Learning From Data
Lecture 3: Generalized Linear Models

Yang Li yangli@sz.tsinghua.edu.cn

October 1, 2022

Ask me a question

What is the difference between probabilistic and non-probabilistic methods?

Today's Lecture

Supervised Learning (Part III)

- Review on linear and logistic regression
- Softmax Regression
- Review: exponential families
- Generalized linear models (GLM)

Written Assignment (WA1) is released. Due on Oct 8th. (Start early!)

Review of Lecture 2

Review of Lecture 2: Linear least square

- Hypothesis function for input feature $x^{(i)} \in \mathbb{R}^n$:

$$h_\theta(x^{(i)}) = \theta^T x^{(i)}, \text{ where } \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, \; x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

- Cost function for m training examples $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)} \right)^2$$

Also known as **ordinary least square regression** model.

How to minimize $J(\theta)$?

- Gradient descent:

$$\text{update rule (batch)} \quad \theta_j \leftarrow \theta_j + \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

$$\text{update rule (stochastic)} \qquad \theta_j \leftarrow \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

- Newton's method

$$\theta \leftarrow \theta - H^{-1} \nabla J(\theta)$$

- Normal equation

$$X^T X \theta = X^T y$$

Review of Lecture 2

Maximum likelihood estimation

- Log-likelihood function:

$$\ell(\theta) = \log \left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \right) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta)$$

  where p is a probability density function.

$$\theta_{MLE} = \operatorname*{argmax}_{\theta} \ell(\theta)$$

(True or False?) Ordinary least square regression is equivalent to the maximum likelihood estimation of θ.

True under the assumptions:

- $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$
- $\epsilon^{(i)}$ are i.i.d. according to $\mathcal{N}(0, \sigma^2)$

Review of Lecture 2: Logistic regression

- Hypothesis function:

$$h_\theta(x) = g(\theta^T x),\ g(z) = \frac{1}{1+e^{-z}} \text{ is the sigmoid function.}$$

- Assuming $y|x;\theta$ is distributed according to Bernoulli($h_\theta(x)$)

$$p(y|x;\theta) = h_\theta(x)^y \left(1 - h_\theta(x)\right)^{1-y}$$

- Log-likelihood function for m training examples:

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

Review of Lecture 2: Multi-Class Classification

Approach 1: Turn multi-class classification to a binary classification problem.

One-Vs-Rest

Learn k classifiers $h_1, \ldots, h_k$. Each h_i classify one class against the rest of the classes.
Given a new data sample x, its predicted label $\hat{y}$:

$$\hat{y} = \operatorname*{argmax}_i h_i(x)$$

Drawbacks of One-Vs-Rest:

- Class imbalance: more negative samples than positive samples when k is large

Approach 2: Multinomial classifier (one model for all classes)

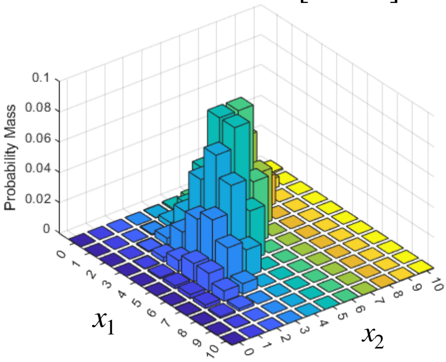
Softmax Regression

Review: Multinomial Distribution

Models the probability of counts for each side of a k-sided die rolled m times, each side with independent probability ϕ_i

$$\phi_1 + \cdots + \phi_k = 1$$

$$k = 3, n = 10 \qquad \phi = \left[\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right]$$



Extend logistic regression: Softmax Regression

Assume $p(y|x)$ is **multinomial distributed**, $k = |\mathcal{Y}|$

Hypothesis function for sample x:

$$h_\theta(x) = \begin{bmatrix} p(y=1|x;\theta) \\ \vdots \\ p(y=k|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x_j}} \begin{bmatrix} e^{\theta_1^T x} \\ \vdots \\ e^{\theta_k^T x} \end{bmatrix} = \text{softmax}(\theta^T x)$$

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{(z_j)}}$$

Parameters: $\theta = \begin{bmatrix} - & \theta_1^T & - \\ & \vdots & \\ - & \theta_k^T & - \end{bmatrix}$

Softmax Regression

Given $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$, the log-likelihood of the Softmax model is

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\
&= \sum_{i=1}^{m} \log \prod_{l=1}^{k} p(y^{(i)} = l|x^{(i)})^{\mathbf{1}\{y^{(i)}=l\}} \\
&= \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\{y^{(i)} = l\} \log p(y^{(i)} = l|x^{(i)}) \\
&= \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\{y^{(i)} = l\} \log \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}
\end{aligned}$$

