18.650: Fundamentals of Statistics	November 17, 2023

Lecture 25 — Logistic Regression

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Regression with binary response

In this lecture, we consider a specific kind of regression in which $Y \in \{0, 1\}$. In other words, we want to predict a yes/no answer: will someone default on their credit default, will a surgery be successful, will someone get heart disease, etc. Figure 1 depicts data of this form, i.e, (X_i, Y_i) pairs where the Y_i's are binary. Since

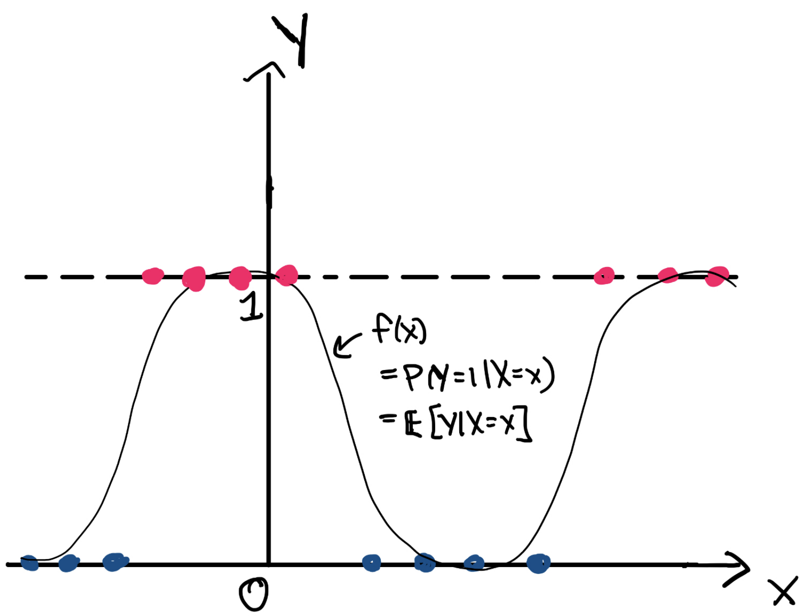

Figure 1: Depiction of (X_i, Y_i) pairs with binary response variable Y_i. The solid line is the function $f(x) = \mathbb{E}[Y|X = x]$.

$Y \mid X = x$ is Bernoulli, with a parameter p depending on x, we can write the conditional distribution as

$$Y \mid X = x \sim \text{Ber}(f(x)).$$

By definition, the regression function is

$$\mathbb{E}[Y|X = x] = f(x).$$

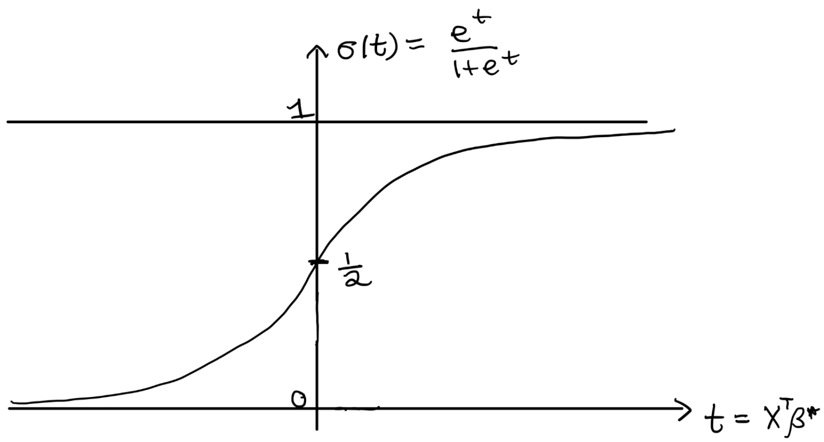

Figure 2: The sigmoid function.

This function is represented by the solid line in Figure 1. In contrast to linear regression, the function $f(x)$ cannot possibly take the form $x^T\beta^*$. This is because linear functions can go to positive and negative infinity, while in our case, $f(x)$ must lie in the unit interval since $f(x)$ denotes a probability.

This is unfortunate, because linear f's have nice interpretability probabilities — for example, the sign of the coefficient β_j^* tells us whether feature j is positively or negatively correlated with the response Y.

