Learning From Data
Lecture 3: Generalized Linear Models

Yang Li   yangli@sz.tsinghua.edu.cn

September 27, 2020

Today's Lecture

Supervised Learning (Part II)

- Review on linear and logistic regression
- Digress on probability: exponential families
- Generalized linear models (GLM)
- Discriminative vs. generative learning

Programming Assignment (PA1) is released. Due on Oct 9th.

Review of Lecture 2

Review of Lecture 2: Linear least square

- Hypothesis function for input feature $x^{(i)} \in \mathbb{R}^n$:
  $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)}$

Review of Lecture 2: Linear least square

- Hypothesis function for input feature $x^{(i)} \in \mathbb{R}^n$:
  $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)}$
- Vector notation: $h_\theta(x^{(i)}) = \theta^T x^{(i)}$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$, $x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$

Review of Lecture 2: Linear least square

- Hypothesis function for input feature $x^{(i)} \in \mathbb{R}^n$:
  $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)}$
- Vector notation: $h_\theta(x^{(i)}) = \theta^T x^{(i)}$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$, $x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$
- Cost function for m training examples $(x^{(i)}, y^{(i)}), i = 1, \ldots, m$:

$$J(\theta) =$$

Review of Lecture 2: Linear least square

- Hypothesis function for input feature $x^{(i)} \in \mathbb{R}^n$:
  $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)}$
- Vector notation: $h_\theta(x^{(i)}) = \theta^T x^{(i)}$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$, $x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$
- 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:

  update rule (batch)

  update rule (stochastic)

- Newton's method

- Normal equation

How to minimize $J(\theta)$?

- Gradient descent:

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

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

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

- Newton's method

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

- Normal equation

$$\underbrace{X^T X \theta = X^T y}$$

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.

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

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.

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

(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.

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

(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: Linear Regression Exercise

$$\begin{bmatrix} | & | & \\ a_1 & a_2 & \cdots \\ | & | & \end{bmatrix}$$

The normal equation for solving ordinary least square is:

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

$Ax = y$

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

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($h_\theta(x)$)

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

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$(h_\theta(x))$

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

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:

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

Review of Lecture 2: Softmax regression

- Hypothesis function:

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

Review of Lecture 2: Softmax regression

- Hypothesis function:

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

- Assume $y|x;\theta$ is distributed according to Multinomial($\underbrace{[h_\theta(x)]}$):

$$\underline{p(y|x;\theta)} = \prod_{l=1}^{k} p(y=l|x;\theta)^{\underbrace{\mathbf{1}\{y=l\}}} \begin{cases} 1 & y=l \\ 0 & \text{o.w.} \end{cases}$$

$p(y|x)^y$

Review of Lecture 2: Softmax regression

- Hypothesis function:

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

- Assume $y|x;\theta$ is distributed according to Multinomial$(h_\theta(x))$:

$$p(y|x;\theta) = \prod_{l=1}^{k} p(y=l|x;\theta)^{\mathbf{1}\{y=l\}}$$

- Log-likelihood function for m training examples:

$$\ell(\theta) = \sum_{i=1}^{m} \sum_{l=1}^{k} \underbrace{\mathbf{1}\{y^{(i)} = l\}} \log \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}$$

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?

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

y

$p(y|x;\theta)$

A class of distributions is in the **exponential family** if it can be written as

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

- y: random variable response variable.
- η : natural/canonical parameter
- $T(y)$: sufficient statistic of the distribution
- $b(y)$:
- $a(\eta)$: log partition function (why?)

Exponential Family

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

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

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

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

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

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

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

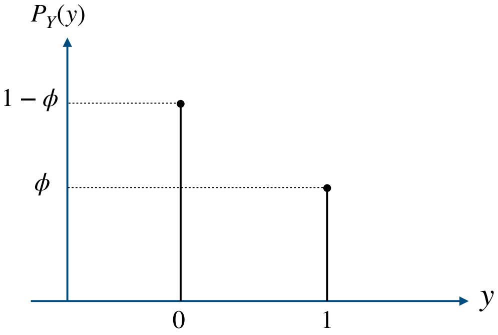
$a(\eta) = \log\left(\sum_y b(y)e^{\eta^T T(y)}\right).$

Exponential Family Examples

Bernoulli Distribution

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

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



Bernoulli Distribution

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

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

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

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

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

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

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

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

$b(y) = 1$ $\quad T(y) = y$ $\quad \eta$

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

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

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

$$\boxed{= \log(1+e^{\eta})}$$

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

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

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

$$e^{\eta} = \phi + e^{\eta}\phi$$

$$= (1+e^{\eta})\phi$$

$$\phi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} \leftarrow \text{sigmoid function.}$$

