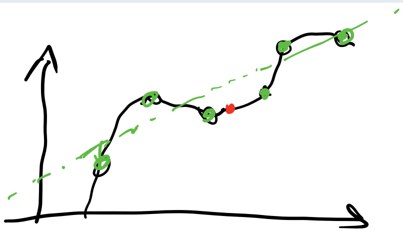
Learning From Data
Lecture 3: Generalized Linear Models

Yang Li yangli@sz.tsinghua.edu.cn

October 8, 2021

Ask me a question (1/2)

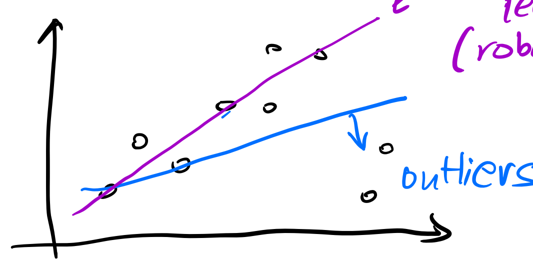
Can linear regression overfit ill-posed data?



ill-posed

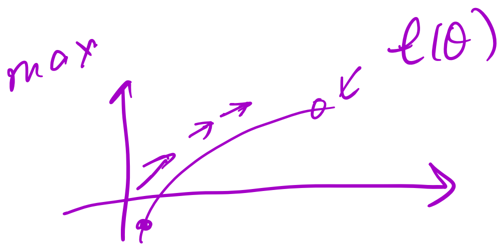
overfitting

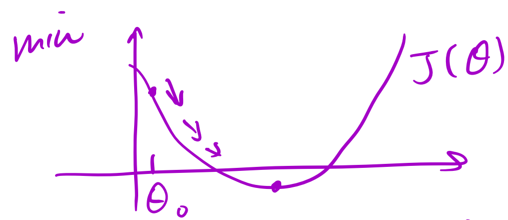
regularized least square (robust)



Ask me a question (2/2)

Why is the gradient update in logistic regression having "+" sign ?





$$\theta := \theta + \alpha \nabla \ell(\theta).$$

gradient ascent

$$\theta := \theta - \alpha \nabla J(\theta)$$

gradient descent.

Today's Lecture

Supervised Learning (Part III)

- Review on linear and logistic regression
- Multi-class classification
- Review: exponential families
- Generalized linear models (GLM)

Written Assignment (WA1) is released. Due on Oct 22nd. (Start early!)

Review of Lecture 2

Review of Lecture 2: Linear least square

$w^T x + b$

intercept / bias

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

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

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 $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$:

OLS.

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

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 $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$:

$$J(\theta) =$$

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

Review of Lecture 2: Linear least square

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

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

- 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:

  update rule (batch)

  update rule (stochastic)

- Newton's method

- Normal equation

How to minimize $J(\theta)$?

- Gradient descent:

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

  update rule (stochastic)

- Newton's method

- Normal equation

How to minimize $J(\theta)$?

- Gradient descent:

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

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

- Newton's method

- Normal equation

How to minimize $J(\theta)$?

- Gradient descent:

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

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

- Newton's method

$\nabla J(\theta) = 0$

$\theta \leftarrow \theta - \frac{f(\theta)}{f'(\theta)}$

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

$H(J(\theta))$

- 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.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

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.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

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

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.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta)$$

(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 $\mathcal{N}(0, \sigma^2)$

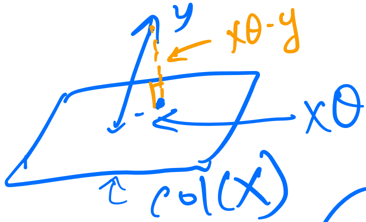
Review of Lecture 2: Linear Regression Exercise

Geometric view:

$X^TX\theta = X^Ty$

$\Rightarrow X^TX\theta - X^Ty = 0$

$X^T(\underbrace{X\theta - y}_{\text{residual}}) = 0$



Solution θ^* always exists since $C(X^T) \subseteq C(X^TX)$.
But the solution is not unique when X^TX is not full rank

The normal equation for solving ordinary least square is:

rank<m

$$\boxed{X^TX\theta = X^Ty}$$

$\theta_0 + \theta_1 x_1 + \theta_2 x_2$

$\rightarrow \theta_0 + \theta_1 x$

When X^TX is invertible, we have $\theta = (X^TX)^{-1}X^Ty$ Now, suppose X^TX is singular. Does the solution exist?

rank deficient

$$\boxed{X^TX} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

(n=3)

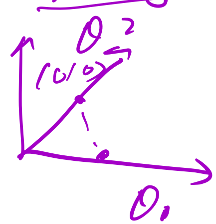
rank=2

$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ 2 training samples

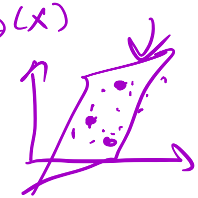
$\underline{Ax = y}$

$X^TX \quad \theta \quad X^Ty$

infinitely many solutions



$h_\theta(x)$



Review of Lecture 2: Logistic regression

- Hypothesis function:

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

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($h_\theta(x)$) $\phi = h_\theta(x)$

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

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$(h_\theta(x))$

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

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:

