Generalized Linear Models

Carlos Valle

Departamento de Informática
Universidad Técnica Federico Santa María

cvalle@inf.utfsm.cl

November 3, 2015

Overview

Motivation

- We have seen both linear regression and logistic regression. These methods are special cases of a broader family models called Generalized Linear Models (GLMs)

The Exponential family

- The distribution of the exponential family can be written in the form:
- $$p(y;\eta) = b(y)e^{\eta^T T(y) - a(\eta)}$$
- η is called the natural or canonical parameter.
- $T(y)$ is the sufficient statistic
- $a(\eta)$ is the log partition function
- $e^{-a(\eta)}$ is a normalization constant for maintaining $p(y;\eta)$ as a distribution.
- The values of T, a and b defines a set of distributions where η is the parameter of these distributions.
- Our goal now, is to model the target distribution as a member of the exponential family.

Bernoulli Distribution

- Recall that a random variable with a Bernoulli distribution takes the value 1 with success probability of p and the value 0 with failure probability of $q = 1 - p$
- Hence, its probability distribution is given by

$$\begin{aligned} f(y,p) &= p^y(1-p)^{1-y} \\ &= e^{y \log p + (1-y)\log(1-p)} \\ &= e^{\log\left(\frac{p}{1-p}\right) y + \log(1-p)} \end{aligned}$$

- Now we choose $\eta = \log\left(\frac{p}{1-p}\right)$,
- $T(y) = y$,
- $b(y) = 1$.

Bernoulli Distribution(2)

- In order to obtain $a(\eta)$ we need to write p in terms of η:

$$
\begin{aligned}
e^{\eta} &= \frac{p}{1-p} \\
e^{\eta} - pe^{\eta} &= p \\
e^{\eta} &= p(1+e^{\eta}) \\
\frac{e^{\eta}}{1+e^{\eta}} &= p
\end{aligned}
$$

- Then, $a(\eta) = -\log(1-p) = \log(1+e^{\eta})$

Normal Distribution

- Note that in linear regression, the value of σ^2 have no effect on the minimization of $J(\beta)$.
- To simplify we use $\sigma^2 = 1$. Then, the probability distribution is expressed as

$$
\begin{aligned}
f(y;\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2} \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} e^{\mu y - \frac{1}{2}\mu^2}
\end{aligned}
$$

- Thus, we choose $\eta = \mu$, $T(y) = y$,$b(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$
- Moreover, $a(\eta) = \mu^2/2 = \eta^2/2$

How to construct GLMs?

- Our general problem is to build a model to estimate the target y using the input features $\mathbf{x}$.
- To obtain a GLM for this task we will make three assumptions:
  1. $y|\mathbf{x};\beta \sim \text{ExponentialFamily}(\eta)$. the distribution of the targets y given $\mathbf{x}$ and β follows some exponential family distribution with parameter η
  2. We want to predict $E[T(y)|\mathbf{x}]$. In the cases of linear regression and logistic regression $T(y) = y$. Therefore, we would like to find a function or hypothesis $f(x) = E[y|\mathbf{x}]$
  3. The natural parameter η and the inputs x are linearly related: $\eta = \beta^T\mathbf{x}$ (Or, if η is a vector, $\eta_j = \beta_j^T$)

Ordinary least squares (OLS)

- Let's consider that the target or response variable y is continuous and $E[y|\mathbf{x}] \sim N(\mu, \sigma^2)$
- As we showed before, $\mu = \eta$. Hence,

$$
\begin{aligned}
f_\beta(\mathbf{x}) &= E[y|\mathbf{x};\beta] \\
&= \mu \\
&= \eta \\
&= \beta^T \mathbf{x}
\end{aligned}
$$

Logistic Regression

- In logistic regression y is modeled as a Bernoulli distribution.
- $E[y|\mathbf{x};\beta] = 1 \cdot P(y=1|\mathbf{x}) + 0 \cdot P(y=0|\mathbf{x}) = P(y=1|\mathbf{x}) = p$

$$
\begin{aligned}
f_\beta(\mathbf{x}) &= E[y|\mathbf{x};\beta] \\
&= p \\
&= \frac{e^\eta}{1+e^\eta} \\
&= \frac{e^{\beta^T\mathbf{x}}}{1+e^{\beta^T\mathbf{x}}}
\end{aligned}
$$

- $g(\eta) = E[T(\eta);\eta]$ is the canonical response function
- For regression is the identify function. While for logistic regression is the sigmoid function.

Softmax Regression

- Consider a multiclass classification problem where $y \in \{1, 2, \ldots, K\}$
- We will model y as a multinomial distribution.
- Let p_k be the success probability of k-th class $k = 1, \ldots, K$
- $\mathbf{p} = (p_1, \ldots, p_K)^T$ is the probability vector.
- However, these parameter would not be independent.
- We will only use $K - 1$ parameters computing $p_K = P(y = K|x; p) = 1 - \sum_{k=1}^{K-1} p_k$

Softmax Regression (2)

- For each input vector x, we can code the targets for the K classes by using an indicator function:
- $T(y) = ((T(y))_1, \ldots, (T(y))_K)^T$, where $(T(y))_k = 1$ If $y = k$, 0 otherwise.
- Thus, $(T(y))_k = I(y = k)$ and $(T(y))_K = 1 - \sum_{k=1}^{K-1} (T(y))_k$,
- where $I(\cdot)$ is the indicator function.

