Generalized Linear Models

Viswa Marepalli

May 2025

This document accompanies my article on GLMs, which can be found here.

Introduction

This document aims to have full derivations on some of the GLM lecture notes from CS229. The lecture notes I'm referencing from can be found here. This document will have minimal explanation. Refer to my article for full explanations (or even better, watch the CS229 lecture).

Bernoulli Distribution to Logistic Regression

Though I did write a fairly complete derivation on my article, it's here for completeness. In fairness, most of these derivations are just some logarithm rules.

$$p(y;\varphi) = \varphi^y (1-\varphi)^{1-y} \quad [1]$$

$$= \exp(\ln(\varphi^y (1-\varphi)^{1-y})) \quad [2]$$

$$= \exp(\ln(\varphi^y) + \ln((1-\varphi)^{1-y})) \quad [3]$$

$$= \exp(y\ln(\varphi) + (1-y)\ln(1-\varphi)) \quad [4]$$

$$= \exp(y\ln(\varphi) + \ln(1-\varphi) - y\ln(1-\varphi)) \quad [5]$$

$$= \exp(y(\ln(\varphi) - \ln(1-\varphi)) + \ln(1-\varphi)) \quad [6]$$

$$= \exp\left(y\ln\left(\frac{\varphi}{1-\varphi}\right) + \ln(1-\varphi)\right) \quad [7]$$

$$b(y) = 1 \quad [8]$$

$$\eta = \ln\left(\frac{\varphi}{1-\varphi}\right) \quad [9]$$

$$a(\eta) = -\ln(1-\varphi) \quad [10]$$

Now, we'll write $a(\eta)$ in terms of η rather than φ. We'll do this by re-arranging the definition we have for η for φ.

$$\eta = \ln\left(\frac{\varphi}{1-\varphi}\right) \quad [11]$$

$$e^\eta = e^{\ln\left(\frac{\varphi}{1-\varphi}\right)} = \frac{\varphi}{1-\varphi} \quad [12]$$

$$e^\eta(1-\varphi) = \varphi \quad [13]$$

$$e^\eta - \varphi e^\eta = \varphi \quad [14]$$

$$\varphi + \varphi e^\eta = e^\eta \quad [15]$$

$$\varphi(1+e^\eta) = e^\eta \quad [16]$$

$$\varphi = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}} \quad [17]$$

Therefore:

$$a(\eta) = -\ln\left(1 - \frac{1}{1+e^{-\eta}}\right) \quad [18]$$

$$= -\ln\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right) = -\ln\left(\frac{1}{1+e^\eta}\right) \quad [19]$$

$$= \ln(1+e^\eta) \quad [20]$$

Finding our hypothesis function now, using the fact that $E(y;\eta) = \frac{\partial}{\partial\eta}[a(\eta)]$:

$$\frac{\partial}{\partial\eta}(\ln(1+e^\eta)) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}} \quad [21]$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \quad [22]$$

Beautiful, right?

Gaussian Distribution to Linear Regression

Version 1

This derivation assumes a variance $\sigma^2 = 1$. The next derivation won't assume a fixed variance.

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \quad [23]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y^2 - 2\mu y + \mu^2)\right) \quad [24]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2 + \mu y - \frac{1}{2}\mu^2\right) \quad [25]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \exp\left(\mu y - \frac{1}{2}\mu^2\right) \quad [26]$$

$$b(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \quad [27]$$

$$\eta = \mu \quad [28]$$

$$a(\eta) = \frac{1}{2}\mu^2 \quad [29]$$

$$E[y;\mu] = \frac{\partial}{\partial \eta}\left(\frac{1}{2}\mu^2\right) \quad [30]$$

Substituting $\eta = \mu$:

$$E(y;\mu) = \frac{\partial}{\partial \eta}\left(\frac{1}{2}\eta^2\right) = \eta = \theta^T x \quad [31]$$

$$h_\theta(x) = \theta^T x \quad [32]$$

Those of you who've studied CS229, will know that this is the familiar hypothesis function that was introduced within the first set of lectures!

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2}\right) \quad [33]$$

Version 2

This derivation introduces some ideas that aren't present in the CS229 lecture notes. This may seem verbose, but it's the way I derived it, so bear with me.