Softmax Regression

Derive the stochastic gradient descent update:

- Find $\nabla_{\theta_l} \ell(\theta)$

$$\nabla_{\theta_l} \ell(\theta) = \sum_{i=1}^{m} \left[\left(\mathbf{1}\{y^{(i)} = l\} - P\left(y^{(i)} = l | x^{(i)}; \theta\right)\right) x^{(i)} \right]$$

Property of Softmax Regression

- Parameters $\theta_1, \ldots \theta_k$ are not independent: $\sum_j p(y=j|x) = \sum_j \phi_j = 1$
- Knowning $k-1$ parameters completely determines model.

Invariant to parameter shift

$$p(y|x;\theta) = p(y|x;\theta - \psi)$$

Proof.

Relationship with Logistic Regression

When K = 2,

$$h_\theta(x) = \frac{1}{e^{\theta_1^T x} + e^{\theta_2^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \end{bmatrix}$$

Replace $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$ with $\theta* = \theta - \begin{bmatrix} \theta_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \theta_1 - \theta_2 \\ 0 \end{bmatrix}$,

$$h_\theta(x) = \frac{1}{e^{\theta_1^T x - \theta_2^T x} + e^{0x}} \begin{bmatrix} e^{(\theta_1 - \theta_2)^T x} \\ e^{0^T x} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{e^{(\theta_1-\theta_2)^T x}}{1+e^{(\theta_1-\theta_2)^T x}} \\ \frac{1}{1+e^{(\theta_1-\theta_2)^T x}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{1+e^{-(\theta_1-\theta_2)^T x}} \\ 1 - \frac{1}{1+e^{-(\theta_1-\theta_2)^T x}} \end{bmatrix} = \begin{bmatrix} g(\theta *^T x) \\ 1 - g(\theta *^T x) \end{bmatrix}$$

When to use Softmax?

- When classes are mutually exclusive: use Softmax
- Not mutually exclusive (a.k.a. **multi-label classification**): multiple binary classifiers may be better

Summary: Linear models

What we've learned so far:

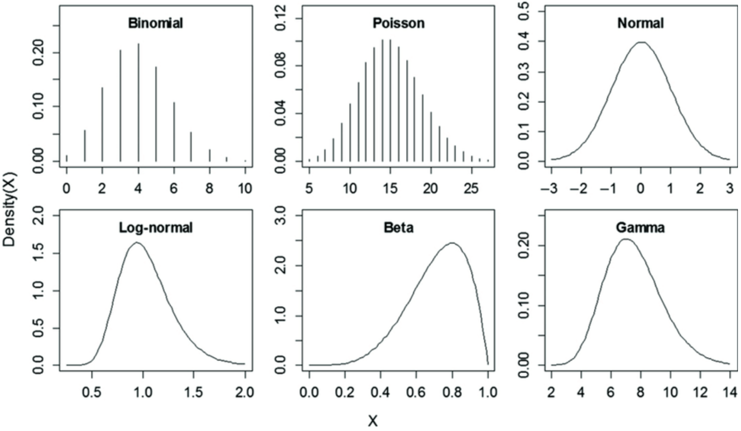
Learning task	Model	$p(y\|x;\theta)$
regression	Linear regression	$\mathcal{N}(h_\theta(x)\ ,\sigma^2)$
binary classification	Logistic regression	Bernoulli($h_\theta(x)$)
multi-class classification	Softmax regression	Multinomial($[h_\theta(x)]$)

Can we generalize the linear model to other distributions?

Generalized Linear Model (GLM): a recipe for constructing linear models in which $y|x;\theta$ is from an **exponential family**.

Review: Exponential Family

Exponential Family of Distributions



Examples of distribution classes in the exponential family.

Exponential Family of Distributions

A class of distributions is in the **exponential family** if its density can be written in the *canonical form*:

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$

- y: random variable
- η : natural/canonical parameter (that depends on distribution parameter(s))
- $T(y)$: sufficient statistic of the distribution
- $b(y)$: a function of y
- $a(\eta)$: log partition function (or "cumulant function")

Exponential Family

Log partition function $a(\eta)$ is the log of a normalizing constant. i.e.

$$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)} = \frac{b(y)e^{\eta^T T(y)}}{e^{a(\eta)}}$$

Function $a(\eta)$ is chosen such that $\sum_y p(y;\eta) = 1$ (or $\int_y p(y;\eta)dy = 1$).