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

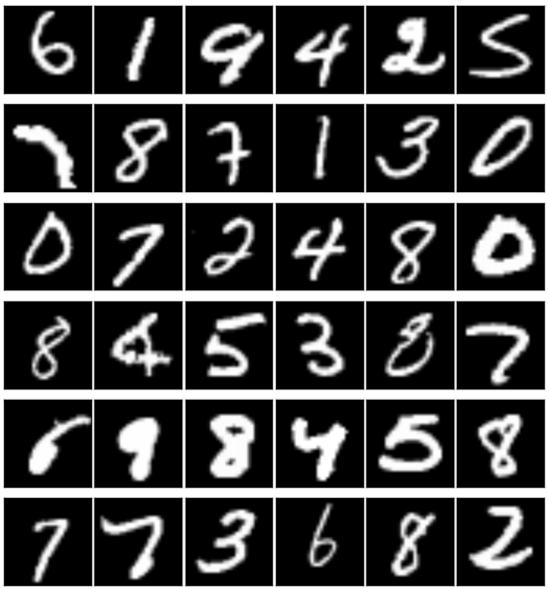
Multi-Class Classification

Multi-class classification

Each data sample belong to one of $k > 2$ different classes.

$$\underline{\mathcal{Y}} = \{1, \ldots, k\}$$

MNIST Samples



$k = 10.$

Given new sample $\underline{x \in \mathbb{R}^k}$, predict which class it belongs.

Naive Approach: Convert to binary classification

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 $\hat{y}$:

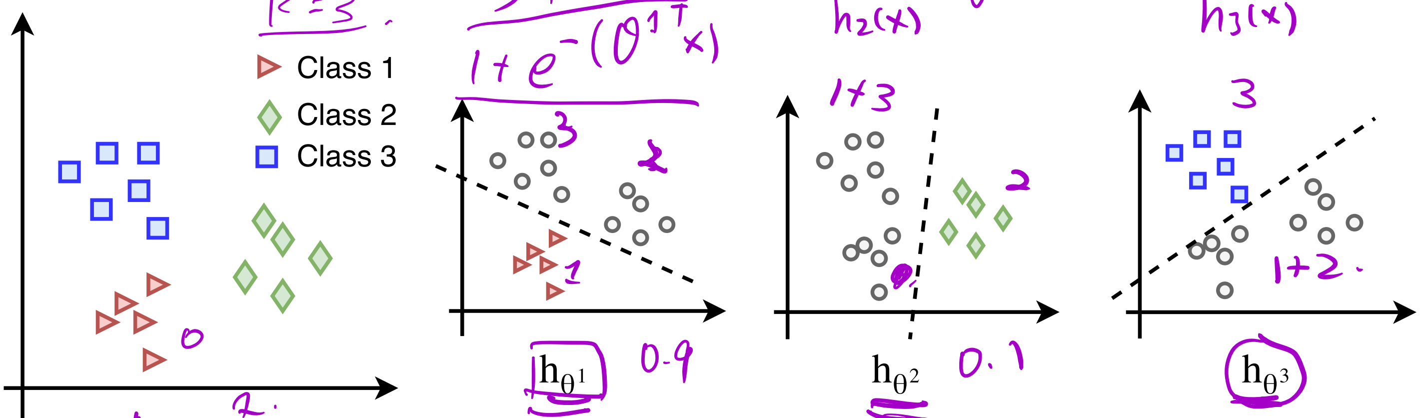
$x \rightarrow \{x^{(1)}, x^{(2)}, x^{(3)}\}$

$$\hat{y} = \underset{i}{\operatorname{argmax}}\, h_i(x)$$

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 2 \\ x_3 & 3 \end{bmatrix} \quad y$$

$k=3$

$$h_1(x) \rightarrow \frac{1}{1+e^{-(\theta^{1T} x)}}$$



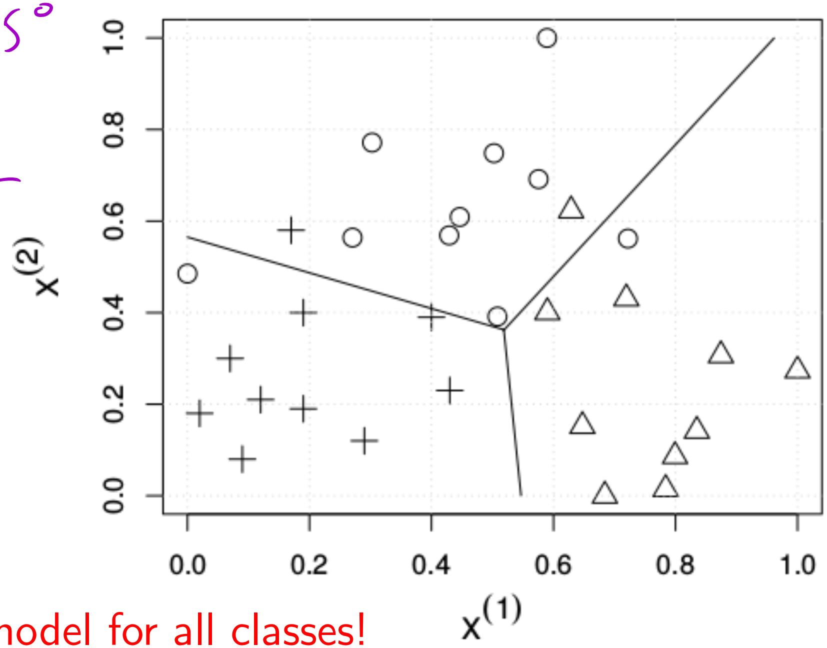
$h_\theta(x)$

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 0 \\ x_3 & 0 \end{bmatrix} \quad \theta^1 \qquad \begin{bmatrix} x_1 & 0 \\ x_2 & 1 \\ x_3 & 0 \end{bmatrix} \quad \theta^2 \qquad \begin{bmatrix} x_1 & 0 \\ x_2 & 0 \\ x_3 & 1 \end{bmatrix} \quad \theta^3$$

Drawbacks of One-Vs-Rest:

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

$k=20$, $m=100$. 5 samples/class $\Rightarrow$ for each binary classifier, ratio of positive to negative classes will be 5% : 95%.