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

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

Gaussian Distribution (unit variance) $\sigma = 1$

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

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

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

$$= \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\right] e^{-\frac{1}{2}(\mu^2 - 2y\mu)}$$

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

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

$\eta = \mu$

$g = g^{-1} = \text{identity}$

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

Gaussian Distribution

$\sigma \neq 1.$

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

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

$T(y)$

$\mu = \eta_1 \cdot \sigma^2$

$\sigma =$

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

Try this before attempting the next written homework

Exponential Family Examples

Poisson distribution: Poisson(λ)

$y \sim \text{Poisson}(\lambda)$.

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*

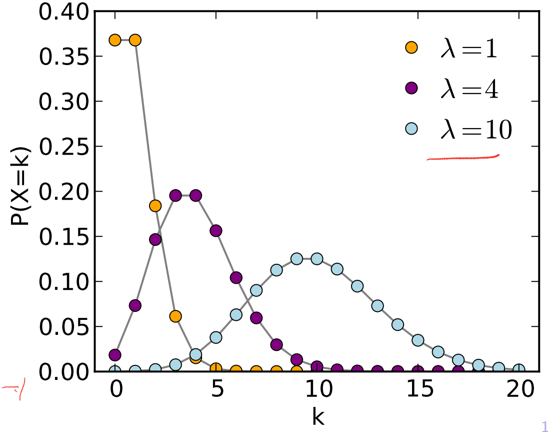
Exponential Family Examples

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

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

Poisson distribution Poisson(λ)

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

$e^{\log P(y;\lambda)}$

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

$$P(y;\lambda) = \frac{1}{y!} e^{\log(\lambda^y e^{-\lambda})}$$

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

$a = \lambda = e^{\eta}$

$\eta = \log\lambda$

$\lambda = e^{\eta}$

$T(y) = y$

Exponential Family Examples

Poisson distribution Poisson(λ)

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

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

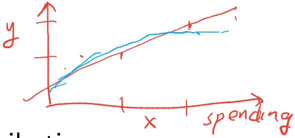
- $\eta = \log \lambda$
- $b(y) = \frac{1}{y!}$
- $T(y) = y$
- $a(\eta) = e^\eta$

Generalized Linear Models

Generalized Linear Models: Intuition

Example 1: Customer Prediction

Predict y, **the number of customers** in the store given x, the recent spending in advertisement.



Problems with linear regression:

- Assumes $y|x;\theta$ has a Normal distribution.
  Poisson *distribution is better for modeling occurrences*
- A constant change in x leads to a constant change in y
  More realistic to have a constant ***rate*** *of increased number of customers* (e.g. doubling or halving y)

Generalized Linear Models: Intuition

Example 2: Purchase Prediction (0, 1) 70%

Predict y, **the probability a customer would make a purchase** given x, the recent spending in advertisement.

Problems with linear regression:

- Assumes $y|x;\theta$ is a Normal distribution.
  Bernoulli *distribution is better for modeling the probability of a binary choice*
- A constant change in x leads to a constant change in y
  More realistic to have a constant change in the ***odds*** *of increased probability* (e.g. from 2 : 1 odds to 4 : 1)

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

Generalized Linear Models: Construction

$T(y) = y$

η

$\underbrace{\mathbb{E}[T(y)]}_{\mu}$

Formal GLM assumptions & design decisions:

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

canonical response

$$\begin{tikzcd} \eta \arrow[r, bend left, "g(\eta)"] & \mathbb{E}_y\left[T(y)\right] \arrow[l, bend left, "g^{-1}(T(y))"] \end{tikzcd}$$

natural parameter

canonical link func

mean of sufficient statistics of y

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

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

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

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

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

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

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

GLM example: ordinary least square

Apply GLM construction rules:

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

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

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 $\eta = \theta^T x$:

$T(y) = g(\eta).$

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

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

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

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:

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

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 $\eta = \theta^T x$:

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

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:

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

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

$g^{-1}(\phi)$

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

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: Customer Prediction

Predict y, **the number of customers** in the store given x, the recent spending in advertisement.

Use GLM to find the hypothesis function...

