DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 7: Probability Interpretation of Logistic Models

Lecturer: Fernando Perez

In the last lecture, we looked at links between probability theory and *regression*. In this lecture, we are going to look at a probabilistic interpretation of *classification*, and specifically binary classification. We begin by recapping the setup of logistic regression, and then give a probabilistic interpretation of the cross-entropy loss as well as a probabilistic justification for using a logistic function as the model for regression problems on categorical data.

1 Binary Classification with Logistic Regression

The setup for binary classification is similar to that of linear regression, with the key difference that the quantity we are trying to predict is categorical rather than quantitative. In the previous lecture on linear regression, we assumed that we had collected data $x_1, ..., x_n$ from which we could extract features $\phi(x_1), ..., \phi(x_n)$. We then used these features to make a quantitative prediction Y using a linear combination of the features. In the classification task that we now seek to address, the prediction Y is no longer a number in $\mathbb{R}$, but is a categorical quantity like e.g. whether ($Y = 1$) or not ($Y = 0$) a picture is of a dog, whether ($Y = 1$) or not ($Y = 0$) it is going to rain today, etc. The setup for binary classification is illustrated in Figure 7.1 below.

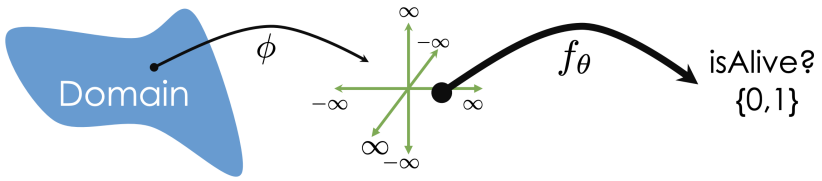


Figure 7.1: Binary classification

In Data 100, you should have seen that a common approach to binary classification is to use a logistic function as the model around a linear function of the features. This is a common sense choice because the data in classification problems is clearly not linear (as shown in Figure 7.2). Generally, the shape of the logistic function more closely matches the desired shape of the data in classification tasks.

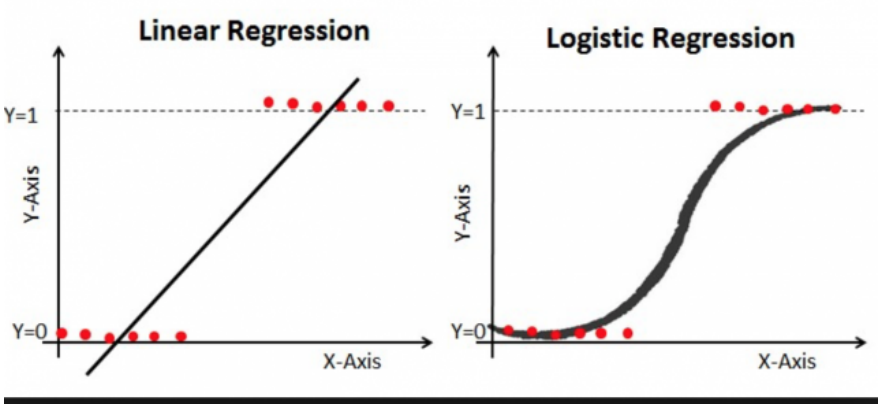


Figure 7.2: Linear Regression vs. Logistic Regression. Note that the linear model does not have the correct shape for the data.

Thus, the parametric model $f_\theta(x)$ for our data in a logistic regression problem is usually in the form given by:

$$f_\theta(x) = \sigma(\phi(x)^T\theta) = \frac{1}{1+e^{-\phi(x)^T\theta}},$$

where once again, we have allowed the features $\phi(x)$ to be d-dimensional, and the parameter vector θ is also d-dimensional.

Remark 7.1. Note that the logistic function is not a linear function in θ.

Given this model, we could proceed as we did in linear regression and find the best parameters θ^* by minimizing the mean squared loss over the data $(x_1, y_1), ..., (x_n, y_n)$:

$$arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

However, since the logistic function is non-linear and actually non-convex, the resulting optimization problem is non-convex (meaning that it may have multiple local minima). This is illustrated in Figure 7.3, where we can see there is a minimum at around 0.5, but also seemingly towards $-\infty$.