To retain the interpretability of linear regression while still getting values within the unit interval, we simply start with $x^T\beta^*$ and squish it into the unit interval by mapping it through a function $\sigma : \mathbb{R} \to [0,1]$. In other words, we take

$$f(x) = \sigma(x^T\beta^*),$$

where σ is a function with the following three properties:

1. σ is an increasing function
2. $\sigma(t) \to 0$ as $t \to -\infty$ and $\sigma(t) \to 1$ as $t \to +\infty$.
3. $\sigma(0) = 1/2$.

See Figure 2 for the most commonly used function with these properties: the sigmoid $\sigma(t) = \frac{e^t}{1+e^t}$.

Definition 1.1: Logistic regression model

The logistic regression model for data (X_i, Y_i), with $X_i \in \mathbb{R}^k$ and binary response $Y_i \in \{0, 1\}$ is the model

$$Y|X = x \sim \text{Ber}\left(\sigma(x^T\beta^*)\right), \quad \text{where} \quad \sigma(t) = \frac{e^t}{1+e^t}.$$

Here, $\beta^* \in \mathbb{R}^k$ is the unknown coefficient vector. The function σ is called either the sigmoid or the logistic function.

Another commonly used σ is $\sigma(t) = \Phi(t)$, where Φ is the standard Gaussian cdf. This choice of σ leads to the so-called probit regression model.

Definition 1.2: Probit regression model

The probit regression model for data (X_i, Y_i), with $X_i \in \mathbb{R}^k$ and binary response $Y_i \in \{0, 1\}$ is the model

$$Y|X = x \sim \text{Ber}\left(\Phi(x^T\beta^*)\right), \tag{1}$$

where $\beta^* \in \mathbb{R}^k$ is the unknown coefficient vector and Φ is the standard Gaussian cdf. The model (1) also has the following alternative representation:

$$Y = \mathbb{1}(X^T\beta^* + Z > 0), \quad \text{where} \quad Z \sim \mathcal{N}(0, 1). \tag{2}$$

Let's show (2) is equivalent to (1). Indeed, since $Y = \mathbb{1}(X^T\beta^* + Z > 0)$ takes value zero or one, it is by definition a Bernoulli random variable. It remains to show the parameter of the Bernoulli is precisely $\Phi(x^T\beta^*)$ when $X = x$. But indeed,

$$\mathbb{P}(x^T\beta^* + Z > 0) = \mathbb{P}(Z > -x^T\beta^*) = \mathbb{P}(Z < x^T\beta^*) = \Phi(x^T\beta^*).$$

The second equality uses the symmetry of the standard Gaussian.

Remark.

> Consider any other random variable $\tilde{Z}$ whose distribution is symmetric about zero and has cdf F. Then F can play the role of the function σ above, since it automatically satisfy the three conditions listed above. Analogously to the probit regression model, we can get $Y|X = x \sim \text{Ber}(F(x^T\beta^*))$ by taking $Y = \mathbb{1}(X^T\beta^* + \tilde{Z} > 0)$.

2 MLE for logistic regression

Logistic regression is commonly used because it has a nice (concave) log likelihood, enabling us to compute MLE efficiently. Let's now compute the log likelihood. Recall: $Y_i|X_i \sim \text{Ber}(\sigma(X_i^T\beta))$, so the pmf is

$$\mathbb{P}(Y_i \mid X_i) = \sigma(X_i^T\beta)^{Y_i}(1-\sigma(X_i^T\beta)^{1-Y_i}).$$

The log likelihood is therefore given by

$$\begin{aligned}
\ell_n(\beta) &= \sum_{i=1}^n \log\left(\sigma(X_i^T\beta)^{Y_i}(1-\sigma(X_i^T\beta))^{1-Y_i}\right) \\
&= \sum_{i=1}^n \left[Y_i \log\sigma(X_i^T\beta) + (1-Y_i)\log(1-\sigma(X_i^T\beta))\right] \\
&= \sum_{i=1}^n \left[Y_i \log\frac{\sigma(X_i^T\beta)}{1-\sigma(X_i^T\beta)} + \log(1-\sigma(X_i^T\beta))\right]
\end{aligned}$$