GLM example: Poisson regression

Apply GLM construction rules:

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

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

2. Derive hypothesis function:

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

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: Softmax regression

Probability mass function of a Multinomial distribution over k outcomes

$\underbrace{\phi_1, \ldots, \phi_k}$

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

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

- $T(y) = \begin{array}{c} \left[\begin{matrix} \boxed{\mathbf{1}\{y=1\}} \\ \vdots \\ \mathbf{1}\{y=k-1\} \end{matrix}\right] \end{array} \begin{array}{l} T(y)_1 \\ \vdots \\ T(y)_{k-1} \end{array}$

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

For the consistency of notation, we use the shorthand $T(y)_i$ to denote $1\{y=i\}$ and write $T(y) = \begin{bmatrix} T(y)_1 \\ \vdots \\ T(y)_{k-1} \end{bmatrix}$

GLM example: Softmax regression

canonical link function

$T(y) = \log \frac{\phi_i}{\phi_k}$ $\quad \phi_i = \frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}}$

Probability mass function of a Multinomial distribution over k outcomes

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

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

$\eta_k = \log \frac{\phi_k}{\phi_k} = 0.$

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

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

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

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

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

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

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

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

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

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

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

$$= \underbrace{1}_{b(y)}\, e^{\begin{bmatrix} \log \frac{\phi_1}{\phi_k} \\ \vdots \\ \log \frac{\phi_{k-1}}{\phi_k} \end{bmatrix}^T T(y) + \log \phi_k}$$

$T(y) = T(y)$

$\eta = \begin{bmatrix} \log \frac{\phi_1}{\phi_k} \\ \vdots \\ \log \frac{\phi_{k-1}}{\phi_k} \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_{k-1} \end{bmatrix}$

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

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 $i = 1 \ldots k-1$

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

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

2. Derive hypothesis function:

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

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

GLM example: Softmax regression

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

η_i

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

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

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 $g(\eta)$

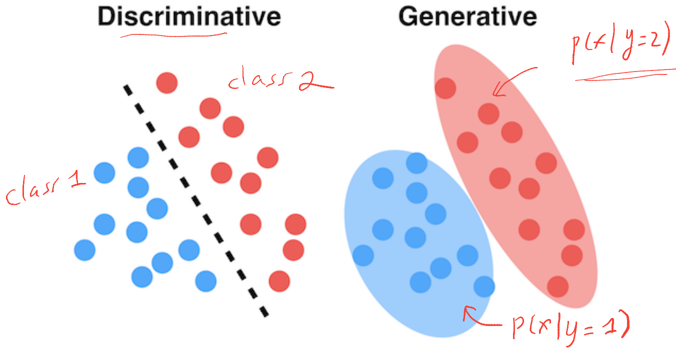
Link function $g^{-1}(\mathbb{E}[T(y);\eta])$

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($\phi_1,\ldots,\phi_k$)	$\{1,\ldots,k\}$	δ_i	$\frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$	$\eta_i = \log\left(\frac{\phi_i}{\phi_k}\right)$

Discriminative & Generative Models

Two Learning Approaches

Classify input data x into two classes $y \in \{0, 1\}$

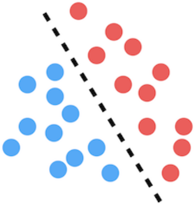


Discriminate between classes of data points

Model the underlying distribution of the data

Discriminative Learning Algorithms

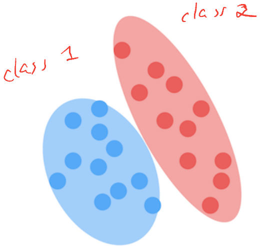
A class of learning algorithms that try to learn the **conditional probability** $p(y|x)$ directly or learn mappings directly from $\mathcal{X}$ to $\mathcal{Y}$.



- e.g. linear regression, logistic regression, k-Nearest Neighbors ...

Generative Learning Algorithms

A class of learning algorithms that model the **joint probability** $p(x, y)$.



- Equivalently, generative algorithms model $p(x|y)$ and $p(y)$
- $p(y)$ is called the **class prior**
- Learned models are transformed to $p(y|x)$ later to classify data using Bayes' rule

Bayes Rule

The posterior distribution on y given x:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Bayes Rule

The posterior distribution on y given x:

posterior

class prior.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Make predictions in a generative model:

$$\boxed{\operatorname*{argmax}_y p(y|x)} = \operatorname*{argmax}_{y\ \theta} \frac{p(x|y)p(y)}{p(x)}$$
$$= \operatorname*{argmax}_y p(x|y)p(y)$$

No need to calculate $p(x)$.

Generative Models

Generative classification algorithms:

- Continuous input: Gaussian Discriminant Analysis
- Discrete input: Naïve Bayes