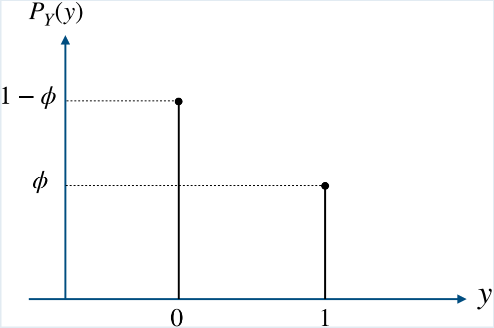
$$a(\eta) = \log\left(\sum_y b(y)e^{\eta^T T(y)}\right)$$

Exponential Family Examples

Bernoulli Distribution

Bernoulli(ϕ): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$



Bernoulli Distribution

Bernoulli(ϕ): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

How to write it in the form of $p(y; \eta) = b(y) e^{\eta^T T(y) - a(\eta)}$?

Exponential Family Examples

Bernoulli Distribution

Bernoulli(ϕ): a distribution over $y \in \{0, 1\}$, such that

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

- $\eta = \log\left(\frac{\phi}{1-\phi}\right)$
- $b(y) = 1$
- $T(y) = y$
- $a(\eta) = \log(1 + e^\eta)$

Exponential Family Examples

Gaussian Distribution (unit variance)

Probability density of a Gaussian distribution $\mathcal{N}(\mu, 1)$ over $y \in \mathbb{R}$:

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

- $\eta = \mu$
- $b(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)$
- $T(y) = y$
- $a(\eta) = \frac{1}{2}\eta^2$

Exponential Family Examples

Two parameter example:

Gaussian Distribution

Probability density of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ over $y \in \mathbb{R}$:

$$p(y; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

- $\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$
- $b(y) = \frac{1}{\sqrt{2\pi}}$
- $T(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$
- $a(\eta) = \frac{\mu^2}{2\sigma^2} + \log \sigma$

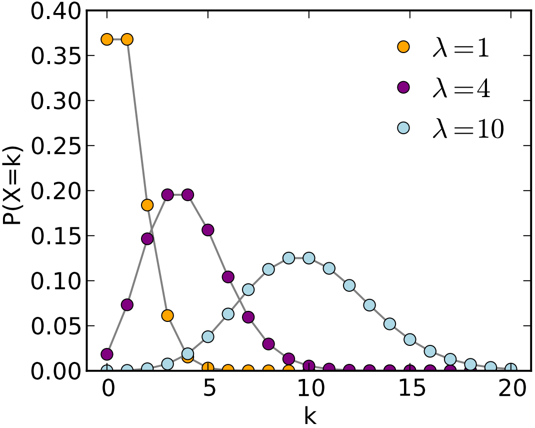
Exponential Family Examples

Poisson distribution: Poisson(λ)

Models the probability that an event occurring $y \in \mathbb{N}$ times in a fixed interval of time, *assuming events occur independently at a constant rate*

Probability density function of Poisson(λ) over $y \in \mathcal{Y}$:

$$p(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$



Exponential Family Examples

Poisson distribution Poisson(λ)

Probability density function of Poisson(λ) over $y \in \mathcal{Y}$:

$$p(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

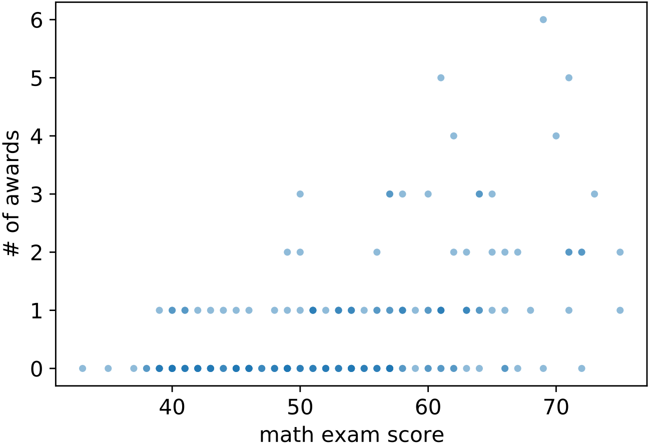
- $\eta = \log \lambda$
- $b(y) = \frac{1}{y!}$
- $T(y) = y$
- $a(\eta) = e^\eta$

Generalized Linear Models

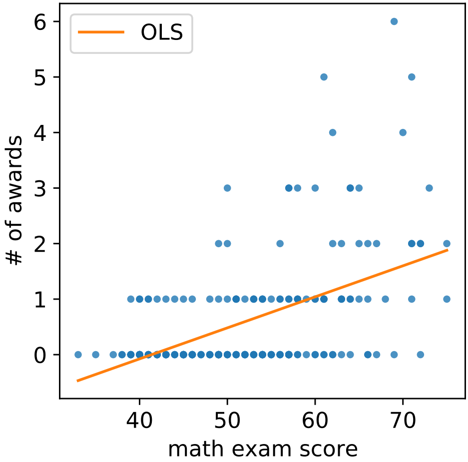
Generalized Linear Models: Intuition

Example 1: Award Prediction

Predict y, **the number of school awards** a student gets given x, the math exam score.