Since $\sigma(t) = e^t/(1+e^t)$, we have $1-\sigma(t) = 1/(1+e^t)$ and $\sigma(t)/(1-\sigma(t)) = e^t$. Using these formulas in the last line above gives

$$\ell_n(\beta) = \sum_{i=1}^n [Y_i \log e^{X_i^T\beta} - \log(1+e^{X_i^T\beta})] = \sum_{i=1}^n [Y_i X_i^T\beta - \log(1+e^{X_i^T\beta})]$$

With this simple form of the log likelihood, it is straightforward to show that $\ell_n(\beta)$ is concave. Therefore, we can find MLE with gradient ascent:

$$\begin{aligned}
&\text{Initialize } \beta^{(0)} \in \mathbb{R}^k \\
&\beta^{(j+1)} = \beta^{(j)} + \eta\nabla\ell_n(\beta^{(j)}), j = 0, 1, 2, \dots.
\end{aligned}$$

See the textbook for a description of another method to find the MLE called IRLS: iteratively reweighted least squares. However, these days gradient ascent is much more common.

3 Multiclass classification

The natural generalization of a binary response is a response $Y \in \{0, 1, \dots, M\}$, i.e. Y can take one of $M+1$ possible labels. E.g. X could be a photo, and Y could classify the photo as depicting "human", "squirrel", "landscape". The pmf of Y

given X is then

$$\mathbb{P}(Y = 0|X = x) = p_0(x)$$
$$\mathbb{P}(Y = 1|X = x) = p_1(x)$$
$$\cdots$$
$$\mathbb{P}(Y = M|X = x) = p_M(x),$$

where $\sum_{\ell=0}^{M} p_\ell(x) = 1$ for all x.

3.1 First modeling attempt

We could consider $M+1$ coefficient vectors β_j^*, $j = 0, 1, \ldots, M+1$, and assume that

$$p_j(x) = \frac{e^{x^T \beta_j^*}}{\sum_{\ell=0}^{M} e^{x^T \beta_\ell^*}}, \quad j = 0, \ldots, M. \tag{3}$$

Note that this choice satisfies the requirement $\sum_{\ell=0}^{M} p_\ell(x) = 1$. However, there are $M + 1$ unknown coefficient vectors $\beta_0^*, \beta_1^*, \ldots, \beta_M^*$. When $M = 1$ (meaning there are $M + 1 = 2$ classes i.e. binary) this model does *not* reduce to standard logistic regression. This is because the model has two unknown vectors β_0^*, β_1^*, while in logistic regression there is only one unknown vector.

The issue is that (3) does not explicitly take into account that there are M, not $M + 1$, degrees of freedom.

3.2 The correct model

To motivate how to choose the model correctly, let's go back to logistic regression. We had

$$f(x) = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}} \implies x^T \beta = \log\left(\frac{f(x)}{1 - f(x)}\right) = \log\left(\frac{p_1(x)}{p_0(x)}\right). \tag{4}$$

Here, we have simply used that the inverse of the sigmoid is $\sigma^{-1}(t) = \log(t/(1-t))$, known as the *logit* function. In the last equality, we have recognized that $f(x)$ is the probability of class 1, i.e. $p_1(x)$ and $1 - f(x)$ is the probability of class 0, i.e. $p_0(x)$.

By analogy to (4), in the multiclass setting we'll assume

$$\log\left(\frac{p_j(x)}{p_0(x)}\right) = x^T \beta_j, \quad j = 1, \ldots, M.$$

This implies

$$p_j(x) = \frac{e^{x^T \beta_j}}{1 + \sum_{\ell=1}^{M} e^{x^T \beta_\ell}}, \quad j = 1, \ldots, M$$

and

$$p_0(x) = \frac{1}{1 + \sum_{\ell=1}^{M} e^{x^T \beta_\ell}}.$$

We see that class 0 does not get a β_0! This model *does* reduce to logistic regression when $M = 1$. It is known as *multiclass logistic regression.*

The log likelihood is more complicated than for logistic regression — it is a function $\ell_n(\beta_1, \beta, \ldots, \beta_M)$ in Mk variables. It turns out to be the negative of the cross entropy loss from machine learning.