Learning From Data
Lecture 3: Generalized Linear Models

Yang Li **yangli@sz.tsinghua.edu.cn**

March 15, 2024

Today's Lecture

Supervised Learning (Part III)

- Softmax Regression
- Review: exponential families
- Generalized linear models (GLM)

Written Assignment (WA1) will be released tonight. Due in two weeks (Start early!)

Softmax Regression

Review: Solve logistic regression via MLE

- Hypothesis function: logistic function

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

$z = \theta^T x$

Review: Solve logistic regression via MLE

- Hypothesis function: logistic function

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

- Assuming $\underline{y|x};\theta$ is distributed according to Bernoulli$(\underbrace{h_\theta(x)}_{\lambda})$

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

Review: Solve logistic regression via MLE

- Hypothesis function: logistic 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($h_\theta(x)$)

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

Review: Solve logistic regression via MLE

- Hypothesis function: logistic 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:

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

- Maximize $\ell(\theta)$ via (stochastic) gradient descent.

$$\frac{\partial l(\theta)}{\partial \theta_j} =$$

Review: Solve logistic regression via MLE

- Hypothesis function: logistic 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:

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

- Maximize $\ell(\theta)$ via (stochastic) gradient descent.

$$\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

Softmax Regression

- Given $x \in \mathbb{R}^n$, find a hypothesis $h_\theta(x)$ that predicts y that takes value in $\{1, \ldots, k\}$
- y can be represented as one-hot vector. e.g. $[0, 1, 0, \ldots, 0]^T$ indicates $y = 2$

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \quad y = 2$$

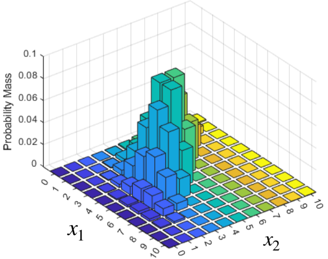
$$h_\theta(x) = \begin{bmatrix} h_{\theta_1}(x) \\ h_{\theta_2}(x) \\ \vdots \\ h_{\theta_k}(x) \end{bmatrix}$$

Review: Multinomial Distribution

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

$$\sum_{i=1}^{k} \phi_i = \phi_1 + \cdots + \phi_k = 1$$

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



Let $y = [y_1, \ldots, y_k]$ be the count of each side, the probability mass function (PMF) is:

$$p(y) = \frac{n!}{y_1! \ldots y_k!} \phi_1^{y_1} \ldots \phi_k^{y_k}$$

When $n = 1$, $\quad y_i \in \{0,1\} \quad \sum_{i=1}^{k} y_i = n = 1$

$$p(y) = \phi_1^{\mathbf{1}\{y=1\}} \ldots \phi_k^{\mathbf{1}\{y=k\}}$$

$$= \prod_{i=1}^{k} \phi_i^{\mathbf{1}\{y=i\}} \qquad \mathbf{1}\{y=i\} = \begin{cases} 1 & y = i \\ 0 & y \neq i \end{cases}$$

Categorical distribution

Extend logistic regression: Softmax Regression

space of output

Assume $p(y|x) \sim Multinomial(h_{\theta_1}(x), \ldots, h_{\theta_k}(x))$ where $k = |\mathcal{Y}|$, $n = 1$

Extend logistic regression: Softmax Regression

Assume $p(y|x) \sim Multinomial(h_{\theta_1}(x), \dots, h_{\theta_k}(x))$ where $k = |\mathcal{Y}|$, $n = 1$

Hypothesis function for sample x:

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

$\theta_i^T x_i$

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

Extend logistic regression: Softmax Regression

Assume $p(y|x) \sim Multinomial(h_{\theta_1}(x), \ldots, h_{\theta_k}(x))$ where $k = |\mathcal{Y}|$, $n = 1$

Hypothesis function for sample x:

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

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

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

Θ

Softmax Regression

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

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

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

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

Softmax Regression

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

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

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

$P\left(\begin{bmatrix} y_1 \\ \vdots \\ y_{k-1} \\ y_k \end{bmatrix}\right) = \text{Multinomial}(\phi_1, \dots, \phi_k)$ $\boxed{\sum_{i=1}^{k} \phi_i = 1}$ y

y_k: # of times $y=k$.

ϕ_1 : prob. of coming up with $y=1$

degree of freedom.

$\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$

- 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.

Invariant to parameter shift

constant. vector.

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

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

Relationship with Logistic Regression

$y = \{1, 2\}$

$\Downarrow$

$y = \{0, 1\}$

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

When K = 2,

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

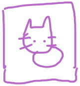
Replace $\boxed{\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}}$ with $\boxed{\theta *} = \theta - \begin{bmatrix} \theta_2 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \theta_1 - \theta_2 \\ 0 \end{bmatrix}$,

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

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

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

When to use Softmax?