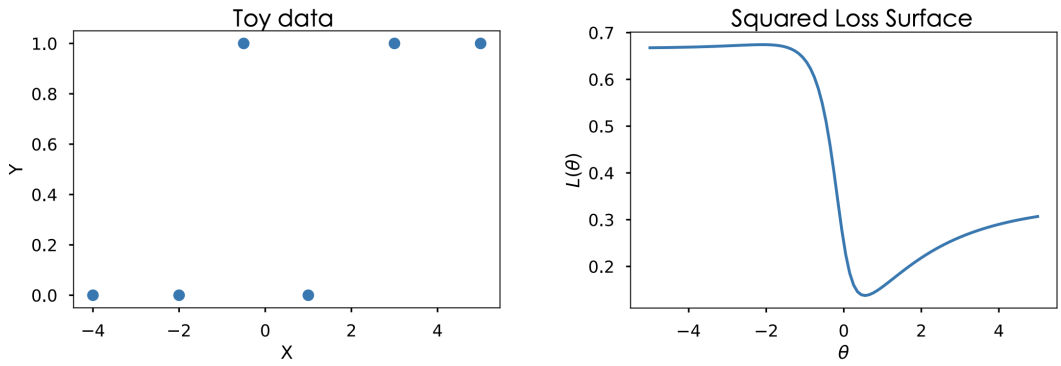


Figure 7.3: Squared loss with a logistic model

To solve this problem we often work with the *cross-entropy* loss, given by:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i^T\theta + \log(f_\theta(x_i)))$$

With this loss, which can be seen as minimizing a KL-divergence (as we will show), the problem is now convex as can be seen by Figure 7.4.

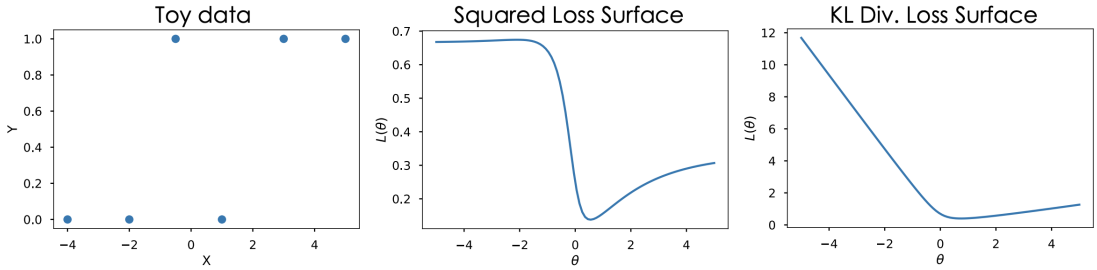


Figure 7.4: Squared loss vs Cross-entropy loss on categorical data with logistic models.

Remark 7.2. If the data is perfectly linearly separable, the optimal weights under the cross-entropy loss may be infinite. To solve this problem, we often solve a regularized version of the problem given by:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i^T\theta + \log(f_\theta(x_i))) + \lambda||\theta||_2^2.$$

where $\lambda \geq 0$ is the regularization parameter.

In the following section, we give a probabilistic interpretation of the cross-entropy loss as well as a probabilistic justification for using a logistic function as the model for regression problems on categorical data.

2 Probabilistic Interpretation of Logistic Models

To provide a probabilistic interpretation of the logistic model as well as that of the cross-entropy loss, we model our data $y_1, ..., y_n$ as being independent, and sampled from their own Bernouilli distributions. That is, $y_i \sim Bernouilli(p_i)$, for $i = 1, ..., n$, where p_i is unknown. We would like to find a maximum likelihood estimate of p_i for each y_i.

Since $y_i \in \{0, 1\}$, the likelihood of y_i is either p_i if $y_i = 1$ or $1 - p_i$ if $y_i = 0$ We can write this compactly as:

$$L(y_i; p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

The likelihood of all of the data is therefore given by:

$$L(y|p) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i},$$

where y is the $n \times 1$ vector of all y_i's and p is the $n \times 1$ vector of all p_i's for $i = 1, ..., n$. As we usually do, it is often most convenient to work with the log-likelihood of the data, ℓ, which simplifies to:

$$\ell(y|p) = \sum_{i=1}^{n}(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)),$$

