Bayesian Generalized Linear Model

Austen

February 19, 2025

1 Generalized Linear Model (GLM)

We have a dataset $D := \{x_i, y_i\}_{i=1}^n$, where x_i is an hour in the day (mapped to a spline basis for flexibility) and y_i is the observed number of cars in that hour. To model the non-linear relationships between time and traffic, the dataset is transformed onto a spline basis. After some trial and error, we discovered a spline basis with 15 knots of second degree splines (resulting regression has 15 DoFs) captured the traffic dynamics well without overfitting. Although y_i is technically discrete and non-negative, we treat it as a continuous, positive quantity.

1.1 Design Matrix and Link Function

Let $X \in \mathbb{R}^{n\times p}$ be the design matrix formed by mapping the raw inputs x_i through spline basis functions (we mentioned 15 knots of a second-degree spline, so $p = 15$). The predicted mean of our response is given by

$$\mu_i = \exp\big(x_i^\top w\big),$$

i.e. a log link function. Thus, our GLM prediction for each sample is

$$\hat{y}(w, x_i) = \exp\big(x_i^\top w\big).$$

1.2 Gamma Distribution

A Gamma-distributed outcome Y_i with shape parameter α (sometimes called k) and mean μ_i can be parameterized in various ways. One common parametrization uses (α, β) where β is a rate parameter, and the mean is α/β. Alternatively, we can specify the shape α and the mean μ_i, in which case the rate is $\beta_i = \alpha/\mu_i$.

The probability density function (pdf) for $y_i > 0$ using shape–rate (α, β_i) is:

$$p(y_i \mid \alpha, \beta_i) = \frac{\beta_i^\alpha}{\Gamma(\alpha)}\, y_i^{\alpha-1}\, e^{-\beta_i y_i}.$$

If we tie $\beta_i = \alpha/\mu_i$, and $\mu_i = \exp(x_i^\top w)$, then:

$$p(y_i \mid w, \alpha) \;=\; \frac{\left(\frac{\alpha}{\mu_i}\right)^\alpha}{\Gamma(\alpha)}\, y_i^{\alpha-1}\, \exp\Big(-\tfrac{\alpha}{\mu_i}\, y_i\Big).$$

Hence, for each data point (x_i, y_i):

$$p(y_i \mid x_i, w, \alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)\,\mu_i^\alpha}\, y_i^{\alpha-1}\, \exp\Big(-\alpha\frac{y_i}{\mu_i}\Big), \quad \text{where} \quad \mu_i = \exp(x_i^\top w).$$

2 Bayesian Formulation

2.1 Posterior Derivation

Using conditional independence, the likelihood for all observations D is:

$$p(D \mid w, \alpha) = \prod_{i=1}^n p(y_i \mid x_i, w, \alpha) = \prod_{i=1}^n \frac{\alpha^\alpha}{\Gamma(\alpha)\,\mu_i^\alpha}\, y_i^{\alpha-1}\, \exp\Big(-\alpha\,\frac{y_i}{\mu_i}\Big).$$

We choose a prior $p(w)$ for the weight vector w. A simple (but commonly used) choice is a Gaussian prior centered at zero:

$$p(w) = \mathcal{N}\big(w \mid 0,\, \Sigma_w\big)$$

for some covariance Σ_w. We also need a prior on α if it is unknown, say a Gamma prior:

$$p(\alpha) = \text{Gamma}(\alpha \mid a_0, b_0).$$

If α is assumed fixed or known, then $p(\alpha)$ reduces to a delta function.

By Bayes' theorem, the posterior is

$$p(w, \alpha \mid D) \;=\; \frac{p(D \mid w, \alpha)\, p(w)\, p(\alpha)}{p(D)}.$$

Since $p(D)$ is just a normalizing constant, we typically write:

$$p(w, \alpha \mid D) \;\propto\; p(w)\, p(\alpha) \prod_{i=1}^{n} p(y_i \mid x_i, w, \alpha).$$

Log-posterior form. Taking logarithms and omitting constants w.r.t. (w, α):

$$\log p(w, \alpha \mid D) \;=\; \log p(w) \;+\; \log p(\alpha) \;+\; \sum_{i=1}^{n} \left[\alpha \ln(\alpha) - \ln \Gamma(\alpha) - \alpha \ln \mu_i + (\alpha - 1) \ln y_i - \alpha \, \frac{y_i}{\mu_i}\right].$$

Recall $\mu_i = \exp(x_i^\top w)$, so $\ln(\mu_i) = x_i^\top w$. For the prior terms:

$$\log p(w) = -\tfrac{1}{2}\,(w^\top \Sigma_w^{-1} w), \quad \text{and e.g.} \quad \log p(\alpha) = (a_0 - 1)\ln(\alpha) - b_0 \alpha \quad \text{(if Gamma prior)}.$$

We then have a posterior in $(p+1)$ dimensions (if α is unknown and w is p-dimensional). This distribution typically has no closed-form solution.

2.2 Inference and Prediction

We can approximate $p(w, \alpha \mid D)$ using:

- **MCMC sampling** (e.g., Metropolis-Hastings, Hamiltonian Monte Carlo) to obtain posterior draws.
- **Variational methods** to get a tractable approximation.
- **Laplace approximation** around the posterior mode, if p is not too large.

Once we have (approximate) draws $(w^{(k)}, \alpha^{(k)})$, we can form the *predictive distribution* for a new input $\hat{x}$ as follows. The Gamma distribution for a new observation $\hat{y}$ given parameters is

$$p(\hat{y} \mid \hat{x}, w, \alpha) = \frac{\left(\frac{\alpha}{\mu}\right)^{\alpha}}{\Gamma(\alpha)}\, \hat{y}^{\,\alpha - 1}\, \exp\!\Big(-\alpha\, \frac{\hat{y}}{\mu}\Big), \quad \text{where } \mu = \exp\big(\hat{x}^\top w\big).$$

So the Bayesian predictive distribution integrates out the parameters:

$$p(\hat{y} \mid \hat{x}, D) = \int p(\hat{y} \mid \hat{x}, w, \alpha)\, p(w, \alpha \mid D)\, dw\, d\alpha.$$

In practice, we approximate via Monte Carlo:

$$p(\hat{y} \mid \hat{x}, D) \;\approx\; \frac{1}{K} \sum_{k=1}^{K} p\big(\hat{y} \mid \hat{x}, w^{(k)}, \alpha^{(k)}\big),$$

where $(w^{(k)}, \alpha^{(k)})$ are samples drawn from $p(w, \alpha \mid D)$.

3 Conclusion

The main decision to make now seems to be how to approximate the posterior in order to generate samples. Additionally, experimenting with different priors may also be interesting. In my previous attempt I was using the standard Gaussian distribution (Identity link function) with a Laplace approximation. In that setting, the posterior approximation is Gaussian with the trained parameters as the mean and some scaling of the inverse Hessian for covariance (Hessian of the minimization objective w.r.t parameters). This didn't work well and now that I want to try and use a more suitable distribution the Laplace approximation derivation may be different.