- human
- animal ✓
- cat ✓

h_{human}

h_{animal}

h_{cat}

- 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

y $y|x \sim$ i.i.d

What we've learned so far:

Learning task	Model	$p(y\|x;\theta)$
- regression continuous $y \in \mathbb{R}^n$	Linear regression	$\mathcal{N}(h_\theta(x), \sigma^2)$
- binary classification $y \in \{0,1\}$	Logistic regression	Bernoulli($h_\theta(x)$)
multi-class classification $y = \{1, \dots, k\}$.	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	$p(y\|x;\theta)$
regression	Linear regression	$\mathcal{N}(h_\theta(x)\ ,\sigma^2)$
binary classification	Logistic regression	Bernoulli($h_\theta(x)$)
multi-class classification	Softmax regression	Multinomial($[h_\theta(x)]$)

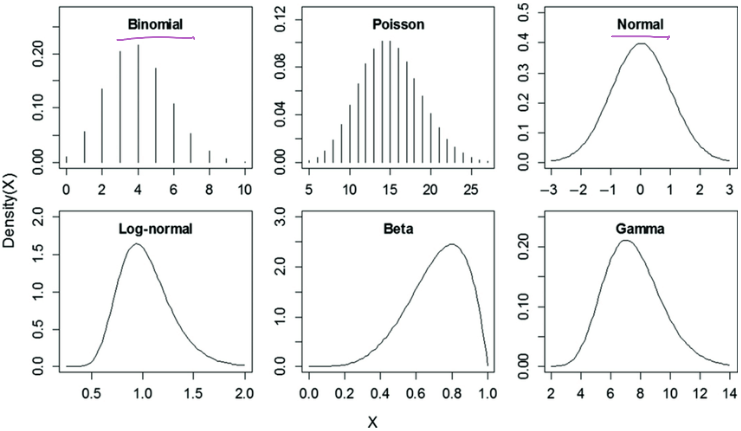
Can we generalize the linear model to other distributions?

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

discrete



Examples of distribution classes in the exponential family.

Exponential Family of Distributions

Bernoulli($\underline{\lambda}$)

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

R.V.

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

sufficient statistic

- y: random variable
- $\theta^T x$ input $\boxed{\eta}$: 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 $\int_y p(y;\eta)dy = 1$).

assume y is discrete

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

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

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

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

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

Gaussian Distribution (unit variance)

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

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

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

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

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

$\eta = \mu$ $\quad T(y)$ $\quad a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$

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 $y \in \mathbb{R}$:

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

sufficient statistics

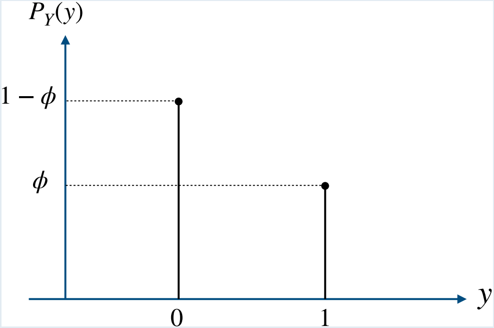
- $\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$
- $T(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$
- $b(y) = \frac{1}{\sqrt{2\pi}}$
- $a(\eta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

Exponential Family Examples

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

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

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

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

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

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

$b(y) = 1$

$T(y) = y.$

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

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

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

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

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*

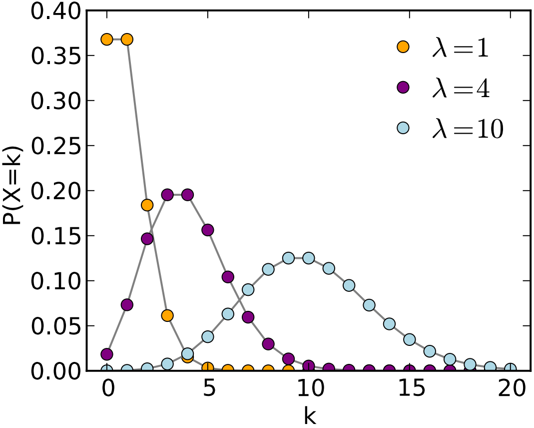
Exponential Family Examples

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 $y \in \mathcal{Y}$: $\underline{e^{\log}}$

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

- $b(y)$
- $T(y)$
- η
- $a(\eta)$

Exponential Family Examples

Poisson distribution Poisson(λ)

