Generalized Linear Models

CMPUT 296: Basics of Machine Learning

Machine Learning Handbook Ch.7

Logistics

- Midterms are marked
  - Grades and feedback available on eClass
- Thought questions #3 will be marked by Thursday
- **Thought questions #4** due **one week from Thursday (Nov 26)**

Recap: Logistic Regression

- **Linear binary classification:** Learn a linear decision boundary
  - All observations on one "side" of boundary are classified as 0, all observations on the other "side" are classified as 1

$$\text{i.e., } f(\mathbf{x};\mathbf{w}) = \begin{cases} 1 & \text{if } \mathbf{w}^T\mathbf{x} > 0 \\ 0 & \text{if } \mathbf{w}^T\mathbf{x} \le 0 \end{cases}$$

- **Logistic regression:** Learn a model $p(y = 1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$
  - Logistic regression because we are learning a mapping from $\mathbf{x}$ to $p(y = 1 \mid \mathbf{x})$
  - **Logistic function** $\sigma(t) = \dfrac{1}{1 + \exp(-t)}$ forces estimate to be a valid probability (i.e., in $[0,1]$)
  - No closed-form solution for MLE; must learn numerically (e.g., SGD)
  - MLE problem is convex; local optimum is also a global optimum

Outline

1. Recap & Logistics
2. Another Linear-ish Regression Scheme
3. Natural Exponential Family Distributions
4. Generalized Linear Models

Probabilistic Approaches

We've now seen two probabilistic approaches to regression:

Linear Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = \omega^\top \mathbf{x}$
2. $p(y \mid \mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$
3. $\mu = \omega^\top \mathbf{x}$

Logistic Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = \sigma(\omega^\top \mathbf{x})$
2. $p(y \mid \mathbf{x}) = \text{Bernoulli}(\alpha)$
3. $\alpha = \sigma(\omega^\top \mathbf{x})$

Question: What do these two approaches have in common?

Example: Population Regression

Suppose we want to predict the **number of sunny days** per year in a city, given some numerical **features** about the city (latitude and longitude).

Questions:

1. Can we directly apply **linear regression** to this problem? Why?
2. Can we directly apply **logistic regression** to this problem? Why?

Exponential Transfer

- The **number of sunny days** is both **integer** and **positive**
- If we try to apply **linear regression** directly, our predictions will sometimes be **negative non-integers**
- If we try to apply **logistic regression** directly, our predictions will **always be between 0 and 1 (and non-integer)**
- What if we replaced the sigmoid function with a different function that forces the expected value to be positive?

$$f(t) = \exp(t) \implies 0 \leq f(t) < \infty$$

- We can apply f to a linear weighting of features to get a positive expected value:

$$\mathbb{E}[y \mid \mathbf{x}] = f(\mathbf{w}^\top \mathbf{x})$$

Poisson Regression

We can use the **exponential transfer function** to define a Poisson model for number of sunny days:

Poisson Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = f(\omega^\top \mathbf{x}) = \exp(\omega^\top \mathbf{x})$
2. $p(y \mid \mathbf{x}) = \text{Poisson}(\lambda)$
3. $\lambda = f(\omega^\top \mathbf{x})$

- Poisson distribution's parameter λ is both the mean and the variance
- Poisson distribution only places positive probability on integers

Poisson Regression: MLE Solution

$$p(y \mid \mathbf{x}, \omega = \mathbf{w}) = \text{Poisson}(\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} = \frac{e^{\mathbf{w}^\top \mathbf{x} y} \exp(-e^{\mathbf{w}^\top \mathbf{x}})}{y!}$$

$$\log p(y \mid \mathbf{x}, \mathbf{w}) = \log \frac{e^{\mathbf{w}^\top \mathbf{x} y} \exp(-e^{\mathbf{w}^\top \mathbf{x} y})}{y!} = \log e^{\mathbf{w}^\top \mathbf{x} y} + \log \exp\left(-e^{\mathbf{w}^\top \mathbf{x}}\right) - \log y!$$

