Learning From Data
Lecture 3: Generalized Linear Models

Yang Li   yangli@sz.tsinghua.edu.cn

9/29/2019

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 $x^{(i)} \in \mathbb{R}^n$:
  $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)}$
- 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) $$\theta_j \leftarrow \theta_j + \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

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

- Newton's method

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

- Normal equation

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

Review of Lecture 2

Maximum likelihood estimation

- Log-likelihood function:

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

  where p is a probability density function.

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

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

True under the assumptions:

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

Review of Lecture 2: Linear Regression Exercise

The normal equation for solving ordinary least square is:

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

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

Review of Lecture 2: Logistic regression

- Hypothesis function:

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

- 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:

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

- 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:

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

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?

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

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

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

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

Exponential Family

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

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

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

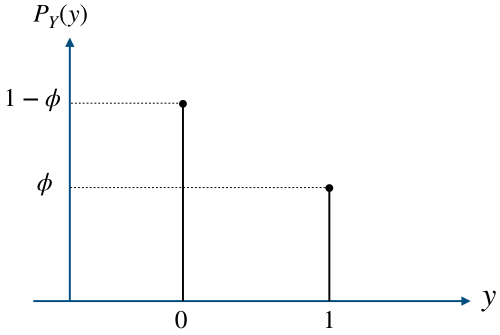
$$a(\eta) = \log\left(\sum_y b(y)e^{\eta^T T(y)}\right)$$

Exponential Family Examples

Bernoulli Distribution

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

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



Bernoulli Distribution

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

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

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

Exponential Family Examples

Bernoulli Distribution

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

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

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

Exponential Family Examples

Gaussian Distribution (unit variance)

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

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

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

Exponential Family Examples

Gaussian Distribution

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

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

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

Try this before attempting the homework

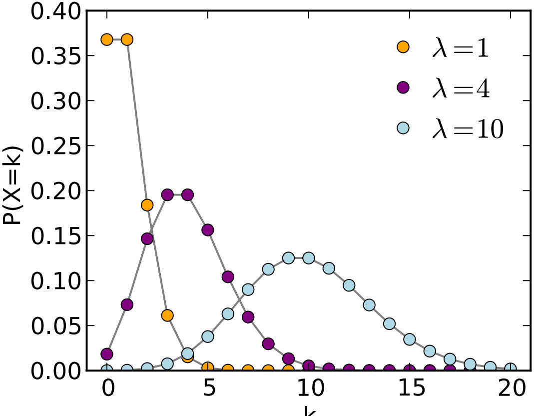
Exponential Family Examples

Poisson distribution: Poisson(λ)

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

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

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



Exponential Family Examples

Poisson distribution Poisson(λ)

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

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

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

Generalized Linear Models

Generalized Linear Models: Intuition

Example 1: 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

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

Formal GLM assumptions & design decisions:

1. $y|x;\theta \sim \text{ExponentialFamily}(\eta)$
   e.g. Gaussian, Poisson, Bernoulli, Multinomial, Beta ...
2. The hypothesis function $h(x)$ is $\mathbb{E}\left[T(y)|x\right]$
   e.g. When $T(y) = y$, $h(x) = \mathbb{E}\left[y|x\right]$
3. The natural parameter η and the inputs x are related linearly:
   η is a number:
   $$\eta = \theta^T x$$
   η is a vector:
   $$\eta_i = \theta_i^T x \quad \forall i = 1, \ldots, n \qquad \text{or} \qquad \eta = \Theta^T x$$

Generalized Linear Models: Construction

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

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

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

  a.k.a. the "mean function"

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

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

GLM example: ordinary least square

Apply GLM construction rules:

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

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

2. Derive hypothesis function:

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

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

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

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

GLM example: logistic regression

Apply GLM construction rules:

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

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

2. Derive hypothesis function:

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

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

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

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

GLM example: Poisson regression

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

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

2. Derive hypothesis function:

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

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

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

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

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

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

GLM example: Softmax regression

Apply GLM construction rules:

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

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

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

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 $\eta_i = \theta_i^T x$:

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

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

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

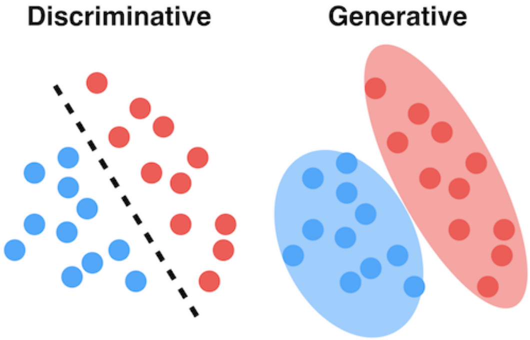
Sufficient statistic	$T(y)$
Response function	$g(\eta)$
Link function	$g^{-1}(\mathbb{E}[T(y);\eta])$

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 $y \in \{0, 1\}$

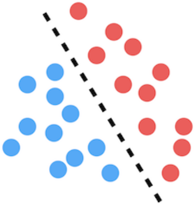


Discriminate between classes of data points

Model the underlying distribution of the data

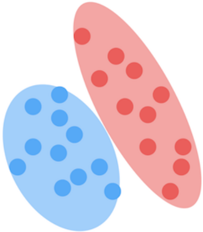
Discriminative Learning Algorithms

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:

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

Make predictions in a generative model:

$$\begin{aligned} \operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y) \end{aligned}$$

No need to calculate $p(x)$.

Generative Models

Generative classification algorithms:

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