Generalized Linear Models

In the regression example, we had $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$, and in the classification one, $y|x;\theta \sim Bernoulli(\phi)$, where for some appropriate definitions of μ and ϕ as functions of x and θ.

In this section, we will show that both of these methods 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.

1. The exponential family

To work our way up to GLMs, we will begin by defining exponential family distributions. We say that a class of distributions is in the exponential family if it can be written in the form

$$p(y;\eta) = b(y)\, exp(\, \eta^T T(y) - a(\eta)\,) \quad - \quad (1)$$

Here, η is called the **natural parameter** (also called the **canonical parameter**) of the distribution; T(y) is the **sufficient statistic** (for the distributions we consider, it will often be the case that T(y) = y); and $a(\eta)$ is the **log partition function**. The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure that 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.

We now show that the Bernoulli and Gaussian distributions are examples of exponentional family distributions.

1.1 Bernoulli distribution

The Bernoulli distribution with mean ϕ, written Bernoulli(ϕ), specifies a distribution over $y \in \{0, 1\}$, so that $p(y = 1; \phi) = \phi$; $p(y = 0; \phi) = 1 - \phi$. As we varying ϕ, we obtain Bernoulli distributions with different means. We now show that this class of Bernoulli distributions, ones obtained by varying ϕ, is in exponential family; i.e., that there is a choice of T, a and b so that Equation (1) becomes exactly the class of Bernoulli distributions.

We write the Bernoulli distribution as:

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

Thus, the natural parameter is given by $\eta = log(\frac{\phi}{1-\phi})$

Interestingly, if we invert this definition for η by solving for ϕ in terms of η, we obtain $\phi = \frac{1}{1+e^{-\eta}}$. This is the familiar sigmoid function! This will come up again when we derive logistic regression as a GLM. To compelete the formulation of the Bernoulli distribution as an exponential family distribution, we also have

$$
\begin{aligned}
T(y) &= y \\
a(\eta) &= -log(1-\phi) \\
&= -log(1-\frac{1}{1+e^{-\eta}}) \\
&= log(1+e^{\eta}) \\
b(y) &= 1
\end{aligned}
$$

This shows that the Bernoulli distribution can be written in the form of Equation (1), using an appropriate choice of T, a and b.

1.2 Gaussian distribution

Lets now move on to consider the Gaussian distribution. Recall that, when deriving liner regression, the value of σ^2 had no effect on our final choice of θ and $h_\theta(x)$. Thus, we can choose an arbitrary value for σ^2 without changing anything. To simplify the derivation below, let's set $\sigma^2 = 1$. We then have:

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}} exp(\,-\frac{1}{2}(y-\mu)^2\,) \\
&= \frac{1}{\sqrt{2\pi}} exp(\,-\frac{1}{2}y^2\,) \cdot exp(\,\mu y - \frac{1}{2}\mu^2\,)
\end{aligned}
$$

Thus, we see that Gaussian is in the exponential family, with

$$
\begin{aligned}
\eta &= \mu \\
T(y) &= y \\
a(\eta) &= \frac{1}{2}\mu^2 \\
&= \frac{1}{2}\eta^2 \\
b(y) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}y^2)
\end{aligned}
$$

There're many other distributions that are members of the exponential family: The multinomial (which we'll see later), the Poisson (for modelling count-data); the gamma and exponential (for modelling continuous, non-negative random variables, such as time-intervals); the beta and the Dirichlet (for distributions over probabilities); and many more. In the next section, we will describe a general "recipe" for constructing models in which y (given x and θ) comes from any of these distributions.

2. Constructing GLMs

2.1 小结

在普通线性回归中，我们假设被解释变量 y 是连续的且服从正态分布，同时，被解释变量 y 的期望与解释变量 x 之间的关系是线性关系。

然而在实际中，被解释变量 y 有可能是离散的，而且有可能不服从正态分布。因此，广义线性模型对普通线性回归进行了推广，放宽了普通线性回归的假设。首先，被解释变量 y 的分布属于某一指数分布族。有很多我们所熟悉的分布都属于指数分布族，包括正态分布、泊松分布、二项分布等等。其次，被解释变量 y 的期望的函数（即 η，link function）与解释变量 x 之间的关系为线性关系。

总之，广义线性模型本质上还是一个线性模型。我们推广的只是被解释变量 y 的分布。

2.2 Assumptions

Suppose you would like to build a model to estimate the number y of customers arriving in your store (or number of page-views on your website) in any given hour, based on certain features x such as store promotions, recent advertising, weather, day-of-week, etc. We know that the Poisson distribution usually gives a good model for numbers of visitors. Knowing this, how can we come up with a model for our problem? Fortunately, the Poisson is an exponential family distribution, so we can apply a Gerneralized Liner Model (GLM). In this section, we will describe a method for constructing GLM models for problems such as these.

More generally, consider a classification or regression problem where we would like to predict the value of some random variable y as a function of x. To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given x and about our model:

1. $y \mid x; \theta \sim ExponentialFamily(\eta)$. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.
2. Given x, our goal is to predict the expected value of $T(y)$. In most of our examples, we will have $T(y) = y$, so this means we would like the prediction $h(x)$ output by our learned hypothesis h to satisfy $h(x) = E[y|x]$. (Note that this assumption is satisfied in the choices for $h_\theta(x)$ for both logistic regression and linear regression, in logistic regression, we had $h_\theta(x) = p(y = 1|x; \theta) = 0) = 0 \cdot p(y = 0|x; \theta) + 1 \cdot p(y = 1|x; \theta) = E[y|x; \theta]$
3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$. (Or, if η is vector-valued, then $\eta_i = \theta_i^T x$)

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 desiging 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.

2.3 Ordinary Least Squares

To show that ordinary least squares is a special case of the GLM family of models, consider the setting where the target variale y (also called the **response variable** in GLM terminology) is continuous, and we model the conditional distribution of y given x as a Gaussian $\mathcal{N}(\mu, \sigma^2)$. (Here, μ may depend x.) So, we let the $ExponentialFamily(\eta)$ distribution above be the Gaussian distribution. As we saw previously, in the formulation of the Gaussian as an exponential family distribution, we had $\mu = \eta$. So, we have

$$\begin{aligned} h_\theta(x) &= E[y|x; \theta] \\ &= \mu \\ &= \eta \\ &= \theta^T x \end{aligned}$$

The first equality follows Assumption 2, above.

The second equality follows from the fact that $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$, and so its expected value is given by μ.

The third equality follows from Assumption 1 (and our earlier derivation showing that $\mu = \eta$ in the formulation of the Gaussian as an exponential family distribution)

The last equality follows from Assumption 3.

2.3 Logistic Regression

We now consider 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 x. In our formulation of the Bernoulli distribution as an exponential family distribution, we had $\phi = \frac{1}{1+e^{-\eta}}$. Furthermore, note that if $y|x; \theta \sim Bernoulli(\phi)$, then $E[y|x; \theta] = \phi$. So, following a similar derivation as the one for ordinary least squares, we get:

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

So, this gives us hypothesis functions of the form $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$. If you are previously wondering how we came up with the form of the logistic function $\frac{1}{1+e^{-z}}$, this gives one answer: Once we assume that y conditioned on x is Bernoulli, it arises as a consequence of the definition of GLMs and exponential family distributions.

```
In [ ]:
```