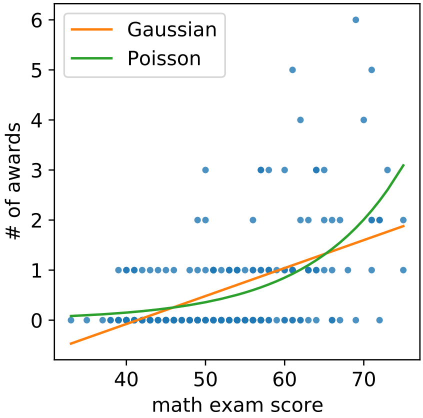
Generalized Linear Models: Intuition



Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution.
- Assumes change in x is proportional to change in y

Generalized Linear Models: Intuition



Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution. ***Poisson** distribution is better for modeling occurrences*
- Assumes change in x is proportional to change in y *More realistic to be proportional to the **rate** of increase in y* (e.g. doubling or halving y)

Generalized Linear Models : Intuition

Generalized Linear Model (GLM): a recipe for constructing linear models in which $y|x;\theta$ is from an exponential family.

Design motivation of GLM

- We can select a distribution for **Response variables** y
- Allow (the **canonical link function** of y) to vary linearly with the input values x

e.g. $log(\lambda) = \theta^T x$

Nelder, John Ashworth, and Robert William Maclagan Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) 135 (3): 37084.

Generalized Linear Models: Construction

Formal GLM assumptions & design decisions:

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\mathbb{E}[T(y)|x]$
   e.g. When $T(y) = y$, $h(x) = \mathbb{E}[y|x]$
3. The natural parameter η and the inputs x are related linearly:

   η is a number:

$$\eta = \theta^T x$$

   η is a vector:

$$\eta_i = \theta_i^T x \quad \forall i = 1, \ldots, n \quad \text{or} \quad \eta = \Theta^T x$$

Generalized Linear Models: Construction

Relate natural parameter η to distribution mean $\mathbb{E}[T(y); \eta]$:

- **Canonical response function** g gives the mean of the distribution

$$g(\eta) = \mathbb{E}[T(y); \eta]$$

  a.k.a. the "mean function"

- g^{-1} is called the **canonical link function**

$$\eta = g^{-1}(\mathbb{E}[T(y); \eta])$$

GLM example: ordinary least square

Apply GLM construction rules:

1. Let $y|x;\theta \sim N(\mu, 1)$

$$\eta = \mu, \; T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \mu = \eta \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

$$h_\theta(x) = \eta = \theta^T x$$

Canonical response function: $\mu = g(\eta) = \eta$ (identity)
Canonical link function: $\eta = g^{-1}(\mu) = \mu$ (identity)

GLM example: logistic regression

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Bernoulli}(\phi)$

$$\eta = \log\left(\frac{\phi}{1-\phi}\right),\ T(y) = y$$

2. Derive hypothesis function:

$$\begin{aligned} h_\theta(x) &= \mathbb{E}\left[T(y)|x;\theta\right] \\ &= \mathbb{E}\left[y|x;\theta\right] \\ &= \phi = \frac{1}{1+e^{-\eta}} \end{aligned}$$

3. Adopt linear model $\eta = \theta^T x$:

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

Canonical response function: $\phi = g(\eta) = \text{sigmoid}(\eta)$
Canonical link function : $\eta = g^{-1}(\phi) = \text{logit}(\phi)$

GLM example: Poisson regression

Example 1: Award Prediction

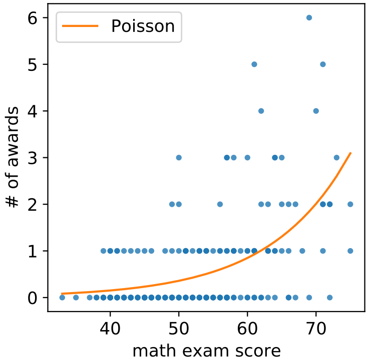
Predict y, **the number of school awards** a student gets given x, the math exam score.

Use GLM to find the hypothesis function...