Learn one model for all classes!

Review: Multinomial Distribution

$(y | x ; \theta)$

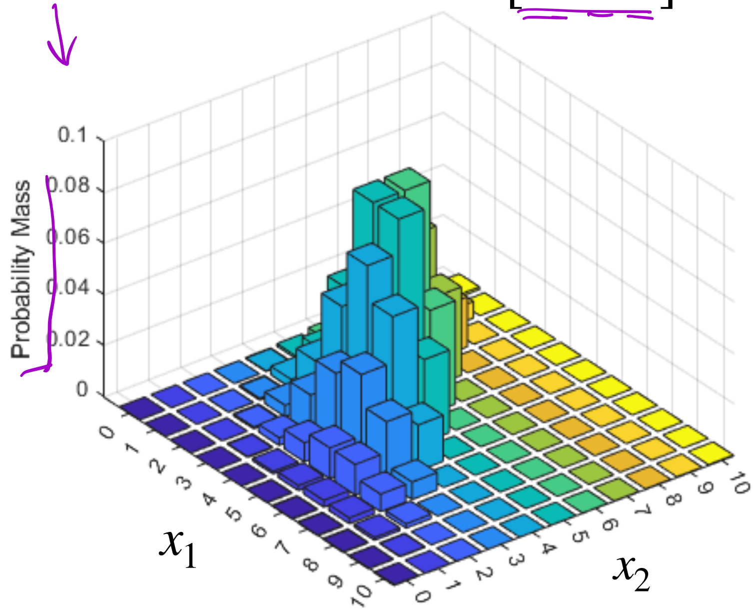
$\{y = 1, \ldots, k\}$

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

$$\phi_1 + \ldots + \phi_k = 1$$

$$\phi = \left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]$$

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



Extend logistic regression: Softmax Regression

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

Extend logistic regression: Softmax Regression

$\Theta = \begin{bmatrix} -\theta_1- \\ -\theta_2- \\ \vdots \\ -\theta_k- \end{bmatrix}$ $\quad x \to \boxed{h_\theta(x)} \to \hat{y}$ $\quad h_{\theta_i}(x) = \dfrac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}}$

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

Hypothesis function for sample x: $(\theta_1, \cdots, \theta_k)$

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

$h_{\theta_1}(x)$, $h_{\theta_j}(x)$, $h_{\theta_k}(x)$

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

Extend logistic regression: Softmax Regression

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

Hypothesis function for sample x:

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

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

$x \in \mathbb{R}^n$.

Parameters: $\theta = \underbrace{\begin{bmatrix} - & \theta_1^T & - \\ & \vdots & \\ - & \theta_k^T & - \end{bmatrix}}_{k \times n}$

Softmax Regression

$h_\theta(x)^{y}(1-h_\theta(x))^{1-y}$ $y \in \{0,1\}$

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

$\ell(\theta_1, \ldots, \theta_k)$

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

$$= \sum_{i=1}^{m} \log \prod_{l=1}^{k} p(y^{(i)} = l|x^{(i)})^{\mathbf{1}\{y^{(i)}=l\}}$$

$$1\{y^i = l\} = \begin{cases} 1 & y^i = l \\ 0 & y^i \neq l. \end{cases}$$

Multinomial(

$$p(y|x,\theta_1,\ldots,\theta_k) = \prod_{l=1}^{k} p(y^i = l|x^i)^{1\{y^i=l\}}$$

$$= \begin{cases} p(y^i = 1|x^i) & \text{if } y^i = 1. \\ p(y^i = 2|x^i) & \text{if } y^i = 2 \\ \vdots & \end{cases}$$

Softmax Regression

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

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

$$= \sum_{i=1}^{m} \log \prod_{l=1}^{k} p(y^{(i)} = l|x^{(i)})^{\mathbf{1}\{y^{(i)}=l\}}$$

$$= \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\{y^{(i)} = l\} \log p(y^{(i)} = l|x^{(i)})$$

13:15

Softmax Regression

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

$$\ell(\theta) \in \mathbb{R}$$

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta) \\
&= \sum_{i=1}^{m} \log \prod_{l=1}^{k} p(y^{(i)} = l | x^{(i)})^{\mathbf{1}\{y^{(i)} = l\}} \\
&= \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\{y^{(i)} = l\} \log \underbrace{p(y^{(i)} = l | x^{(i)})}_{\downarrow \text{softmax}} \\
&= \underbrace{\sum_{i=1}^{m}}_{m \text{ samples}} \underbrace{\sum_{l=1}^{k}}_{k\ \theta_j\text{'s}} \mathbf{1}\{y^{(i)} = l\} \log \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}
\end{aligned}
$$

Softmax Regression

$$\Theta = \begin{bmatrix} - \ \theta_1^T \ - \\ \vdots \\ \theta_l^T \\ \vdots \\ \theta_K^T \end{bmatrix}$$

Derive the stochastic gradient descent update:

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

not to be confused with the ℓ in the previous page.

