Regression Tasks - Generalized Linear Models

Astrini Sie

asie@seattleu.edu

Week 4b

ECEGR4750 - Introduction to Machine Learning
Seattle University

October 12, 2023

Recap and Updates

- Lab Take Home Assignment due this Thursday at 11.59pm
- Office Hours
  - T, Th 12-1p at Bannan 224
  - W 7-9p via Zoom
  - F 9-9.45a via Zoom
- Zoom Link: `https://seattleu.zoom.us/j/7519782079?pwd=cnhCM2tPcHJKVWwxZVArS2VHSUNJZz09`
  - Meeting ID: 751 978 2079
  - Passcode: 22498122

- Linear Regression (Regression):
  - Closed form solution: OLS
  - Numerical solution: LMS (GD, BGD, SGD)
  - Effect of noise on regression
- Logistic Regression (Binary Classification):
  - Numerical solution: MLE
  - Another numerical solution: Newton's Method

Overview

Generalized Linear Models

We have seen two regression examples that are distinguished by their types of output distribution:

- **Linear Regression:** $y \in \mathbb{R}$
- **Logistic Regression / Binary Classification:** $y \in 0, 1$

Generalized Linear Models

We have seen two regression examples that are distinguished by their types of output distribution:

- **Linear Regression:** $y \in \mathbb{R}$
- **Logistic Regression / Binary Classification:** $y \in 0, 1$

Which also means:

- **Linear Regression:** y is a Gaussian distribution: $y|x;\theta \sim \mathcal{N}(\mu, \sigma^2)$
- **Logistic Regression / Binary Classification:** y is a Bernoulli distribution: $y|x;\theta \sim Bernoulli(\phi)$

where μ and ϕ are functions of x and θ

Generalized Linear Models

We can generalize this approach for broader family of models based on their output distribution, called **Generalized Linear Models (GLMs)**

Generalized Linear Models

We can generalize this approach for broader family of models based on their output distribution, called **Generalized Linear Models (GLMs)**

For example:

- Gaussian (y = real numbers)
- Bernoulli (y = binary)
- Multinomial (y = multi-class)
- Poisson (y = counts)
- Beta & Dirichlet (y = probabilities)

Generalized Linear Models

We can generalize this approach for broader family of models based on their output distribution, called **Generalized Linear Models (GLMs)**

For example:

- Gaussian (y = real numbers)
- Bernoulli (y = binary)
- Multinomial (y = multi-class)
- Poisson (y = counts)
- Beta & Dirichlet (y = probabilities)

We use the exponential family to unify inference and learning for many important models

Exponential Family

"If P has a special form, then inference and learning come for free"

Exponential Family of Generalized Linear Models

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\} \tag{1}$$

where y, $a(\eta)$, and $b(y)$ are scalars, and $T(y)$ have the same dimension as η.

- η is the **natural parameter** or **canonical parameter** of the distribution.
- $T(y)$ is the **sufficient statistic**, where often $T(y) = y$.
- $b(y)$ is the **base measure**.
- $a(\eta)$ is the **log partition function**. $e^{-a(\eta)}$ plays the role of a normalization constant to ensure the distribution $p(y;\eta)$ sums or integrates over y to 1.

Exponential Family

"If P has a special form, then inference and learning come for free"

Exponential Family of Generalized Linear Models

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}$$

where y, $a(\eta)$, and $b(y)$ are scalars, and $T(y)$ have the same dimension as η.

A fixed choice of T, a, and b defines a family of distributions that is parameterized by η. As we vary η, we get different distributions within this family.

Let's look at a couple examples...

Gaussian Distribution

(Linear Regression)

Gaussian distribution with a mean of μ and variance of σ^2, over $y \in \mathbb{R}$ is written as $\mathcal{N}(\mu, \sigma^2)$.

