MAT8034: Machine Learning

Generalized Linear Models

Fang Kong

https://fangkongx.github.io/Teaching/MAT8034/Spring2025/index.html

Part of slide credit: Stanford CS229

Outline

- The exponential family
  - Motivation/Intuition
  - Examples
- Generalized linear models (GLMs)
  - Design ideas
  - Workflow

The exponential family

Motivation

- In the regression problem $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$
- In the classification problem $y|x;\theta \sim \text{Bernoulli}(\phi)$

- Whether these distributions can be uniformly represented?

- If P has a a special form, then inference and learning come for free

The exponential family

- $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$

- y: data label (scalar)
- η: natural parameter
- $T(y)$: sufficient statistic
- $b(y)$: base measure, depend on y, but not η (scalar)
- $a(\eta)$: log partition function (scalar)

$$1 = \sum_y P(y;\eta) = e^{-a(\eta)} \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

$$\Longrightarrow \quad a(\eta) = \log \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

Example 1: Bernoulli distribution

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

- Bernoulli(ϕ)

- $p(y=1;\phi)=\phi;\ p(y=0;\phi)=1-\phi$

- $$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

$$\eta = \log(\phi/(1-\phi))$$

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^\eta) \\ b(y) &= 1 \end{aligned}$$

Example 2: Gaussian distribution with $\sigma^2 = 1$

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

- Gaussian(μ, 1)

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

Thus, we see that the Gaussian is in the exponential family, with

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2). \end{aligned}$$

An observation

- Notice that for a Gaussian with mean μ we had

$$\eta = \mu, T(y) = y, a(\eta) = \frac{1}{2}\eta^2.$$

- We observe something peculiar:

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial_\eta^2 a(\eta) = 1 = \sigma^2 = \mathrm{var}(y)$$

- That is, derivatives of the log partition function is the expectation and variance. Same for Bernoulli.

Is this true in general?

Log Partition Function

- Yes! Recall that

$$a(\eta) = \log \sum_y b(y) \exp\left\{\eta^T T(y)\right\}$$

- Then, taking derivatives

$$\nabla_\eta a(\eta) = \frac{\sum_y T(y) b(y) \exp\{\eta^T T(y)\}}{\sum_y b(y) \exp\{\eta^T T(y)\}} = \mathbb{E}[T(y); \eta]$$

- Note: $\nabla^2_\eta a(\eta) = \mathrm{var}[T(y); \eta]$, you can check!

- Takeaway: In this way, once we're in the exponential family, we get inference "for free" meaning in the same way for every member

Quiz: Gaussian distribution with σ^2

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

- Gaussian(μ, σ^2) ?

Some Facts About Exponential Models

- There are many canonical exponential family models:
  - Binary $\mapsto$ Bernoulli
  - Multiple Classses $\mapsto$ Multinomial
  - Real $\mapsto$ Gaussian
  - Counts $\mapsto$ Poisson
  - $\mathbb{R}_+ \mapsto$ Gamma, Exponential
  - Distributions $\mapsto$ Dirichlet
- In this course, we'll use $T(y) = y$.

The GLMs

Three assumptions/design choices

1. $y \mid x; \theta \sim \text{ExponentialFamily}(\eta)$. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.

2. Given x, our goal is to predict the expected value of $T(y)$ given x. In most of our examples, we will have $T(y) = y$, so this means we would like the prediction $h(x)$ output by our learned hypothesis h to satisfy $h(x) = \mathrm{E}[y|x]$. (Note that this assumption is satisfied in the choices for $h_\theta(x)$ for both logistic regression and linear regression. For instance, in logistic regression, we had $h_\theta(x) = p(y = 1|x;\theta) = 0 \cdot p(y = 0|x;\theta) + 1 \cdot p(y = 1|x;\theta) = \mathrm{E}[y|x;\theta]$.)

3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$. (Or, if η is vector-valued, then $\eta_i = \theta_i^T x$.) Design choice

How linear regression belongs to GLMs?

- Consider the label $y \sim N(\mu, \sigma^2)$

$h_\theta(x)$	$=$	$E[y \vert x;\theta]$	Assumption 2
	$=$	μ	Gaussian distribution
	$=$	η	Assumption 1
	$=$	$\theta^T x.$	Assumption 3

How logistic regression belongs to GLMs?

- Consider the label $y \sim \text{Bernoulli}(\phi)$

$$
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] && \text{Assumption 2} \\
&= \phi && \text{Bernoulli distribution} \\
&= 1/(1+e^{-\eta}) && \text{Assumption 1} \\
&= 1/(1+e^{-\theta^T x}) && \text{Assumption 3}
\end{aligned}
$$

- Another reason for the definition of logistic regression

Workflow of GLMs

- Model formulation

$$\begin{array}{ccccc} \underline{\text{Model Parameter}} & & \underline{\text{Natural Parameter}} & & \underline{\text{Canonical}} \\ & & & & \phi : \text{Bernoulli} \\ \theta & \xmapsto{\theta^T x} & \eta & \xmapsto{g} & \mu : \text{Gaussian} \\ & & & & \lambda : \text{Poisson} \end{array}$$

- Maximum log-likelihood $\max_{\theta} \log p(y \mid x; \theta)$

- Gradient ascent to optimize $\theta^{(t+1)} = \theta^{(t)} + \alpha \left(y^{(i)} - h_{\theta^{(t)}}(x^{(i)}) \right) x^{(i)}$

Summary

- The exponential family
  - Motivation/Intuition
  - Examples
- Generalized linear models (GLMs)
  - Design ideas
  - Workflow