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

$\theta_l \in \mathbb{R}^{n \times 1}$

try this at home.

Property of Softmax Regression

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

$$\begin{bmatrix} \theta_1 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} - \begin{bmatrix} \psi \\ \psi \\ \psi \\ \vdots \end{bmatrix} \quad \psi \mathbb{1}_n$$

Invariant to scalar addition

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

Proof. RHS, $p(y|x;\theta-\psi)$

$$p(y=l|x;\theta-\psi) = \frac{e^{(\theta_l-\psi)^T x}}{\sum_{j=1}^{k} e^{(\theta_j-\psi)^T x}} = \frac{e^{\theta_l^T x} \cdot \cancel{e^{-\psi^T x}}}{\sum_{j=1}^{k} e^{\theta_j^T x} \cdot e^{-\psi^T x}} = \frac{e^{\theta_l^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}}$$

$$= p(y=l|x;\theta)$$

Relationship with Logistic Regression

$$\frac{e^{\theta_i^T x}}{\sum_{j=1} e^{\theta_j^T x}} \quad i = 1, 2.$$

When K = 2,

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

Relationship with Logistic Regression

For any $\psi \in \mathbb{R}$, $p(y|x;\theta) = p(y|x;\theta-\psi)$

$$\left[\frac{100}{1}\ \frac{5}{2}\ \frac{5}{3}\right]$$

$[\ p(y)$

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

let $\psi = \theta_2$

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

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

$\theta^* = \theta_1 - \theta_2$

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

$\leftarrow$ logistic regression

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:

$\varepsilon \sim$

Learning task	Model	$p(y \mid x; \theta)$
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?

Summary: Linear models

What we've learned so far:

Learning task	Model	$p(y \mid x; \theta)$
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 $y \mid x; \theta$ is from an **exponential family**.

Review: Exponential Family

Exponential Family

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

R.V.

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

sufficient statistic

- 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.

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

$e^{-a(\eta)}$

Function $a(\eta)$ is chosen such that $\sum_y p(y;\eta) = 1$ y is discrete

(or $\int_y p(y;\eta)dy = 1$). y is cont.

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

$$\sum_y p(y;\eta)$$

$$= \sum_y \frac{b(y)e^{\eta^T T(y)}}{e^{a(\eta)}} = 1$$

$$\frac{1}{e^{a(\eta)}} \sum_y b(y)e^{\eta^T T(y)} = 1.$$

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

Exponential Family Examples

Bernoulli Distribution

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

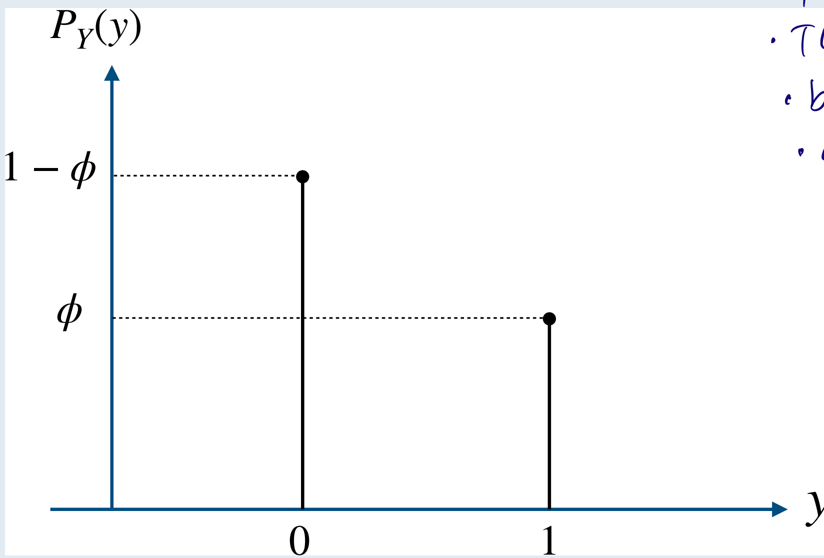
$$p(y; \phi) = \underline{\phi^y} \, \underline{(1 - \phi)^{1-y}}$$

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

- η
- $T(y)$
- $b(y)$
- $a(y)$

?

PMF.



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

y

natural parameter.

a function of ϕ.

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

$$= e^{y\log\phi + (1-y)\log(1-\phi)}$$

$$= e^{y\log\phi + \log(1-\phi) - y\log(1-\phi)}$$

$$= e^{y\log\frac{\phi}{1-\phi} + \log(1-\phi)}$$

$$= e^{\underbrace{y}_{T(y)}\underbrace{\log\frac{\phi}{1-\phi}}_{\eta} - \underbrace{(-\log(1-\phi))}_{a(\eta)}}$$

$b(y) = 1.$

$$\eta = \log\frac{\phi}{1-\phi}$$

$$e^\eta = \frac{\phi}{1-\phi}$$

$$e^\eta - e^\eta\phi = \phi$$

$$\phi = \frac{e^\eta}{1+e^\eta} = \underbrace{\frac{1}{1+e^{-\eta}}}_{\text{sigmoid}}$$

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

$$= -\log\left(1 - \frac{1}{1+e^{-\eta}}\right)$$

$$= -\log\left(\frac{1}{1+e^{\eta}}\right)$$

$$= \log(1+e^\eta)$$

Exponential Family Examples

Bernoulli Distribution

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

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

- $\eta =$
- $b(y) =$
- $T(y) =$
- $a(\eta) =$

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)

$y.\ \eta \leftrightarrow \mu.$

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

$$\underline{b(y) \cdot e^{\eta^T T(y) - a(\eta)}} \qquad p(y;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

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

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} e^{-\frac{1}{2}(\mu^2 - 2y\mu)}.$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}}_{b(y)} e^{\mu y - \frac{\mu^2}{2}} \qquad a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2},$$

$\eta = \mu$ $\qquad T(y) = y$

Exponential Family Examples

Gaussian Distribution (unit variance)

σ.