$$= \mathbf{w}^\top \mathbf{x} y - e^{\mathbf{w}^\top \mathbf{x}} - \log y!$$

$$\mathbf{w}_{\text{MLE}} = \arg\max_{\mathbf{w}} p(\mathcal{D} \mid \mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^n p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^n \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^n \mathbf{w}^\top \mathbf{x}_i y_i - e^{\mathbf{w}^\top \mathbf{x}_i} \cancel{- \log y_i!}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^n e^{\mathbf{w}^\top \mathbf{x}_i} - \mathbf{w}^\top \mathbf{x}_i y_i$$

There is no closed-form solution to this optimization problem.

Question: How can we find $\mathbf{w}_{\text{MLE}}$?

Natural Exponential Family Distributions

The Gaussian (Normal), Bernoulli, and Poisson distributions are all examples of distributions from the (natural) exponential family

The **natural exponential family** of distributions are distributions with the form:

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

- θ is the **parameter** of the distribution
- $a : \mathbb{R} \to \mathbb{R}$ is the **log-normalizer** function
- $b : \mathbb{R} \to \mathbb{R}$ is the **base measure** function

Natural Exponential Family
Example: Poisson

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

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

$$= \exp\left(\log \lambda^y\right) \exp(-\lambda) \exp\left(\log \frac{1}{y!}\right)$$

$$= \exp\left(\log \lambda^y - \lambda + \log \frac{1}{y!}\right)$$

$$= \exp\left(y \log \lambda - \lambda - \log y!\right)$$

$$\theta = \log \lambda$$

$$a(\theta) = \exp(\theta) = \lambda$$

$$b(y) = -\log y!$$

Natural Exponential Family Example: Bernoulli

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

$$\begin{aligned} p(y \mid \alpha) &= \alpha^y (1-\alpha)^{1-y} \\ &= \frac{\alpha^y}{(1-\alpha)^{y-1}} \\ &= \frac{\alpha^y}{(1-\alpha)^y}(1-\alpha) \\ &= \left(\frac{\alpha}{1-\alpha}\right)^y (1-\alpha) \\ &= \exp\left(\log\left(\frac{\alpha}{1-\alpha}\right)^y + \log(1-\alpha)\right) \end{aligned}$$

$$\begin{aligned} &= \exp\left(y \log\frac{\alpha}{1-\alpha} + \log(1-\alpha)\right) \\ &= \exp\left(y \log\frac{\alpha}{1-\alpha} - \log\frac{1}{1-\alpha}\right) \end{aligned}$$

$$\theta = \log\frac{\alpha}{1-\alpha}$$

$$a(\theta) = \log\left(1 + \exp(\theta)\right) = \log\frac{1}{1-\alpha}$$

$$b(y) = 0$$

Natural Exponential Family

Example: Gaussian with $\sigma = 1$

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

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

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

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

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

$$\theta = \mu$$

$$a(\theta) = \frac{\theta^2}{2}$$

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

Log-Normalizer

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

The **log-normalizer** ensures that the probability density integrates to 1:

$a(\theta) = \log z$ where $z = \int_{\mathcal{Y}} \exp(\theta y + b(y))\, dy$, so

$$\int_{\mathcal{Y}} \exp(\theta y - a(\theta) + b(y))\, dy = \int_{\mathcal{Y}} \exp(\theta y - \log z + b(y))\, dy$$

$$= \frac{1}{\exp(\log z)} \int_{\mathcal{Y}} \exp(\theta y + b(y))\, dy$$

$$= \frac{1}{\int_{\mathcal{Y}} \exp(\theta y + b(y))\, dy} \int_{\mathcal{Y}} \exp(\theta y + b(y))\, dy \quad = 1$$

Properties

$$\frac{\partial a(\theta)}{\partial \theta} = \mathbb{E}[Y]$$

$$\frac{\partial^2 a(\theta)}{\partial \theta^2} = \mathbb{V}[Y]$$

Generalized Linear Models

Linear Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = \omega^\top \mathbf{x}$
2. $p(y \mid \mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$
3. $\mu = \omega^\top \mathbf{x}$

