The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure the distribution $p(y; \eta)$ sums/integrates over y to 1.

A fixed choice of T, a and b defines a family (or set) of distributions that is parameterized by η; as we vary η, we then get different distributions within this family.

Bernoulli Distribution

A part of exponential family

The Bernoulli distribution with mean ϕ, written Bernoulli(ϕ), specifies a distribution over $y \in \{0, 1\}$ such that

$$
\begin{aligned}
p(1; \phi) &= \phi \\
p(0; \phi) &= 1 - \phi
\end{aligned}
$$

As we vary ϕ, we obtain Bernoulli distributions with different means.

We can write the Bernoulli distribution as

$$\begin{aligned} p(y;\phi) &= \phi^y (1-\phi)^{1-y} \\ &= \exp(\log(\phi^y (1-\phi)^{1-y})) \\ &= \exp(y \log \phi + (1-y) \log(1-\phi)) \\ &= \exp(y \log \phi - y \log(1-\phi) + \log(1-\phi)) \\ &= \exp\left(y \log\left(\frac{\phi}{1-\phi}\right) + \log(1-\phi)\right) \end{aligned}$$

Consider

$$\eta = \log\left(\tfrac{\phi}{1-\phi}\right)$$

which can be rewritten as

$$\phi = \tfrac{1}{1-e^{-\eta}}$$

$$p(y;\phi) = \exp\left(y\log\left(\frac{\phi}{1-\phi}\right) + \log(1-\phi)\right)$$

compare this with

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

we obtain

$b(y) = 1$ $\quad \eta = \log\left(\frac{\phi}{1-\phi}\right)$ $\quad T(y) = y$ $\quad a(\eta) = -\log(1-\phi)$

This shows that the Bernoulli distribution can be written in the form of $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$, using an appropriate choice of T, a and b.

Gaussian Distribution

A part of exponential family

To simplify the derivation below, lets set $\sigma^2 = 1$.

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) . \exp\left(\frac{-(-2y\mu + \mu^2)}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) . \exp\left(y\mu - \frac{\mu^2}{2}\right)
\end{aligned}
$$

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) . \exp\left(y\mu - \frac{\mu^2}{2}\right)$$

compare this with

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

we obtain

$$b(y) \;= \frac{1}{\sqrt{2\pi}} \;\exp\left(\frac{-y^2}{2}\right)$$

$$\eta = \mu$$

$$T(y) = y$$

$$a(\eta) \;= \eta^2/2$$

This shows that the Gaussian distribution can be written in the form of $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$, using an appropriate choice of T, a and b.

There are many other distributions that are members of the exponential family:

- The multinomial distribution
- The Poisson distribution
- The gamma and the exponential distribuion
- The beta and the dirichlet distribution

Constructing GLMs

Consider a classification or regression problem where we would like to predict the value of some random variable y as a function of $\mathbf{x}$.

To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given $\mathbf{x}$ $(y|\mathbf{x})$ and about our model:

$y|\mathbf{x};\mathbf{w} \sim \text{ExponentialFamily}(\eta)$. that is, the distribution of y given $\mathbf{x}$ parameterized by $\mathbf{w}$ follows some exponential family distribution, with parameter η.

Our goal is to predict the expected value of $T(y)$ given $\mathbf{x}$. Since in most examples, $T(y) = y$, our prediction $h(\mathbf{x})$ need to satisfy the following equality: $h(\mathbf{x}) = E[y|\mathbf{x}]$.

The natural parameter η and the inputs $\mathbf{x}$ are related linearly: $\eta = \mathbf{w}^T\mathbf{x}$.

- The third of these assumptions might seem the least well justified of the above, and it might be better thought of as a design choice in our recipe for designing GLMs, rather than as an assumption per se.
- These three assumptions/design choices will allow us to derive a very elegant class of learning algorithms, namely GLMs, that have many desirable properties such as ease of learning.
- The resulting models are often very effective for modelling different types of distributions over y.