Probability density of a Gaussian distribution $\mathcal{N}(\mu, \underline{1})$ 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

η

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

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

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

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*

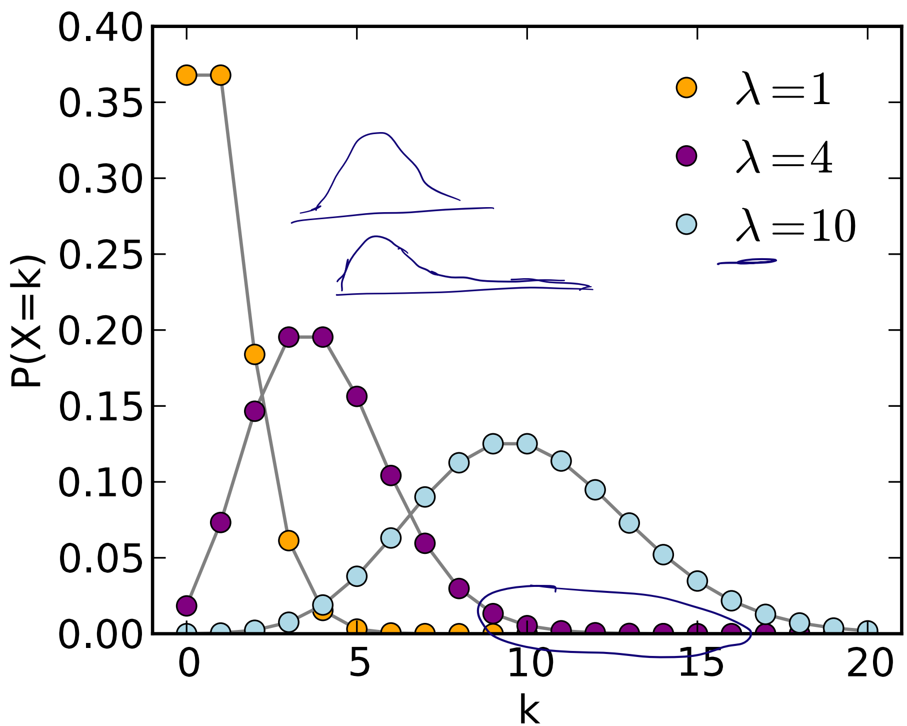
Exponential Family Examples

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

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

Poisson distribution Poisson(λ)

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

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

$\lambda = e^\eta$.

$\eta = \log \lambda$

$T(y) = y$

$b(y) = \frac{1}{y!}$

$a(\eta) = e^\eta$.

$\frac{1}{y!} e^{\log(\lambda^y e^{-\lambda})} = \frac{1}{y!} e^{y \underbrace{\log \lambda}_{\eta} + (-\lambda)}$

$T(y)$.

$a = \lambda = e^\eta$.

Exponential Family Examples

Poisson distribution Poisson(λ)

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

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

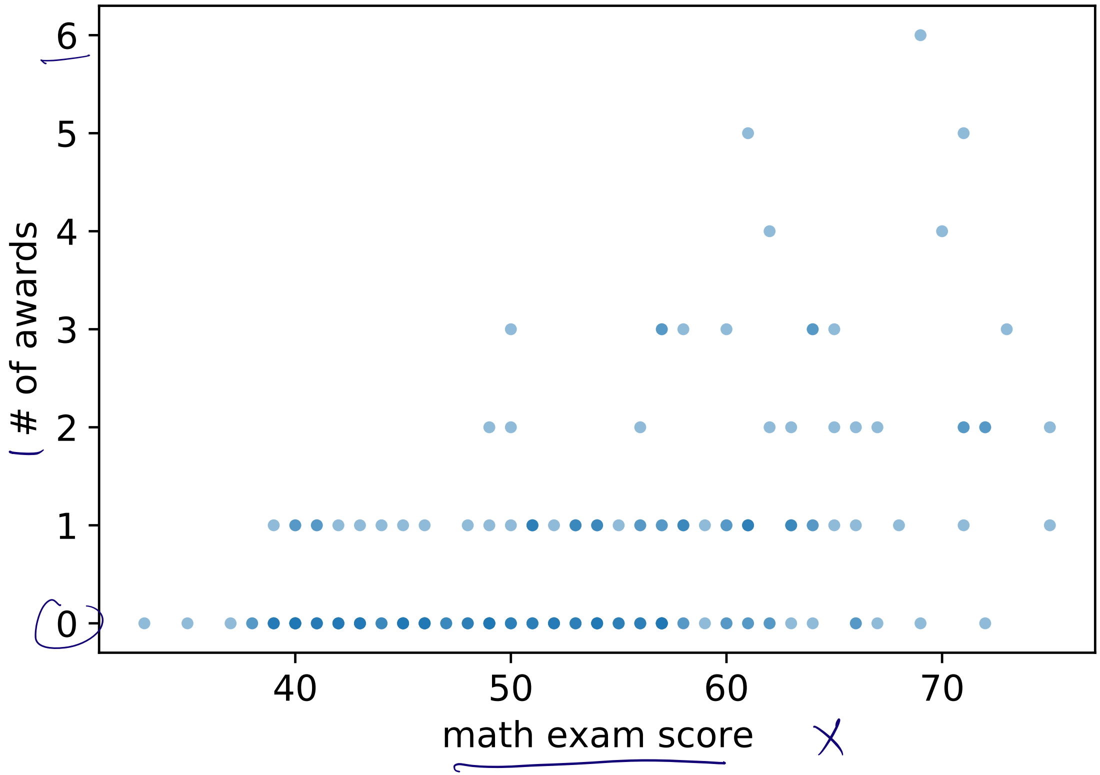
- $\eta = \log \lambda$
- $b(y) = \frac{1}{y!}$
- $T(y) = y$
- $a(\eta) = e^\eta$