Logistic Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = \sigma(\omega^\top \mathbf{x})$
2. $p(y \mid \mathbf{x}) = \text{Bernoulli}(\alpha)$
3. $\alpha = \sigma(\omega^\top \mathbf{x})$

Poisson Regression

1. $\mathbb{E}[y \mid \mathbf{x}] = f(\omega^\top \mathbf{x}) = \exp(\omega^\top \mathbf{x})$
2. $p(y \mid \mathbf{x}) = \text{Poisson}(\lambda)$
3. $\lambda = f(\omega^\top \mathbf{x})$

- Linear, logistic, and Poisson regression are all **generalized linear models**
- A generalized linear model is a model where
  1. $\mathbb{E}[y \mid \mathbf{x}] = f(\mathbf{w}^\top \mathbf{x})$
  2. $p(y \mid \mathbf{x})$ is an **exponential family distribution**
- The transfer function f is typically the derivative of the **log-normalizer** of p
  - i.e., the transfer function and the distribution family are **chosen together**

Solving Generalized Linear Models

- **Question:** Can we analytically solve GLMs?
- GLMs are typically solved using (stochastic) gradient descent:

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

$$\log p(y \mid \theta) = \theta y - a(\theta) + b(y)$$

$$\begin{aligned}
\arg\max_{\mathbf{w}} \log p(y \mid \theta) &= \arg\max_{\mathbf{w}} \theta y - a(\theta) + b(y) \\
&= \arg\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{x} y - a(\mathbf{w}^\top \mathbf{x}) \cancel{+ b(y)} \\
&= \arg\min_{\mathbf{w}} a(\mathbf{w}^\top \mathbf{x}) - \mathbf{w}^\top \mathbf{x} y
\end{aligned}$$

Solving GLMs (2)

$$\arg\min_{\mathbf{w}} c_i(\mathbf{w}) = \arg\min_{\mathbf{w}} a(\mathbf{w}^\top \mathbf{x}_i) - \mathbf{w}^\top \mathbf{x}_i y_i$$

$$\frac{\partial c_i(\mathbf{w})}{\partial w_j} = \frac{\partial}{\partial w_j}\left(a(\mathbf{w}^\top \mathbf{x}_i) - \mathbf{w}^\top \mathbf{x}_i y_i\right)$$

$$= \frac{\partial a(\mathbf{w}^\top \mathbf{x}_i)}{\partial w_j} - \frac{\partial \mathbf{w}^\top \mathbf{x}_i y_i}{\partial w_j}$$

$$= \left(\frac{\partial a(\mathbf{w}^\top \mathbf{x}_i)}{\partial \mathbf{w}^\top \mathbf{x}_i} - \frac{\partial \mathbf{w}^\top \mathbf{x}_i y_i}{\partial \mathbf{w}^\top \mathbf{x}_i}\right)\frac{\partial \mathbf{w}^\top \mathbf{x}_i}{w_j}$$

$$= (f(\mathbf{w}^\top \mathbf{x}_i) - y_i)x_{ij}$$

So the **stochastic gradient descent** update would be:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \left(f(\mathbf{w}^\top \mathbf{x}_i) - y_i\right) \mathbf{x}_i$$

Summary

- Linear and logistic regression are both **generalized linear models**
    - So is Poisson regression
- A **generalized linear model** is a model where:
    1. $\mathbb{E}[y \mid \mathbf{x}] = f(\mathbf{w}^\top \mathbf{x})$
    2. $p(y \mid \mathbf{x})$ is an **exponential family distribution**
- An **exponential family distribution** is a distribution that can expressed as

$$p(y \mid \theta) = \exp\left(\theta y - a(\theta) + b(y)\right)$$

- The transfer function for a GLM is typically $f = \dfrac{\partial a}{\partial \theta}$