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

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

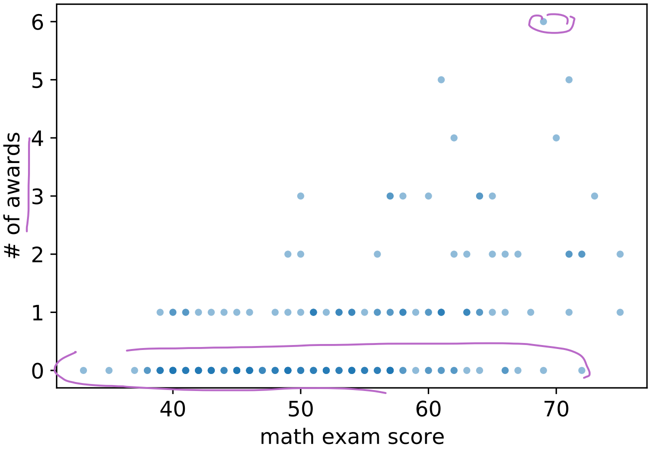
- $\eta = \log \lambda$
- $b(y) = \frac{1}{y!}$
- $T(y) = y$
- $a(\eta) = e^\eta$

Generalized Linear Models

Generalized Linear Models: Intuition

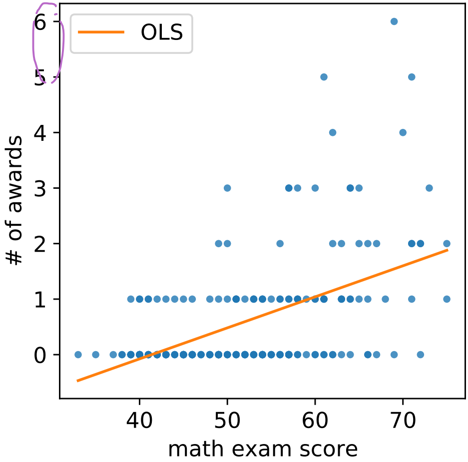
Example 1: Award Prediction

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



Generalized Linear Models: Intuition

y is discrete

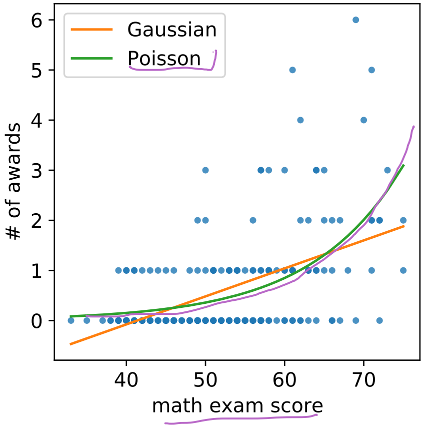


Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution.
- Assumes change in x is → 2x proportional to change in y → 2y

6.45

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

$\eta = \log(\lambda) \in$ link function.

$\boxed{\eta}$ $\lambda = \mathbb{E}[y|x]$

$\propto x$

$\log(\lambda) \propto x$ $\lambda = e^{\eta}$ response function.

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. $log(\lambda) = \boxed{\theta^T x}$

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: e.g. $\begin{bmatrix} y \\ y^2 \end{bmatrix}$

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\boxed{\mathbb{E}[T(y)|x]}$
   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: (1D)

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

   η is a vector: (nD)

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

Generalized Linear Models: Construction

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

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

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

a.k.a. the “mean function”

Generalized Linear Models: Construction

e.g. $g^{-1}(\lambda) = \log(\lambda)$

$g(\eta) = e^{\eta}$.

g^{-1} link function

$$\boxed{\eta} \underset{g}{\overset{g^{-1}}{\rightleftarrows}} \mathbb{E}[T(y)|x] = \lambda$$

$\propto x$

canonical response

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

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

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

$T(y) = y \quad \rightarrow \quad \mathbb{E}[y|x]$

  a.k.a. the "mean function"

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

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

GLM example: ordinary least square

Apply GLM construction rules:

identity link

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

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

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:

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

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 $\eta = \theta^T x$:

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

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:

logit link function.

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

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

GLM example: logistic regression

Apply GLM construction rules:

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

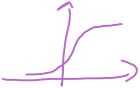
$$\eta = \log\left(\frac{\phi}{1-\phi}\right),\ T(y) = y$$

2. Derive hypothesis function:

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

response function

GLM example: logistic regression



Apply GLM construction rules:

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

link g^{-1}

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

2. Derive hypothesis function:

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

response g

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

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

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

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

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

log link function

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

link $\eta = \log(\lambda)$, $T(y) = y$

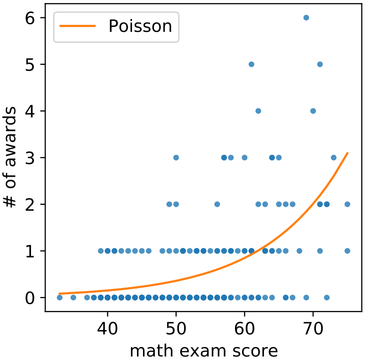
2. Derive hypothesis function:

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

response

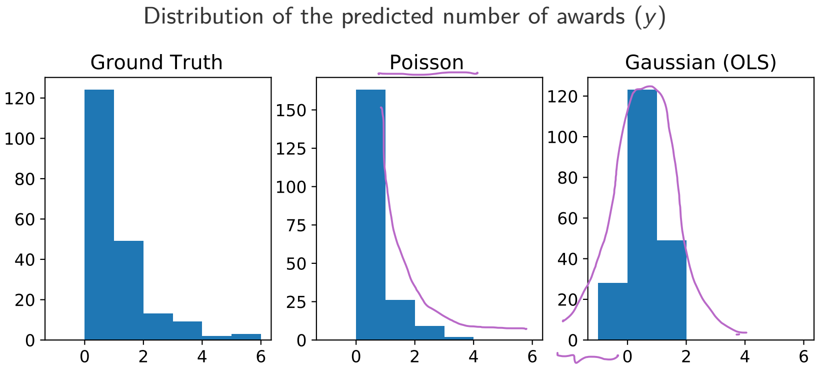
3. Adopt linear model $\eta = \theta^T x$:

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



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

GLM example: Poisson regression



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

GLM example: Softmax regression

Multinomial$(\phi_1, \dots, \phi_k)$.

Probability mass function of a Multinomial distribution over k outcomes

$\sum_{i=1}^{k} 1\{y_i = 1\} = 1$

$\sum_{i=1}^{k} \partial(y_i) = 1$

$\partial(y_k) = 1 - \sum_{i=1}^{k-1} \partial(y_i)$

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

$\rightarrow \partial(y_i) = \begin{cases} 1 & y=i \\ 0 & \text{o.w.} \end{cases}$

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

$p(y;\phi) = \left(\prod_{i=1}^{k-1} \phi_i^{\partial(y_i)}\right) \phi_k^{\partial(y_k)} = e^{\log\left(\prod_{i=1}^{k-1} \phi_i^{\partial(y_i)}\right) \phi_k^{\partial(y_k)}}$, $\phi_k^{1-\sum_{i=1}^{k-1}\partial(y_i)}$

$= e^{\sum_{i=1}^{k-1} \partial(y_i) \log \phi_i + (1 - \sum_{i=1}^{k-1} \partial(y_i)) \log \phi_k}$

$\log \phi_k - \log \phi_k \sum_{i=1}^{k-1} \partial(y_i)$

$= e^{\sum_{i=1}^{k-1} \partial(y_i)(\log \phi_i - \log \phi_k) + \log \phi_k}$

$= e^{\sum_{i=1}^{k-1} \partial(y_i) \log \frac{\phi_i}{\phi_k} + \log \phi_k}$

$= e^{\begin{bmatrix} \partial y_1 \\ \vdots \\ \partial y_{k-1} \end{bmatrix} \cdot \begin{bmatrix} \log \frac{\phi_1}{\phi_k} \\ \vdots \\ \log \frac{\phi_{k-1}}{\phi_k} \end{bmatrix} + \log \phi_k}$

$T(y) = \begin{bmatrix} \partial(y_1) \\ \vdots \\ \partial(y_{k-1}) \end{bmatrix} = \begin{bmatrix} 1\{y=1\} \\ \vdots \\ 1\{y=k-1\} \end{bmatrix}$

link

$\eta = \begin{bmatrix} \log \frac{\phi_1}{\phi_k} \\ \vdots \\ \log \frac{\phi_{k-1}}{\phi_k} \end{bmatrix} \Rightarrow \eta_i = \log \frac{\phi_i}{\phi_k} \Rightarrow e^{\eta_i} = \frac{\phi_i}{\phi_k}$. $(\phi_i = \phi_k e^{\eta_i})$

$a(y) = \log \phi_k$

$\sum_{i=1}^{k} \phi_i = \sum_{i=1}^{k} \phi_k e^{\eta_i} = 1$

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

$b(y) = 1$.

GLM example: Softmax regression

$\phi_i = \frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}}$ } canonical response.

Probability mass function of a Multinomial distribution over k outcomes

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

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

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

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

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

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

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 $i = 1 \ldots k-1$

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

$h_\theta(x) = \mathbb{E}[\,T(y)|x\,]$

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 $y|x;\theta \sim \text{Multinomial}(\phi_1, \ldots, \phi_k)$, 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*

2. Derive hypothesis function:

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

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

GLM example: Softmax regression

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

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

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

GLM example: Softmax regression

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

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

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

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

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

GLM Summary

Sufficient statistic $T(y)$

Response function $g(\eta)$

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

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

GLM is effective for modelling different types of distributions over y