Generalized Linear Models

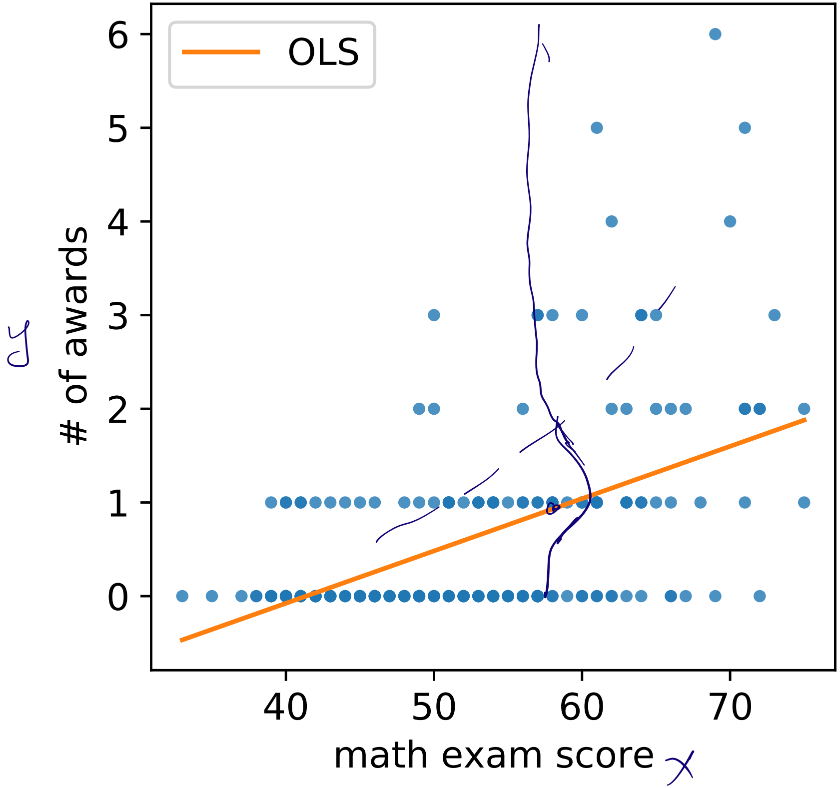
Generalized Linear Models: Intuition

Example 1: Award Prediction

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



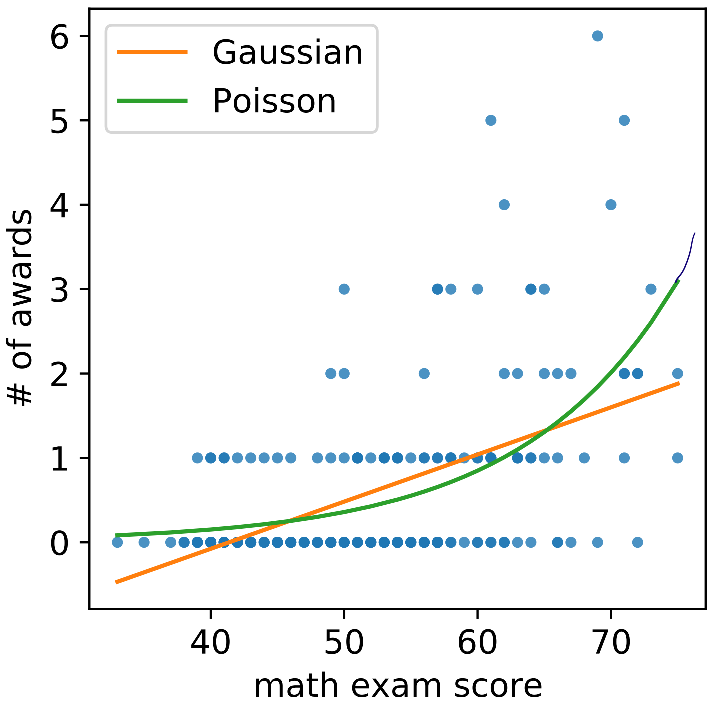
Generalized Linear Models: Intuition



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

- **Response variables** y can have arbitrary distributions
- Allow arbitrary function of y (the **link function**) to vary linearly with the input values x

$y = \theta^T x.$

$g(y) = \theta^T x$

Generalized Linear Models: Construction

If $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$

$\mathbb{E}[y|x]$

$\mathcal{N}(\mu, \sigma^2)$

Formal GLM assumptions & design decisions:

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$ ← natural parameters $\phi_1, \phi_2, \ldots, \phi_k$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\mathbb{E}[T(y)|x]$ Sufficient statistics.
   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:
   natural/canonical
   $$\eta = \theta^T x$$
   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

natural param

mean

μ, ϕ

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

$p(y|x)$

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"

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

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

  a.k.a. the "mean function"

- g^{-1} is called the **canonical <u>link function</u>**

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

GLM example: ordinary least square

Apply GLM construction rules:

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

$$\underbrace{\eta = \mu,}\ \underbrace{T(y) = y}$$

GLM example: ordinary least square

Apply GLM construction rules:

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

$$\boxed{\eta = \mu} \quad T(y) = y$$

$$g = \text{Identity}$$

$$\mu$$

$$\begin{tikzcd} \eta \arrow[r, bend right, "g^{-1}"'] & \mathbb{E}[T(y)|x] \arrow[l, bend right, "g"'] \end{tikzcd}$$

2. Derive hypothesis function:

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

$$\Rightarrow h_\theta(x) = \eta$$

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 $\underbrace{\eta = \underbrace{\theta^T x}}$:

$$\underbrace{h_\theta(x)} = \eta = \underbrace{\theta^T x}$$

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

$$\phi = \underline{\underline{g(\eta)}} = \frac{1}{1+e^{-\eta}}$$

response

Apply GLM construction rules:

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

g^{-1}

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

GLM example: logistic regression

Apply GLM construction rules:

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

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

$P(y) = \phi^{y}(1-\phi)^{1-y}$.

2. Derive hypothesis function:

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

sigmoid.

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:

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

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

$$h_\theta(x) = \frac{1}{1+e^{\underbrace{-\theta^T x}_{\eta}}} \Bigg\}\ \text{logistic function}$$

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: $\phi = g(\eta) = \text{sigmoid}(\eta)$

GLM example: logistic regression

Apply GLM construction rules:

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

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

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 : $\eta = g^{-1}(\phi) = \text{logit}(\phi)$ $\log \frac{\phi}{1-\phi}$

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 $y|x;\theta \sim \text{Poisson}(\lambda)$

$$\eta \overset{g^{-1}}{=} \log(\lambda),\ T(y) = y$$

2. Derive hypothesis function:

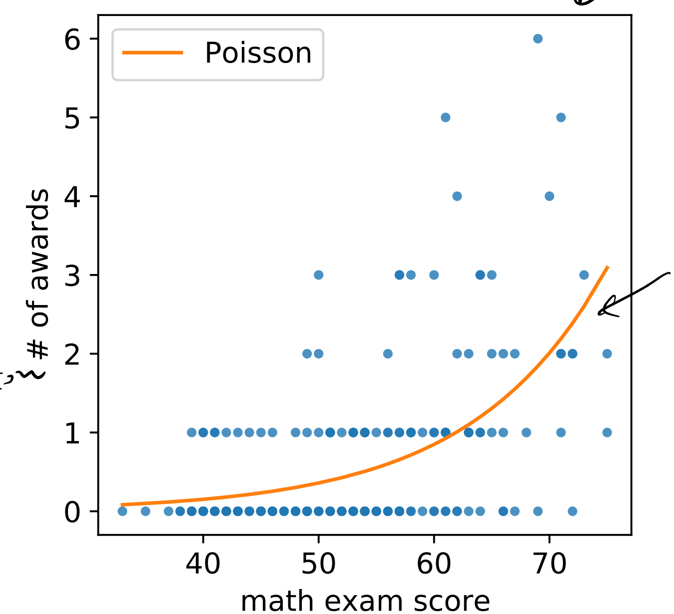
$\mathbb{E}[T(y)|x;\theta]$

$$h_\theta(x) = \mathbb{E}[y|x;\theta]$$
$$= \lambda = e^\eta$$

$g(\eta)$.
response function

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

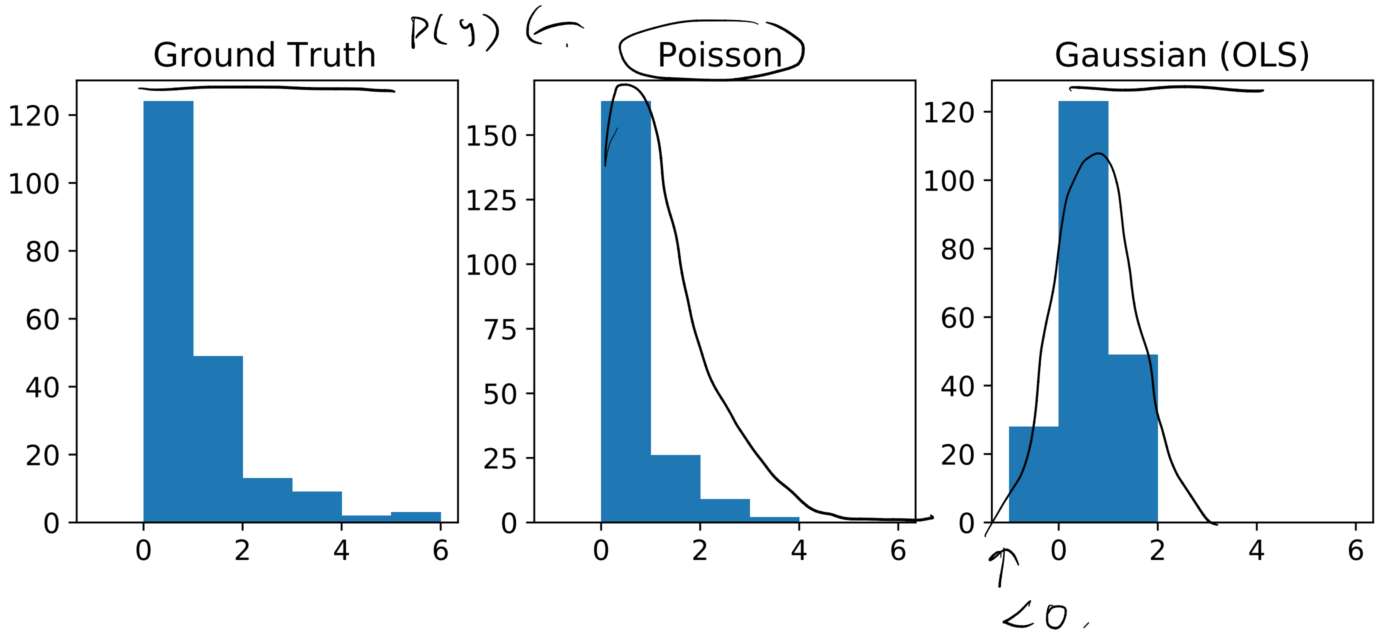
$$\boxed{h_\theta(x) = e^{\theta^T x}}$$