Softmax Regression (3)

- The multinomial distribution for y is given by

$$
\begin{aligned}
f(y, \mathbf{p}) &= \prod_{k=1}^{K} p_k^{y_k} \\
&= \prod_{k=1}^{K} p_k^{I(y=k)} \\
&= \prod_{k=1}^{K} p_k^{(T(y))_k} \\
&= e^{\sum_{k=1}^{K} (T(y))_k \log (p_k)} \\
&= e^{(T(y))_1 \log (p_1) + \cdots + (T(y))_{k-1} \log (p_{K-1}) + \left(1 - \sum_{k=1}^{K-1} (T(y))_k\right) \log(p_K)} \\
&= e^{(T(y))_1 \log (p_1/p_K) + (T(y))_2 \log (p_2/p_K) + \cdots + (T(y))_{K-1} \log (p_{K-1}/p_K) + \log(p_K)} \\
&= b(y) e^{\eta^T T(y) - a(\eta)}
\end{aligned}
$$

Softmax Regression (4)

- Now, we choose $\eta = (\log(p_1/p_K), \ldots, \log(p_{K-1}/p_K))^T$
- $b(y) = 1$
- Define $\eta_K = \log(p_K/p_K) = 0$
- Then

$$\begin{aligned} \eta_k &= \log(p_k/p_K) \\ e^{\eta_k} &= p_k/p_K \\ p_K \eta_k &= p_k \end{aligned} \tag{1}$$

Softmax Regression (5)

- Adding all p_k we have

$$p_K \sum_{k=1}^{K} \eta_k = \sum_{k=1}^{K} p_k = 1$$

- This implies $p_K = 1/\sum_{k=1}^{K} e^{\eta_k}$, if we substitute this result in equation (1) we have

$$p_k = \frac{e^{\eta_k}}{\sum_{k=1}^{K} e^{\eta_k}}$$

- This is called the softmax.
- Finally we set $\eta_k = \beta_k^T \mathbf{x}$ and $\eta_K = 0$, implying $\beta_K^T = 0$.

Softmax Regression (6)

- Hence the conditional distribution $y|x$ is given by

$$
\begin{aligned}
p(y=k|x;\beta) &= p_k \\
&= \frac{e^{\eta_k}}{\sum_{k=1}^{K} e^{\eta_k}} \\
&= \frac{e^{\beta_k^T}}{1+\sum_{k=1}^{K-1} e^{\beta_k^T}}, k = 1, \ldots, K-1
\end{aligned}
$$

Hence, our hypothesis is

$$
\begin{aligned}
f_\beta(x) &= E[T(y)|x;\beta] \\
&= E[(I(y=1), I(y=2), \ldots, I(y=K))^T|x;\beta] \\
&= (p_1, \ldots, p_K)^T \\
&= \left(\frac{e^{\beta_1^T \mathbf{x}_m}}{1+\sum_{k=1}^{K-1} e^{\beta_k^T \mathbf{x}_m}}, \frac{e^{\beta_2^T \mathbf{x}_m}}{1+\sum_{k=1}^{K-1} e^{\beta_k^T \mathbf{x}_m}}, \ldots, \frac{e^{\beta_{K-1}^T \mathbf{x}_m}}{1+\sum_{k=1}^{K-1} e^{\beta_k^T \mathbf{x}_m}}\right)^T
\end{aligned}
$$

How to find β?

- Note that $p(y = K|x; \beta) = \frac{1}{1+\sum_{k=1}^{K-1} e^{\beta_k^T}}$
- As in binary logistic regression we would like to learn the parameter vector β which minimizes the log-likelihood:

$$J(\beta) = -\left[\sum_{m=1}^{M}\sum_{k=1}^{K} T_{mk} \log f_k(\mathbf{x}_m)\right]$$

- Where $T_{mk} = I\left(y_m = k\right)$
- $f_k(\mathbf{x}_m) = \frac{e^{\beta_k^T \mathbf{x}_m}}{1+\sum_{k=1}^{K-1} e^{\beta_k^T \mathbf{x}_m}}$
- It can be minimized by using gradient ascent or Newton-Raphson.

How to find β? (2)

- It can be shown that:

$$\nabla_{\beta_k} J(\beta) = \sum_{m=1}^{M} (f_k(\mathbf{x}_m) - T_{mk})\mathbf{x}_m$$

- And

$$\frac{\partial^2 J(\beta)}{\partial \beta_k \beta_j} = -\sum_{m=1}^{M} f_k(\mathbf{x}_m)(I_{kj} - f_j(\mathbf{x}_m))\mathbf{x}_m \mathbf{x}_m^T$$

Discriminative approaches

- So far, our classification algorithms have modeled $p(y|x;\beta)$ to build a decision boundary based on the input features.
- Recall our credit classification problem, to classify a new credit, the algorithm makes its prediction according to which side of the decision boundary falls the new instance.
- These types of algorithms are call discriminative.
- In the next chapter, we will study different approach for constructing algorithms.

Any questions?