Ordinary Least Squares

Assume that the target variable y (also called the response variable in GLM terminology) is continuous, and we model the conditional distribution of y given $\mathbf{x}$ as a Gaussian:

$$y|\mathbf{x};\mathbf{w} \sim \mathcal{N}(\mu,\ \sigma^2)$$

We know that Gaussian distribution is an exponential family distribution with $\mu = \eta$. This leads to

$$\begin{aligned} h_{\mathbf{w}}(\mathbf{x}) &= E[y|\mathbf{x};\mathbf{w}] && \text{(by Assumption 2)} \\ &= \mu && (\text{Since }, y|\mathbf{x} \sim N(\mu,\sigma^2)) \\ &= \eta && \text{(by Assumption 1)} \\ &= \mathbf{w}^T\mathbf{x} && \text{(by Assumption 3)} \end{aligned}$$

It shows that ordinary least squares is a special case of the GLM family of models.

Logistic Regression

Here we are interested in binary classification, so $y \in \{0, 1\}$.

Given that y is binary-valued, it therefore seems natural to choose the Bernoulli family of distributions to model the conditional distribution of y given $\mathbf{x}$.

In our formulation of the Bernoulli distribution as an exponential family distribution, we had

$$\phi = 1/(1 + e^{-\eta})$$

Note that if $y|\mathbf{x}; \mathbf{w} \sim$ Bernoulli(ϕ), then $E[y|\mathbf{x}; \mathbf{w}] = \phi$

$$\begin{aligned} h_{\mathbf{w}}(\mathbf{x}) &= E[y|\mathbf{x};\mathbf{w}] \\ &= \phi \\ &= \frac{1}{1+e^{-\eta}} \\ &= \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \end{aligned}$$

So, this gives us hypothesis functions of the form $h_{\mathbf{w}}(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$

The function g giving the distribution's mean as a function of the natural parameter $(g(\eta) = E[T(y);\eta])$ is called the canonical response function.

Its inverse, g^{-1}, is called the canonical link function.

Thus, the canonical response function for the Gaussian family is just the identify function; and the canonical response function for the Bernoulli is the logistic function.

Softmax Regression

Consider a classification problem in which the response variable y can take on any one of k values, so $y \in \{1, 2, ..., k\}$

We will thus model it as distributed according to a multinomial distribution.

To parameterize a multinomial over k possible outcomes, one could use k parameters $\phi_1, \phi_2, \ldots, \phi_k$ specifying the probability of each of the outcomes.

These parameters would be redundant, or more formally, they would not be independent (since knowing any $k-1$ of the ϕ's uniquely determines the last one, as they must satisfy

$$\sum_{i=1}^{k} \phi_i = 1.$$

So, we will instead parameterize the multinomial with only $k-1$ parameters, $\phi_1, \phi_2, \ldots, \phi_{k-1}$, where

- $\phi_i = p(y = i; \phi)$
- $p(y = k) = 1 - \sum_{i=1}^{k-1} \phi_i$
- For notational convenience, we will also let $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$

To express the multinomial as an exponential family distribution, we will define $T(y) \in \mathbb{R}^{k-1}$ as follows:

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad T(3) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots \quad T(k-1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad T(k) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$T(y)$ is now a $k-1$ dimensional vector, rather than a real number.

We will write $(T(y))_i$ to denote the i-th element of the vector $T(y)$.

Indicator function: An indicator function $1\{\cdot\}$ takes on a value of 1 if its argument is true, and 0 otherwise$(1\{\mathrm{True}\} = 1, 1\{\mathrm{False}\} = 0)$.

For example: $1\{2 = 3\} = 0,\;\; 1\{5 - 2 = 3\} = 1$

Also we have that $E[(T(y))_i] = p(y = i) = \phi_i$.

$$
\begin{aligned}
p(y;\phi) &= \phi_1^{1\{y=1\}}.\phi_2^{1\{y=2\}} \ldots \phi_k^{1\{y=k\}} \\
&= \phi_1^{1\{y=1\}}.\phi_2^{1\{y=2\}} \ldots \phi_k^{1-\sum_{i=1}^{k-1} 1(y=i)} \\
&= \phi_1^{(T(y))_1}.\phi_2^{(T(y))_2} \ldots \phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i} \\
&= \exp\log(\phi_1^{(T(y))_1}.\phi_2^{(T(y))_2} \ldots \phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i}) \\
&= \exp\left[(T(y))_1 \log\phi_1 + (T(y))_2 \log\phi_2 + \ldots + (1 - \sum_{i=1}^{k-1}(T(y))_i)\log\phi_k\right] \\
&= \exp[(T(y))_1 \log(\phi_1/\phi_k) + (T(y))_2 \log(\phi_2/\phi_k) + \ldots + \\
&\qquad\qquad (T(y))_{k-1}\log(\phi_{k-1}/\phi_k) + \log\phi_k] \\
&= b(y)\exp(\eta^T T(y) - a(\eta))
\end{aligned}
$$