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

$x \cdot y = x^T y$

Probability mass function of a Multinomial distribution over k outcomes

$$a(y)_k = 1 - \sum_{i=1}^{k-1} a(y)_i \qquad \underbrace{p(y;\phi) = \prod_{i=1}^{k} \phi_i^{1\{y=i\}}} \qquad \underbrace{\sum_{i=1}^{K} \phi_i = 1}$$

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

$$\boxed{a(y)_i} = 1\{y=i\}$$

$$a(y) = \begin{bmatrix} a(y)_1 \\ \vdots \\ a(y)_k \end{bmatrix} = \boxed{\begin{bmatrix} 1\{y=1\} \\ \vdots \\ 1\{y=k\} \end{bmatrix}}$$

$$p(y;\phi) = \left(\prod_{i=1}^{k-1} \phi_i^{1\{y=i\}}\right) \phi_k^{1\{y=k\}}$$

$$= e^{\sum_{i=1}^{k-1} a(y)_i \log\phi_i + a(y)_k \log\phi_k}$$

$$a(y)_k = 1 - \sum_{i=1}^{k-1} a(y)_i$$

$$= e^{\sum_{i=1}^{k-1} a(y)_i \log\phi_i + \log\phi_k - \sum_{i=1}^{k-1} a(y)_i \log\phi_k}$$

$$= e^{\sum_{i=1}^{k-1} a(y)_i \log\frac{\phi_i}{\phi_k} + \log\phi_k}$$

$b(y) = 1$

$a(\eta) = -\log\phi_k$

$\eta_i = \log\frac{\phi_i}{\phi_k}$

$\phi_i =$

$$= 1 \cdot e^{a(y)^T \begin{bmatrix} \log\frac{\phi_1}{\phi_k} \\ \vdots \\ \log\frac{\phi_{k-1}}{\phi_k} \end{bmatrix}}\ \eta$$

$T(y) \rightarrow a(y)$

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

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

$$T(y)_i = \mathbf{1}\{y=i\} = \begin{cases} 0 & y \neq i \\ 1 & y = i \end{cases}$$

$\eta_i = \log \frac{\phi_i}{\phi_k}$ } $g^{-1}(\eta)$.

$e^{\eta_i} = \frac{\phi_i}{\phi_k}$

$\phi_i = \phi_k \cdot e^{\eta_i}$

Since $\sum_{i=1}^{K} \phi_i = 1$

$\sum_{i=1}^{K} \phi_k \cdot e^{\eta_i} = 1$.

$\boxed{\phi_k = \frac{1}{\sum_{i=1}^{K} e^{\eta_i}}}$

$a(\eta)$.

$= -\log \phi_k$.

$= -\log \left(\frac{1}{\sum_{i=1}^{K} e^{\eta_i}} \right)$

$= \log \sum_{i=1}^{K} e^{\eta_i}$

$\Rightarrow$. $\phi_i = \phi_k \cdot e^{\eta_i}$

$= \frac{e^{\eta_i}}{\sum_{i=1}^{K} e^{\eta_i}}$ } $g(\eta)$

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

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

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

  $$T(y)_i = \mathbf{1}\{y=i\} = \begin{cases} 0 & y \neq i \\ 1 & y = i \end{cases}$$

- $a(\eta) = -\log(\phi_k) = \log \sum_{i=1}^{k} e^{\eta_i}$

- $$\eta = \begin{bmatrix} \log\left(\frac{\phi_1}{\phi_k}\right) \\ \vdots \\ \log\left(\frac{\phi_{k-1}}{\phi_k}\right) \end{bmatrix}$$

- $b(y) = 1$

GLM example: Softmax regression

Apply GLM construction rules:

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

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

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

GLM example: Softmax regression

Apply GLM construction rules:

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

recall that

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

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

$\downarrow$ inverse

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

2. Derive hypothesis function:

T(y)

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

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

GLM example: Softmax regression

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

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

η_i

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

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

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

$\phi_1 + \cdots + \phi_{k-1} \neq 1$

$\eta_1, \dots, \eta_{k-1}$ are independent parameters

GLM Summary

Sufficient statistic $T(y)$

Response function $g(\eta)$

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

response

link

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