The value of the variance σ^2 has no effect on the final choice of θ and $h_\theta(x)$. Thus, to simplify derivation, let's set $\sigma^2 = 1$

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

Gaussian Distribution

Let's derive Equation 1 from Equation 2. First, multiply out the square and group terms:

Gaussian Distribution

Let's derive Equation 1 from Equation 2. First, multiply out the square and group terms:

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}(y-\mu)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}(y^2 - 2\mu y + \mu^2)\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)\right) \exp\left\{\mu y - \tfrac{1}{2}\mu^2\right\}
\end{aligned}
$$

This equation is the same as Equation 1:

$$
P(y;\eta) = b(y) \exp\left\{\eta^T T(y) - a(\eta)\right\}
$$

Generalized Linear Model for Gaussian Distribution

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

where:

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

$$a(\eta) = \frac{1}{2}\mu^2 = \frac{1}{2}\eta^2$$

$$T(y) = y$$

$$\eta = \mu$$

Bernoulli Distribution

(Binary Classification or Logistic Regression)

Bernoulli distribution with a mean of ϕ, over $y \in 0, 1$ is written as *Bernoulli*(ϕ).

$$p(y = 1; \phi) = \phi$$
$$p(y = 0; \phi) = 1 - \phi$$

Which can be written as (see previous lecture):

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y} \tag{3}$$

Bernoulli Distribution

Let's derive Equation 1 from Equation 3. First, take the log of $p(y;\phi)$ and exp of the log ($p(y;\phi) = \exp(\log(p(y;\phi)))$:

Bernoulli Distribution

Let's derive Equation 1 from Equation 3. First, take the log of $p(y;\phi)$ and exp of the log ($p(y;\phi) = \exp(\log(p(y;\phi)))$:

$$\begin{aligned}
p(y;\phi) &= \exp(\log(p(y;\phi))) \\
&= \exp(\log(\phi^y(1-\phi)^{1-y})) \\
&= \exp(y\log\phi + (1-y)\log(1-\phi)) \\
&= \exp(y\log\phi - y\log(1-\phi) + \log(1-\phi)) \\
&= \exp\left(y\log\left(\tfrac{\phi}{1-\phi}\right) + \log(1-\phi)\right)
\end{aligned}$$

Bernoulli Distribution

Generalized Linear Model for Bernoulli Distribution

$$p(y;\phi) = \exp\left(y \log\left(\tfrac{\phi}{1-\phi}\right) + \log(1-\phi)\right) \tag{4}$$

This Equation 4 is in the same form as Equation 1:

$$P(y;\eta) = b(y)\exp\left\{\eta^T T(y) - a(\eta)\right\}$$

where:

$$b(y) = 1$$
$$a(\eta) = -\log(1-\phi)$$
$$T(y) = y$$
$$\eta = \log\frac{\phi}{1-\phi}$$

Bernoulli Distribution

Let's express $a(\eta)$ in terms of η and verify that it is a function of η.

Bernoulli Distribution

Let's express $a(\eta)$ in terms of η and verify that it is a function of η.

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

Plug ϕ back into $a(\eta)$:

Bernoulli Distribution

Let's express $a(\eta)$ in terms of η and verify that it is a function of η.

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

Plug ϕ back into $a(\eta)$:

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

This equation is the same form as Equation 1

Bernoulli Distribution

Generalized Linear Model for Bernoulli Distribution

$$p(y;\phi) = \exp\left(y\log\left(\tfrac{\phi}{1-\phi}\right) + \log(1-\phi)\right)$$

where:

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

Multinomial Distribution

(Multi-Class Classification)

Multi-class Classification: y can take on any of k values

Given a training set $\{(x^{(i)}, y^{(i)}$ for $i = 1, \ldots, n\}$, let $y^{(i)} \in 1, 2, \ldots, k$.

