DS 102 Discussion 8

Monday, April 6, 2020

1. **Generalized Linear Models**

In this discussion, we'll review some of the time-series models presented in Lecture 22 (April 2) and see how they are examples of **generalized linear models** (GLMs). As their name implies, GLMs generalize the linear regression problem we know and love to model situations where 1) the distribution of the output given the input is not simply Gaussian and 2) the mean of that distribution is not simply a linear function of the input.

A quick refresher on the linear regression model we're familiar with: we have d-dimensional input T, and the scalar output X is given by

$$X = \beta^T T + \epsilon \tag{1}$$

where d-dimensional β are parameters, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise term that may capture, for example, measurement noise or variance due to other unobserved factors (we use X to denote the output, to be consistent with Section 22.4 of Lecture 22). Equivalently, we can state that the conditional distribution of X given T is:

$$X \mid T \sim \mathcal{N}(\beta^T T, \sigma^2). \tag{2}$$

There are two key types of assumptions in (2) that GLMs relax.

1. **Output distribution**. In the above linear regression model, we assumed that given the input, the output variable is distributed as a Gaussian random variable. Furthermore, we assumed that the input only determines the mean of this conditional distribution. That is, we have

$$X \mid T \sim \mathcal{N}(\mu(T), \sigma^2 I) \tag{3}$$

where $\mu(T) := \mathbb{E}[X \mid T]$ is called the **mean function** because it describes, as a function of T, the mean of the conditional distribution of X given T.

2. **Link function**. In the above linear regression model, we further assumed that the mean function is simply a linear function of the input: $\mu(T) = \beta^T T$.

To relax these two assumptions and model more interesting phenomena, for a GLM we need to specify two components:

1. **Output distribution**. No longer shackled to the Gaussian, we can pick other distributions to model the conditional distribution of the output X given the input T, depending on what is appropriate for the application. However, as in the linear regression case, we keep the assumption that the input only determines the mean $\mathbb{E}[X \mid T]$ of that distribution, and does not affect any other parameters.

2. **Link function**. To describe the mean function, we can pick any invertible function of a linear function of the inputs. That is, we have

$$g(\mu(T)) = \beta^T T \tag{4}$$

where g is called the **link function**. (In linear regression, the link function is just the identity.)

We'll now review an application similar to the one in Section 22.4 of Lecture 22, and show how it can be cast as a GLM. Let X_t, the observed output variable, denote the number of COVID-19 hospitalizations on day t. We'd like to model the relationship between the output X_t and the input t.

(a) As noted in the lecture, epidemiology tells us that in some settings, exponential growth for the mean of X_t is reasonable:

$$\mathbb{E}[X_t] = \alpha \exp(\gamma t). \tag{5}$$

Find an appropriate link function. That is, as shown in (4) find a function of the mean function that is equivalent to just a linear function of the input.

> **Solution:** Our mean function is given as
>
> $$\mu(T) = \mathbb{E}[X_T \mid T] = \mathbb{E}[X_T] = \gamma \exp(rT).$$
>
> We want to find a function g such that (4) holds for some parameters β. We can take the log:
>
> $$\log \mu(T) = \log \alpha + \gamma T = \beta_0 + \beta_1 T. \tag{6}$$
>
> where $\beta_0 = \log \alpha$ and $\beta_1 = \gamma$.

(b) To complete the specification of the GLM, we need an output distribution. Since we are modeling integer-valued X_t, what are natural choices for the conditional distribution of X_T given T?

> **Solution:** The Poisson or negative binomial distributions.

(c) Show how the GLM we've described is equivalent to the model in Section 22.4 of Lecture 22:

$$X_t \sim \text{Poisson}(Z_t) \tag{7}$$
$$Z_{t+1} = (1 + r)Z_t. \tag{8}$$

That is, express α and γ in our GLM as functions of r and Z_0.

Solution: Unrolling the recursion in the model gives

$$Z_t = (1+r)^t Z_0$$

as the mean function. Substituting in our exponential growth expression for the mean function gives

$$\begin{aligned} \alpha \exp(\gamma t) &= (1+r)^t Z_0 && (9) \\ \alpha &= Z_0 && (10) \\ \gamma &= \log(1+r). && (11) \end{aligned}$$

2. **GLM for continuous data**. The examples in Lecture 22 involved discrete data. That is, the observed output variable X_t was always a positive integer (*e.g.*, number of hospitalizations on day t). However, what happens if X_t can be any real number? Then the Poisson GLM wouldn't make much sense as a model.

   Consider the following example of a time series with a continuous output. Suppose a rocket has been launched in Florida, and we in California start observing the rocket at time $t = 0$. We want to measure the rocket's distance from Earth at some time t in the future. At each time step, we obtain a noisy measurement of this distance.

   Let $\beta_0 \in \mathbb{R}$ be the initial distance from Earth (in miles) of the rocket when we started observing it at time $t = 0$. Suppose the rocket is moving away from Earth at a constant rate of $\beta_1 \in \mathbb{R}$ miles per time step t. Let X_t denote our observation of the rocket's distance from Earth, which is noisy due to weather, measurement error, etc. Assume that for all t our observation noise is normally distributed with a standard deviation of $\sigma = 50$ miles.

   (a) What is the distribution of X_t, the observed distance of the rocket at time t? Write this distribution in terms of β_0, β_1, and σ.

   Solution: We can model the observed distance of the rocket at time t as:

   $$X_t \sim \mathcal{N}(\beta_0 + t\beta_1, \sigma^2)$$

   That is, without noise in our observations, the actual position of the rocket at time t is $\beta_0 + t\beta_1$. The position that we observe is centered around this true position with a standard deviation of $\sigma = 50$ miles.

   (b) Suppose we don't know β_0 or β_1, and we observe $X_0, X_1, ..., X_T$ positions of the rocket from California. Our goal is to predict the future positions of the rocket by estimating β_0 and β_1. First, we'll cast our model as a GLM with output X_t and input t. What is the output distribution and link function?

Solution: This is a linear regression model. That is, the model $X_t \sim \mathcal{N}(\beta_0 + t\beta_1, \sigma)$ is equivalent to a GLM with output distribution $\mathcal{N}(\mu(t), \sigma^2)$, where $\mu(t) = \beta_0 + \beta_1 t$. The link function is the identity function.

(c) Given the data $X_0, X_1, ..., X_T$, how might we solve for β_0 and β_1 in the above GLM?

Solution: We can solve for the maximum likelihood estimates (MLEs) of β_0 and β_1 over the observed $X_0, X_1, ..., X_T$. For GLMs in general, there are various algorithms for maximum-likelihood estimation of the parameters of different output distributions and link functions.

For the linear regression model in particular, however, the MLE is simply the OLS estimate of β_0 and β_1 with inputs $t = 0, 1, \ldots, T$ and outputs $X_0, X_1, \ldots, X_T$.