GLM example: Poisson regression

Apply GLM construction rules:

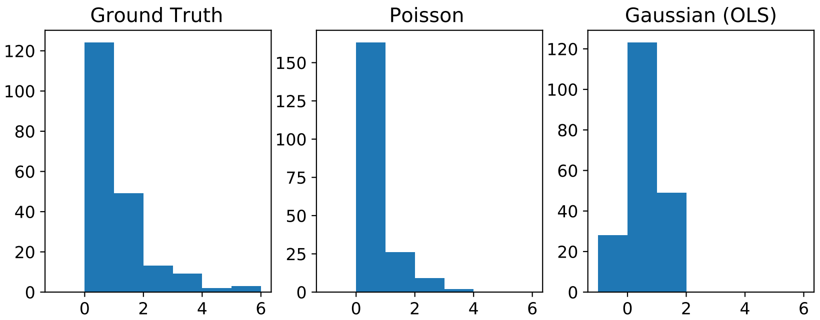
1. Let $y|x;\theta \sim \text{Poisson}(\lambda)$
$$\eta = \log(\lambda),\ T(y) = y$$
2. Derive hypothesis function:
$$h_\theta(x) = \mathbb{E}\left[y|x;\theta\right] = \lambda = e^\eta$$
3. Adopt linear model $\eta = \theta^T x$:
$$h_\theta(x) = e^{\theta^T x}$$



Canonical response function: $\lambda = g(\eta) = e^\eta$
Canonical link function : $\eta = g^{-1}(\lambda) = \log(\lambda)$

GLM example: Poisson regression

Distribution of the predicted number of awards (y)



Poisson regression successfully captures the long tail of $P(y)$

GLM example: Softmax regression

Probability mass function of a Multinomial distribution over k outcomes

$$p(y;\phi) = \prod_{i=1}^{k} \phi_i^{\mathbf{1}\{y=i\}}$$

Derive the exponential family form of Multinomial$(\phi_1, .., \phi_k)$: Note: $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$ is not a parameter

- $T(y) = \begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix}$

  $T(y)_i = \mathbf{1}\{y=i\} = \begin{cases} 0 & y \neq i \\ 1 & y = i \end{cases}$
- $a(\eta) = -\log(\phi_k)$
- $\eta = \begin{bmatrix} \log\left(\frac{\phi_1}{\phi_k}\right) \\ \vdots \\ \log\left(\frac{\phi_{k-1}}{\phi_k}\right) \end{bmatrix}$
- $b(y) = 1$

GLM example: Softmax regression

Apply GLM construction rules:

1. Let $y|x;\theta \sim \text{Multinomial}(\phi_1, \ldots, \phi_k)$, for all $i = 1 \ldots k-1$

$$\eta_i = \log\left(\frac{\phi_i}{\phi_k}\right), \; T(y) = \begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix}$$

   Compute inverse: $\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$ ← *canonical response function*

2. Derive hypothesis function:

$$h_\theta(x) = \mathbb{E}\left[\begin{bmatrix} \mathbf{1}\{y=1\} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{bmatrix} \middle| x;\theta\right] = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{k-1} \end{bmatrix}$$

$$\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

GLM example: Softmax regression

3. Adopt linear model $\eta_i = \theta_i^T x$:

$$\phi_i = \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \text{ for all } i = 1 \dots k-1$$

$$h_\theta(x) = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ \vdots \\ e^{\theta_{k-1}^T x} \end{bmatrix}$$

Canonical response function: $\phi_i = g(\eta) = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$

Canonical link function : $\eta_i = g^{-1}(\phi_i) = \log\left(\frac{\phi_i}{\phi_k}\right)$

GLM Summary

Sufficient statistic $T(y)$
Response function $g(\eta)$
Link function $g^{-1}(\mathbb{E}[T(y);\eta])$

Exponential Family	$\mathcal{Y}$	$T(y)$	$g(\eta)$	$g^{-1}(\mathbb{E}[T(y);\eta])$
$\mathcal{N}(\mu, 1)$	$\mathbb{R}$	y	η	μ
Bernoulli(ϕ)	$\{0, 1\}$	y	$\frac{1}{1+e^{-\eta}}$	$\log\frac{\phi}{1-\phi}$
Poisson(λ)	$\mathbb{N}$	y	e^{η}	$\log(\lambda)$
Multinomial($\phi_1, \ldots, \phi_k$)	$\{1, \ldots, k\}$	$\mathbf{1}\{y = i\}$	$\frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$	$\eta_i = \log\left(\frac{\phi_i}{\phi_k}\right)$

GLM is effective for modelling different types of distributions over y