

CEE 697M: Probabilistic Machine Learning

M2 Linear Methods: Splines, GAMs and GLMs

Jimi Oke

UMassAmherst

College of Engineering

October 7, 2025

Outline

- ① Exponential family
- ② GLMs
- ③ Fitting a GLM

The exponential family

A probability distribution belongs to the exponential family if its density can be modeled as:

$$p(\mathbf{y}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{y}) \exp [\boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y})] = h(\mathbf{y}) \exp [\boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y}) - A(\boldsymbol{\eta})] \quad (1)$$

where:

- $Z(\boldsymbol{\eta})$ is the partition function (normalization constant)
- $h(\mathbf{y})$ is the base measure (scaling constant; typically 1)
- $\boldsymbol{\eta}$ are the natural/canonical parameters
- $\mathcal{T}(\mathbf{y})$ are the sufficient statistics
- $A(\boldsymbol{\eta}) = \ln Z(\boldsymbol{\eta})$ is the log-partition function

The log-likelihood is then given by:

$$\log p(\mathbf{y}|\boldsymbol{\eta}) = \log h(\mathbf{y}) + \boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y}) - A(\boldsymbol{\eta}) + \text{const} \quad (2)$$

Properties of exponential family

- Generalization: we define $\boldsymbol{\eta} = f(\boldsymbol{\phi})$, thus:

$$p(\mathbf{y}|\boldsymbol{\phi}) = h(\mathbf{y}) \exp \left[f(\boldsymbol{\phi})^\top \mathcal{T}(\mathbf{y}) - A(f(\boldsymbol{\phi})) \right] \quad (3)$$

- If $f(\boldsymbol{\phi})$ is nonlinear, then the model is in the curved exponential family
- If $\boldsymbol{\eta} = f(\boldsymbol{\phi}) = \boldsymbol{\phi}$, the model is in **canonical form**
- If $\mathcal{T}(\mathbf{y}) = \mathbf{y}$, the model is in the natural exponential family

$$p(\mathbf{y}|\boldsymbol{\eta}) = h(\mathbf{y}) \exp \left[\boldsymbol{\eta}^\top \mathbf{y} - A(\boldsymbol{\eta}) \right] \quad (4)$$

Bernoulli distribution in exponential family form (1/2)

The Bernoulli distribution is given by:

$$p(y|\mu) = \mu^y(1 - \mu)^{1-y}, \quad y \in \{0, 1\}, \quad 0 < \mu < 1 \quad (5)$$

where $\mu = \mathbb{E}(y)$ is the probability of success. Rewriting:

$$\begin{aligned} p(y|\mu) &= (1 - \mu) \left(\frac{\mu}{1 - \mu} \right)^y = (1 - \mu) \exp \left[y \log \left(\frac{\mu}{1 - \mu} \right) \right] \\ &= (1 - \mu) \exp \left[y \log \left(\frac{\mu}{1 - \mu} \right) - 0 \right] \end{aligned}$$

Comparing to the exponential family form:

$$\begin{aligned} h(y) &= 1 - \mu \quad (\text{base measure}) \\ \mathcal{T}(y) &= y \quad (\text{sufficient statistic}) \\ \eta &= \log \left(\frac{\mu}{1 - \mu} \right) \quad (\text{natural parameter}) \\ A(\eta) &= 0 \quad (\text{log-partition function}) \end{aligned}$$

Cumulant generating function

- Cumulants $\kappa_n(\mathbf{y})$ are functions of the central moments of a distribution
- For example, $\kappa_1(\mathbf{y}) = \mathbb{E}(\mathbf{y})$ and $\kappa_2(\mathbf{y}) = \mathbb{V}(\mathbf{y})$
- Higher order cumulants are polynomial functions of the central moments
- The cumulants of a distribution are defined by the cumulant generating function (CGF):

$$K_{\mathbf{y}}(t) = \log \mathbb{E}(\exp(t\mathbf{y})) \quad (6)$$

where $\mathbb{E}(\exp(t\mathbf{y}))$ is the moment generating function (MGF) of \mathbf{x}