For example, we want to choose whether the Iris in the picture belongs to one of the 3 classes: 'Setosa', 'Virginica', 'Versicolor'.
In this case, $k = 3$.
We can perform a **one-hot encoding**, in which $y \in \{0, 1\}^k$, and $\sum_{j=1}^{k} y_j = 1$.

$$\underset{'Setosa'}{\begin{pmatrix}1\\0\\0\end{pmatrix}} \quad \underset{'Virginica'}{\begin{pmatrix}0\\1\\0\end{pmatrix}} \quad \underset{'Versicolor'}{\begin{pmatrix}0\\0\\1\end{pmatrix}}$$

Find the prediction, which is the distribution over the k classes.

Define the Softmax function as our hypothesis. *Softmax* : $\mathbb{R}^k \to \mathbb{R}^k$ turns $(t_1, \ldots, t_k) = (\theta_1^T x, \ldots, \theta_k^T x)$ into a probability vector with non-negative entries that sum up to 1:

$$\begin{bmatrix} P(y = 1|x;\theta) \\ \vdots \\ P(y = k|x;\theta) \end{bmatrix} = softmax(t_1, \ldots, t_k) = \begin{bmatrix} \frac{exp(\theta_1^T x)}{\sum_{j=1}^k exp(\theta_j^T x)} \\ \vdots \\ \frac{exp(\theta_k^T x)}{\sum_{j=1}^k exp(\theta_j^T x)} \end{bmatrix}$$

where $x, \theta_j \in \mathbb{R}^{d+1}$ for $j = 1, \ldots, k$

Multinomial Distribution

Loss Function

We can shorten $\phi_i = \frac{exp(\theta_i^T x)}{\sum_{j=1}^k exp(\theta_j^T x)}$, hence:

$$P(y = i|x; \theta) = \phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)}$$

How do we train for θ?

Let's define cross-entropy loss $\ell_{ce} : \mathbb{R}^k$:

Cross-Entropy Loss

$$\ell_{ce} = -\sum_{i=1}^{k} t_j \log(\phi_j)$$

where t_j is the truth label for class j and ϕ_j is the softmax probability.

Multinomial Distribution

Loss Function

The gradient of the loss function is:

$$\frac{\partial \ell_{ce}(\theta)}{\partial \theta_j} = \sum_{j=1}^{k} \left(\phi_j^{(i)} - 1\{y^{(i)} = j\} \right) \dot{x}^{(i)}$$

where $\phi_j = \frac{exp(\theta_i^T x)}{\sum_{j=1}^{k} exp(\theta_j^T x)}$ is the probability that the model predicts class j for sample $x^{(i)}$.

We can iterate θ using gradient descent methods to minimize the loss function $\ell(\theta)$.

Multinomial Distribution

Predicting the Output

After obtaining the final values of θ, calculate ϕ_j for each class j. The class with the greatest value of ϕ will be returned as the predicted class.

Constructing GLMs

There are many types of distributions (Exponential Families):

- Gaussian (y = real numbers)
- Bernoulli (y = binary)
- Multinomial (y = multi-class)
- Poisson (y = counts)
- Beta & Dirichlet (y = probabilities)

We can create a general rule to construct a GLM.

Constructing GLMs

Given inputs $x \in \mathbb{R}^{d+1}$ (where d is the number of features) and a target y, create a model $h_\theta(x)$

Constructing GLMs

Given inputs $x \in \mathbb{R}^{d+1}$ (where d is the number of features) and a target y, create a model $h_\theta(x)$

1. $y|x, \theta \sim ExponentialFamily(\eta)$
   Given feature x and weight θ, the distribution of target y follows some exponential family distribution with a parameter of η.

Constructing GLMs

Given inputs $x \in \mathbb{R}^{d+1}$ (where d is the number of features) and a target y, create a model $h_\theta(x)$

1. $y|x, \theta \sim ExponentialFamily(\eta)$
   Given feature x and weight θ, the distribution of target y follows some exponential family distribution with a parameter of η.
2. $\eta = \theta^T x$, in which $\theta, x \in \mathbb{R}^{d+1}$
   Assume a linear model, in which the inputs x and the natural parameter η are linearly related.

