Generalized Linear Models
(Conditional Models via Exp-Family)

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Generalized Linear Models

- (Probabilistic) Linear Regression: when response $\boldsymbol{y}$ is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic Regression: when response $\boldsymbol{y}$ is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

$$\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$$

Note: Probabilistic Linear Regression and Logistic Regression are also GLMs

- In both, the model depends on the inputs $\boldsymbol{x}$ via a linear model $\boldsymbol{w}^\top \boldsymbol{x}$
- Generalized Linear Models (GLM) allow modeling other types of responses, e.g.,
  - Counts (e.g., predicting the hourly hits on a website)
  - Positive reals (e.g., predicting depth of different pixels in a scene, or stock prices)
  - Fractions between 0 and 1 (e.g., predicting proportion of crude oil convertible to gasoline)
- Note: Can convert responses to real values and apply standard regression, but it is better to model them directly (e.g., for better interpretability of the model)

Generalized Linear Models: Formally

The reason why GLMs can model a wide variety of responses

- GLMs model the response using an exponential family distribution

Response $\boldsymbol{y}$ assumed univariate but vector GLMs also exist

Scalar natural param (depends on input $\boldsymbol{x}$)

Scalar suff-stats $\phi(y) = y$

$$p(y|\eta) = h(y)\exp(\eta y - A(\eta))$$

- The inputs $\boldsymbol{x}$ only appear via a linear model $\xi = \boldsymbol{w}^\top \boldsymbol{x}$ and the overall pipeline is

For prob. linear regression with Gaussian lik, f is identify since mean $\mu = \mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{w}^\top \boldsymbol{x}$

For logistic regression, f is sigmoid since mean $\mu = \mathbb{E}[y|\boldsymbol{x}] = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$

Response Function

$$w,\ x \to \xi \xrightarrow{f} \mu \xrightarrow{\psi} \eta$$

Link Function

For GLM with Canonical Response Function (next slide), $\psi = f^{-1}$ and thus nat. param. $\eta = \xi = \boldsymbol{w}^\top \boldsymbol{x}$

Conditional mean of response $\mu = \mathbb{E}[y|\boldsymbol{x}] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$

Natural parameter $\eta = \psi(\mu)$

f known as "inverse link function" in this case

- Note: Some GLM are represented via exponential <u>dispersion</u> family given by

If σ^2 is fixed, it is the standard exponential family

$$p(y|\eta,\sigma^2) = h(y,\sigma^2)\exp\left[\frac{\eta y - A(\eta)}{\sigma^2}\right]$$

Called the "dispersion parameter"

Examples: Gaussian GLM, Gamma GLM

Recall cumulant results of exp-fam

$\mathbb{E}[y] = A'(\eta)$

$\text{var}[y] = A''(\eta)\sigma^2$

Generalized Linear Models: Examples

- Consider the overdispersed GLMs

$$p(y|\eta,\sigma^2) = h(y,\sigma^2)\exp\left[\frac{\eta y - A(\eta)}{\sigma^2}\right] = \exp\left[\frac{\eta y - A(\eta)}{\sigma^2} + \log h(y,\sigma^2)\right]$$

Note that here we expressed the Gaussian in the overdispersed GLM form unlike how we did it earlier when discussing exp-family

- Consider a linear regression model with Gaussian likelihood

$$p(y|\boldsymbol{x},\boldsymbol{w},\sigma^2) \propto \exp\left[-\frac{(y-\boldsymbol{w}^\top\boldsymbol{x})^2}{2\sigma^2}\right] = \exp\left[-\frac{y^2 + (\boldsymbol{w}^\top\boldsymbol{x})^2 - 2y\boldsymbol{w}^\top\boldsymbol{x}}{2\sigma^2}\right] = \exp\left[\frac{y\boldsymbol{w}^\top\boldsymbol{x} - (\boldsymbol{w}^\top\boldsymbol{x})^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2}\right]$$

- Comparing the expressions, $\eta = \boldsymbol{w}^\top\boldsymbol{x}$, $A(\eta) = \frac{\eta^2}{2}, \log h(y,\sigma^2) = -y^2/2\sigma^2$