- In the exponential family, the log-partition function $A(\boldsymbol{\eta})$ is the CGF of the sufficient statistics $\mathcal{T}(\mathbf{y})$
- Thus, the cumulants can be obtained by differentiating $A(\boldsymbol{\eta})$:

$$\begin{aligned} \kappa_1(\mathcal{T}(\mathbf{y})) &= \mathbb{E}(\mathcal{T}(\mathbf{y})) = \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) \\ \kappa_2(\mathcal{T}(\mathbf{y})) &= \text{Cov}(\mathcal{T}(\mathbf{y})) = \nabla_{\boldsymbol{\eta}}^2 A(\boldsymbol{\eta}) \end{aligned}$$

Unique global maximum of the likelihood

From the CGF properties, we have:

$$\nabla_{\boldsymbol{\eta}}^2 A(\boldsymbol{\eta}) = \text{Cov}(\mathcal{T}(\mathbf{y})) > 0 \quad (7)$$

This implies that the log-partition function $A(\boldsymbol{\eta})$ is strictly convex. Thus, the log-likelihood

$$\log p(\mathbf{y}|\boldsymbol{\eta}) = \log h(\mathbf{y}) + \boldsymbol{\eta}^\top \mathcal{T}(\mathbf{y}) - A(\boldsymbol{\eta}) + \text{const} \quad (8)$$

is guaranteed to have a unique global maximum.

The generalized linear model (GLM)

- Conventional linear regression models have the form:

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}(y|\mathbf{x}^\top \mathbf{w}, \sigma^2) \quad (9)$$

where

- y_i is a continuous response
- \mathbf{x}_i is a vector of quantitative and/or qualitative explanatory variables
- Generalized linear models (GLMs) were introduced to extend this framework to allow y_i to be modeled by other exponential family distributions besides the normal/Gaussian, e.g.
 - exponential
 - binomial/multinomial (with fixed number of trials)
 - Poisson
- In the GLM framework:
 - The mean of y_i is given by μ_i
 - μ_i can be specified by a nonlinear function of $\mathbf{x}_i^\top \mathbf{w}$
 - Note that the simple linear regression is a special case of GLM in which $\mu_i = \mathbf{x}_i^\top \mathbf{w}$ and y_i follows a Gaussian distribution

GLM formulation

The GLM is a version of the exponential family distribution in which the natural parameters η_n are a **linear function** of the output. It is given by:

$$p(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \exp \left[\frac{y_n \eta_n - A(\eta_n)}{\sigma^2} + \log h(y_n, \sigma^2) \right] \quad (10)$$

where:

- $\eta_n = \mathbf{w}^\top \mathbf{x}_n$ is the natural parameter (input)
- $y_n = \mathcal{T}(y_n)$ is the sufficient statistic
- $A(\eta_n)$ is the log-partition function (or log normalizer)
- $h(y_n, \sigma^2)$ is the base measure
- σ^2 is the dispersion parameter (typically known or set to 1)

Link and mean functions

Recalling that the mean and variance of the sufficient statistics $\mathcal{T}(y_n) = y_n$ are given by the first and second derivatives of the log-partition function $A(\eta_n)$, we have:

$$\mathbb{E}(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = A'(\eta_n) = \ell^{-1}(\eta_n) \quad (11)$$

$$\text{Var}(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = A''(\eta_n) \sigma^2 \quad (12)$$

We define the **mean function** as

$$\mu_n = \ell^{-1}(\eta_n) \quad (13)$$

and the **link function** as its inverse:

$$g(\mu_n) = \ell(\mu_n) \quad (14)$$

The link function is thus the inverse of the mean function.