Constructing GLMs

Given inputs $x \in \mathbb{R}^{d+1}$ (where d is the number of features) and a target y, create a model $h_\theta(x)$

1. $y|x, \theta \sim ExponentialFamily(\eta)$
   Given feature x and weight θ, the distribution of target y follows some exponential family distribution with a parameter of η.
2. $\eta = \theta^T x$, in which $\theta, x \in \mathbb{R}^{d+1}$
   Assume a linear model, in which the inputs x and the natural parameter η are linearly related.
3. $h_\theta(x) = \mathrm{E}[y|x; \theta]$
   Predict the expected value $T(y)$ given x. In our examples, we assume $T(y) = y$. This means, prediction will satisfy $h_\theta(x) = \mathrm{E}[y|x; \theta]$. Note that, the expected value is equivalent to the arithmetic mean.

Constructing GLMs

Terminologies

Model Parameter	Natural Parameter	Canonical Parameter
θ	η	$\mathrm{E}[T(y);\eta]$ $\begin{pmatrix} \phi\text{: Bernoulli} \\ \mu\text{: Gaussian} \\ \lambda\text{: Poisson} \end{pmatrix}$

- Linear Function: $\eta = \theta^T x$
  relates the model parameters θ and the natural parameter η.
- Canonical Response Function: $g(\eta) = \mathrm{E}[T(y);\eta]$
  expresses the mean of the distribution $\mathrm{E}[T(y);\eta]$ as a function of the natural parameter η. The canonical function g varies depending on the distribution $ExponentialFamily(\eta)$.
- Canonical Link Function: g^{-1}

Constructing GLMs

Examples

1. Bernoulli Distribution
    1. $y|x, \theta \sim Bernoulli(\phi)$

Constructing GLMs

Examples

1. Bernoulli Distribution
    1. $y|x, \theta \sim Bernoulli(\phi)$
    2. $g(\eta) = \mathrm{E}[T(y); \eta]$
       $1/(1 + e^{-\eta}) = \phi$

Constructing GLMs

Examples

1. Bernoulli Distribution
   1. $y|x, \theta \sim Bernoulli(\phi)$
   2. $g(\eta) = \mathrm{E}[T(y); \eta]$
      $1/(1 + e^{-\eta}) = \phi$
   3. $h_\theta(x) = \mathrm{E}[y|x; \theta]$
      $h_\theta(x) = \phi$ (mean of a Bernoulli distribution)
      $h_\theta(x) = 1/(1 + e^{-\eta})$ (substituting link function from 2)
      $h_\theta(x) = 1/(1 + e^{-\theta^T x})$ (substituting linear function $\eta = \theta^T x$)

Constructing GLMs

Examples

1. Bernoulli Distribution
   1. $y|x,\theta \sim Bernoulli(\phi)$
   2. $g(\eta) = \mathrm{E}[T(y);\eta]$
      $1/(1+e^{-\eta}) = \phi$
   3. $h_\theta(x) = \mathrm{E}[y|x;\theta]$
      $h_\theta(x) = \phi$ (mean of a Bernoulli distribution)
      $h_\theta(x) = 1/(1+e^{-\eta})$ (substituting link function from 2)
      $h_\theta(x) = 1/(1+e^{-\theta^T x})$ (substituting linear function $\eta = \theta^T x$)
2. Gaussian Distribution:
   1. $y|x,\theta \sim \mathcal{N}(\mu,\sigma^2)$
   2. $g(\eta) = \mathrm{E}[T(y);\eta]$
      $\eta = \phi$
   3. $h_\theta(x) = \mathrm{E}[y|x;\theta]$
      $h_\theta(x) = \mu$ (mean of a Gaussian distribution)
      $h_\theta(x) = \eta$ (substituting link function from 2)
      $h_\theta(x) = \theta^T x$ (substituting linear function $\eta = \theta^T x$)

References

Chris Re, Andrew Ng, and Tengyu Ma (2023)
CSE229 Machine Learning
Stanford University