where

$$\eta = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix},$$

$$a(\eta) = \log \phi_k$$

$$b(y) = 1$$

Therefore, multinomial distribution is a part of the exponential family distribution.

The link function is given (for $i = 1, ..., k$) by

$$\eta_i = \log \frac{\phi_i}{\phi_k}$$

For convenience, we have also defined $\eta_k = \log \frac{\phi_k}{\phi_k} = 0$

To invert the link function and derive the response function, we have that

$$
\begin{aligned}
e^{\eta_i} &= \frac{\phi_i}{\phi_k} \\
\phi_k e^{\eta_i} &= \phi_i && \text{...}(1) \\
\phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k} \phi_i = 1 \\
\phi_k &= \frac{1}{\sum_{i=1}^{k} e^{\eta_i}} && \text{...}(2)
\end{aligned}
$$

By equations (1) and (2), we have

$$\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

This function mapping from the $\eta' s$ to the ϕ's is called the softmax function.

Recall from the assumption (3) that $\eta_i' s$ are linearly related to $\mathbf{x}$, we have

$$\eta_i = \mathbf{w}_i^T \mathbf{x} \; (\text{for } i = 1, 2, \ldots, k-1)$$

where $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k-1} \in \mathbb{R}^{n+1}$ are the parameters of the model.

For notational convenience, we can also define $\mathbf{w}_k = 0$, so that $\eta_k = \mathbf{w}_k^T \mathbf{x} = 0$.

Hence, our model assumes that the conditional distribution of y given $\mathbf{x}$ is given by

$$
\begin{aligned}
p(y = i \middle| \mathbf{x}; \mathbf{w}) &= \phi_i \\
&= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\
&= \frac{e^{\mathbf{w}_i^T \mathbf{x}}}{\sum_{j=1}^{k} e^{\mathbf{w}_j^T \mathbf{x}}}
\end{aligned}
$$

This model, which applies to classification problems where $y \in \{1, 2, \ldots, k\}$, is called softmax regression.

It is a generalization of logistic regression.

$$
\begin{aligned}
h_{\mathbf{w}}(\mathbf{x}) &= E[T(y)|\mathbf{x};\mathbf{w}] \\
&= E\left[\left.\begin{array}{c} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{array}\right|\mathbf{x};\mathbf{w}\right] \\
&= \left[\begin{array}{c} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{array}\right]
\end{aligned}
$$

$$
\begin{aligned}
h_{\mathbf{w}}(\mathbf{x}) &= E[T(y)|\mathbf{x};\mathbf{w}] \\
&= \left[\begin{array}{c} \dfrac{\exp(\mathbf{w}_1^T\mathbf{x})}{\sum_{j=1}^{k}\exp(\mathbf{w}_j^T\mathbf{x})} \\ \dfrac{\exp(\mathbf{w}_2^T\mathbf{x})}{\sum_{j=1}^{k}\exp(\mathbf{w}_j^T\mathbf{x})} \\ \vdots \\ \dfrac{\exp(\mathbf{w}_{k-1}^T\mathbf{x})}{\sum_{j=1}^{k}\exp(\mathbf{w}_j^T\mathbf{x})} \end{array}\right]
\end{aligned}
$$

In other words, our hypothesis will output the estimated probability that $p(y = i|\mathbf{x}; \mathbf{w})$, for every value of $i = 1, ..., k$

Even though $h_{\mathbf{w}}(\mathbf{x})$ as defined above is only $k-1$ dimensional, clearly

$$p(y = k|\mathbf{x}; \mathbf{w}) = 1 - \sum_{i=1}^{k-1} \phi_i$$