We'll first begin by defining the complete Gaussian distribution:

$$p(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad [34]$$

Notice how there's now two variables, μ and σ^2, that the PDF depends on. This means our natural parameter η is now a 2 dimensional vector, and can be written as $\vec{\eta} = \binom{\eta_1}{\eta_2}$. Remember, η and the sufficient statistic $T(y)$ are both of the same dimensions ($\eta, T(y) \in \mathbb{R}^n$), therefore $T(y)$ can be written as $\vec{T}(y) = \binom{T_1(y)}{T_2(y)}$.

Expanding the inner component of exp(...):

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-y^2 + 2\mu y - \mu^2}{2\sigma^2}\right) \quad [35]$$

Splitting up the fraction:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \quad [36]$$

Hopefully you can see that our natural parameters come out pretty clearly, but our base measure $b(y)$ is a little shrewd.

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \tag{37}$$

It looks like we have our natural parameters η_1 and η_2, and our sufficient statistics $T_1(y)$ and $T_2(y)$, but this still can't be concrete until we can manipulate this to also find $a(\eta)$, so let's do that.

Another way of writing exponential family form is:

$$p(y;\eta) = b(y)\frac{\exp(\eta^T T(y))}{\exp(a(\eta))} \tag{38}$$

I'll start by rewriting Equation 37 slightly:

$$\frac{1}{\sqrt{2\pi}} \frac{\exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2}\right)}{\sigma\left(\exp\left(\frac{\mu^2}{2\sigma^2}\right)\right)} \tag{39}$$

It is important to notice that σ has switched from the first fraction to the second fraction.

The denominator on the second fraction can be re-written as so using logarithms to bring σ into the exponent:

$$\exp(\ln(\sigma)) \exp\left(\frac{\mu^2}{2\sigma^2}\right) = \exp\left(\frac{\mu^2}{2\sigma^2} + \ln(\sigma)\right) \tag{40}$$

The final manipulated form of the Gaussian Distribution is then as follows:

$$p(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2}\right)}{\exp\left(\frac{\mu^2}{2\sigma^2} + \ln(\sigma)\right)} \tag{41}$$

From Equation 38 it's hopefully pretty clear to see our canonical parameters:

$$b(y) = \frac{1}{\sqrt{2\pi}} \tag{42}$$

$$\eta_1 = \frac{\mu}{\sigma^2} \Longleftrightarrow T_1(y) = y \tag{43}$$

$$\eta_2 = -\frac{1}{2\sigma^2} \Longleftrightarrow T_2(y) = y^2 \tag{44}$$

$$a(\eta) = \frac{\mu^2}{2\sigma^2} + \ln(\sigma) \tag{45}$$

Rewriting $a(\eta)$ in terms of η_1 and η_2 (an easy but fun algebra task for the reader). Keep in mind, the component inside the ln is not imaginary, $\eta_2 < 0$, therefore it simply evaluates to be positive:

$$a(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{\ln(-2\eta_2)}{2} \tag{46}$$

Now, we know to find our hypothesis function, we must find $E(y;\eta) = \frac{\partial}{\partial\eta}[a(\eta)]$, but what if we have a n-dimensional η. Then, we must compute the partials with respect to each component:

$$E[T_1(y);\eta] = E[y;\eta] = \frac{\partial}{\partial\eta_1}[a(\eta)] = -\frac{\eta_1}{2\eta_2} = -\frac{\frac{\mu}{\sigma^2}}{2\left(-\frac{1}{2\sigma^2}\right)} = \mu \quad [47]$$

This is what we get in the original Version 1 derivation, and is also what is expected.

$$E[T_2(y);\eta] = E[y^2;\eta] = \frac{\partial}{\partial\eta_2}[a(\eta)] = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2 \quad [48]$$

This has been derived as an additional step of validation.

Rewriting Equation 48 in terms of the variance σ^2:

$$\sigma^2 = E[y^2;\eta] - \mu^2 = E[y^2;\eta] - E[y;\eta]^2 \quad [49]$$

This follows the traditional relationship of variance $\sigma^2 = E(X^2) - [E(X)]^2$, convincing us that our derivation has worked.

Formally now:

$$E[y;\eta] = \mu = \eta = \theta^T x \quad [50]$$