Linear regression (1/2)

Linear regression has the form:

$$p(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \mathbf{w}^\top \mathbf{x}_n)^2\right) \quad (15)$$

Taking logs:

$$\log p(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - \mathbf{w}^\top \mathbf{x}_n)^2 \quad (16)$$

Setting $\eta_n = \mathbf{w}^\top \mathbf{x}_n$, we can write in GLM form as:

$$\log p(y_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \frac{y_n \eta_n - \eta_n^2/2}{\sigma^2} - \frac{1}{2} \left(\frac{y_n^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \quad (17)$$

Linear regression (2/2)

If we set:

$$A(\eta_n) = \eta_n^2/2 \quad (18)$$

$$h(y_n, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}y_n^2\right) \quad (19)$$

then we can write:

$$\log p(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \frac{y_n\eta_n - A(\eta_n)}{\sigma^2} + \log h(y_n, \sigma^2) \quad (20)$$

And thus, the cumulants are given by:

$$\mathbb{E}(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = A'(\eta_n) = \eta_n = \mathbf{w}^\top \mathbf{x}_n \quad (21)$$

$$\text{Var}(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = A''(\eta_n)\sigma^2 = \sigma^2 \quad (22)$$

GLM components

A GLM can be considered as consisting of three parts:

- **Random component:** this is the probability distribution of the response variable
- **Systematic component:** specifies the explanatory variables within the linear combination of their coefficients ($\mathbf{X}\mathbf{w}$)
- **Link function $g(\mu)$:** defines the relationship between the random and systematic components:
 - Simple linear regression (**identity** link function):

$$g(\mu_n) = g(\mathbb{E}(y_n)) = \mathbf{x}_n^\top \mathbf{w} \quad (23)$$

- Binary logistic regression (**logit** link function):

$$g(\mu_n) = g(p(\mathbf{x}_n)) = \text{logit}(p(\mathbf{x}_n)) = \ln \left(\frac{p(\mathbf{x}_n)}{1 - p(\mathbf{x}_n)} \right) = \mathbf{x}_n^\top \mathbf{w} \quad (24)$$

Assumptions of GLM

- The observations of the response variable \mathbf{y} are i.i.d.
- Response variable y_n is typically exponentially distributed (not restricted to being normally distributed)
 - Implies that errors need not be normally distributed (but should be independent)
- Link function is linear with respect to the coefficients (w_d)
 - Relationship between response and explanatory variables does not have to be linear
 - Explanatory variables can be nonlinear transformations of original values (as in simple linear regression)
- Variance may not homogeneous (i.e. homoscedasticity is not a requirement)
- Parameters are estimated via MLE

Commonly used GLM models and their components

Model	Random component	Link function
Linear regression	Gaussian	Identity: $g(\mu_n) = \mu_n = \mathbf{w}^\top \mathbf{x}_n$
Binary logistic regression	Bernoulli	Logit: $g(\mu_n) = \log\left(\frac{\mu_n}{1-\mu_n}\right)$
Probit regression	Bernoulli	Probit: $g(\mu_n) = \Phi^{-1}(\mu_n)$
Multinomial logit/logistic	Categorical	Multinomial logit: $g(\mu_{nc}) = \log\left(\frac{\mu_{nc}}{\mu_{nC}}\right)$
Poisson regression	Poisson	Log: $g(\mu_n) = \log(\mu_n)$

Note that in all cases, the link function always results in:

$$g(\mu_n) = \mathbf{w}^\top \mathbf{x}_n \quad (25)$$

Its job is to “link” the response to the systematic component via a suitable transformation that results in a linear function of the \mathbf{w} ’s.

MLE of GLM parameters

The negative log-likelihood (ignoring constant terms) is given by

$$\text{NLL}(\mathbf{w}) = -\log p(\mathcal{D}|\mathbf{w}) = -\sum_{n=1}^N \log p(y_n|\mathbf{x}_n, \mathbf{w}) = \sum_{n=1}^N \frac{A(\eta_n)}{\sigma^2} - \frac{y_n \eta_n}{\sigma^2} \quad (26)$$

If we set $\ell_n = \eta_n y_n - A(\eta_n)$, then the NLL can be written as:

$$\text{NLL}(\mathbf{w}) = -\sum_{n=1}^N \frac{\ell_n}{\sigma^2} \quad (27)$$

where $\eta_n = \mathbf{w}^\top \mathbf{x}_n$.

The gradient of the NLL (for a single term) is then given by:

$$\mathbf{g}_n = \frac{y_n - \mu_n}{\sigma^2} \mathbf{x}_n \quad (28)$$

where $\mu_n = A'(\eta_n) = \ell^{-1}(\eta_n)$ is the mean function.