- Can likewise express other models for exp-family distributions $p(y|\boldsymbol{x})$
  - Regardless of the form, all will have $\eta = \boldsymbol{w}^\top\boldsymbol{x}$

GLM with Canonical Response Function

- For GLM with Canon Resp Func (a.k.a., canonical GLM)

$$p(y|\eta) = h(y)\exp(\eta y - A(\eta)) = h(y)\exp(y\mathbf{w}^\top \mathbf{x} - A(\eta))$$

The simple form of canonical GLM (nat. param just a linear function $\boldsymbol{w}^\top \boldsymbol{x}$) makes parameter estimation via MLE/MAP easy since gradient and Hessian have simple expressions (though the Hessian may be expensive to compute/invert)

- Consider doing MLE (assuming N i.i.d. responses). The log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n)\exp(y_n \mathbf{w}^\top \mathbf{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \mathbf{w}^\top \sum_{n=1}^{N} y_n \mathbf{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Gradient of log-lik w.r.t. $\boldsymbol{w}$

Exp of suff-stats $\mathbb{E}[y_n]$

Corrective updates for $\boldsymbol{w}$

$$\mathbf{g} = \sum_{n=1}^{N}\left(y_n \mathbf{x}_n - A'(\eta_n)\frac{d\eta_n}{d\mathbf{w}}\right) = \sum_{n=1}^{N}(y_n \mathbf{x}_n - \mu_n \mathbf{x}_n) = \sum_{n=1}^{N}(y_n - \mu_n)\mathbf{x}_n$$

The Hessian can also be shown to be

$$\mathbf{H} = -\nabla \mathbf{g} = \sum_{n=1}^{N} f'(\eta_n)\boldsymbol{x}_n \boldsymbol{x}_n^\top$$

- Note $\mu_n = f(\xi_n) = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and $f = \psi^{-1}$ ("inverse link") depends on the model
  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is sigmoid function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}$
  - Count-valued y (Poisson regression): f is exp, i.e., $\mu_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$
  - Non-negative y (gamma regression): f is inverse negative i.e., $\mu_n = -1/(\boldsymbol{w}^\top \boldsymbol{x}_n)$

Fully Bayesian Inference for GLMs

- Most GLMs, except linear regression with Gaussian lik. and Gaussian prior, do not have conjugate pairs of likelihood and priors (recall logistic regression)

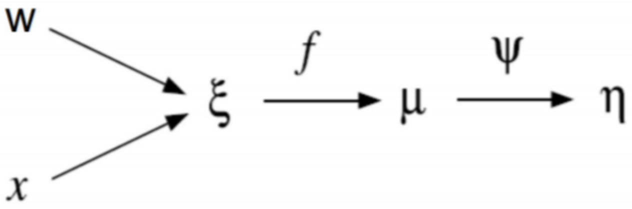
- Posterior over the weight vector $\boldsymbol{w}$ is intractable

- Approximate inference methods needed, e.g.,

  - Laplace approximation (have already seen): Easily applicable since derivatives (first and second) can be easily computed (note that we need $\boldsymbol{w}_{MAP}$ and Hessian)

  - MCMC or variational inference (will see later)

Various Types of GLMs



Type of response	Type of GLM	Link Function Ψ	Response Function f (Inv Link Func if canon. GLM) (Operates on $\xi = \boldsymbol{w}^\top \boldsymbol{x}$)	Activation
Real	Gaussian	Identity	Identity	Linear
Binary	Logistic	Log-odds: $\log \frac{\mu}{1-\mu}$	Sigmoid	Sigmoid
Binary	Probit	Inv CDF: $\Phi^{-1}(\mu)$	Φ (CDF of N(0,1))	Probit
Categorical	Multinoulli	Log-odds: $\log \frac{\mu_k}{1-\mu_k}$	Softmax	Softmax
Count	Poisson	$\log \mu$	exp	
Non-negative real	gamma	Negative of inverse	Negative of inverse	
Binary	Gumbel	Gumbel Inv CDF: log(-log())	Gumbel CDF: exp(-exp(-))	

.. and several others (exponential, inverse Gaussian, Binomial, Tweedie, etc)

Coming Up

- Generative models for supervised learning