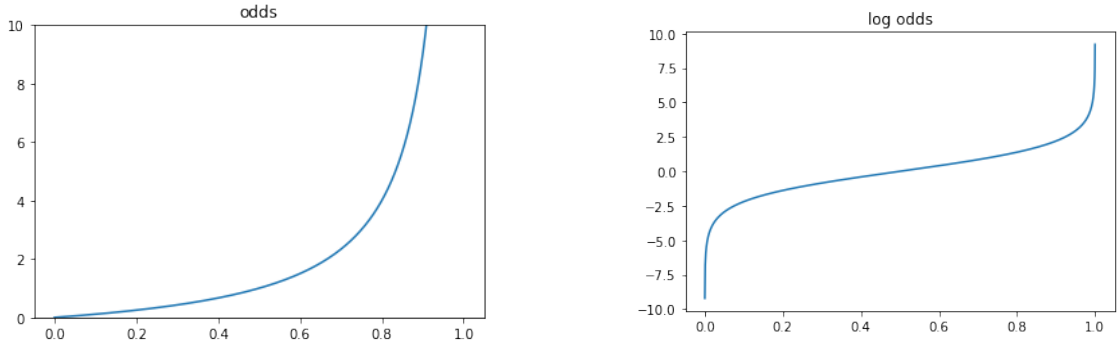
Rearranging, we can write the log-likelihood of the data as:

$$\ell(y|p) = \sum_{i=1}^{n}(y_i \log \frac{p_i}{1 - p_i} + \log(1 - p_i)),$$

Given the log-likelihood in this form, we can see that the first term is the logarithm of the "odds" ratio which is given by:

$$\text{"odds" ratio} = \frac{p_i}{1 - p_i}$$

If we plot the "odds" ratio, as can be seen in Figure 7.5a, we can see that this is not a symmetric function of p_i around 0.5. However, if we analyze the log-odds ratio, as seen in Figure 7.5b, we can see that it *is* symmetric around $p_i = 0.5$. As such this seems like the correct function to model for a problem of learning the value of p_i for different data.



(a) (b)

Figure 7.5: Odds ratio vs. Log-odds ratio

We remark that on the interval $0.1 - 0.9$, the log-odds ratio is roughly linear, so we decide to model the log-odds ratio as a linear function of our data $x_1, ..., x_n$. Solving for the value of p_i if we model the log-odds as a function of x_i recovers the logistic function we analyzed before.

$$\begin{aligned}
\log \frac{p_i}{1-p_i} &= \phi(x_i)^T\theta \\
\frac{p_i}{1-p_i} &= e^{\phi(x_i)^T\theta} \\
p_i &= \frac{e^{\phi(x_i)^T\theta}}{1+e^{\phi(x_i)^T\theta}} \\
p_i &= \frac{1}{1+e^{-\phi(x_i)^T\theta}}
\end{aligned}$$

Substituting this into the expression for the log-likelihood gives:

$$\ell(y|p) = \sum_{i=1}^{n} (y_i \phi(x_i)^T\theta + \sigma(\phi(x_i)^T\theta),$$

If we were to find the maximum likelihood value of θ, this is exactly equivalent to minimizing the cross-entropy loss of our model on the data. Therefore, we can see that logistic regression is essentially just maximum likelihood estimation where the log-odds is view as a linear function of the data.

2.1 Logistic function as the posterior probability of data coming from one of two gaussians

Another way of deriving the logistic function is by computing the posterior probability of data coming from one of two Gaussian distributions. Suppose you receive a sample x which, with

probability $1-p$ came from Gaussian $Y=0$ with mean μ_0 and variance σ^2 and with probability p came from a Gaussian $Y=1$ with mean μ_1 and variance σ^2. you compute the posterior probability that the sample came from Gaussian 1. Remember that a Gaussian distribution has density given by:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To compute the posterior probability we can use Bayes' rule and then simplify:

$$\begin{aligned} P(Y=1|x) &= \frac{P(x|Y=1)P(Y=1)}{P(x|Y=1)P(Y=1)+P(x|Y=0)P(Y=0)} \\ &= \frac{\pi e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\pi e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + (1-\pi)e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}} \end{aligned}$$

Multiplying the top and bottom by $e^{\frac{(x-\mu_0)^2}{2\sigma^2}}$ and rearraging gives:

$$\begin{aligned} P(Y=1|x) &= \frac{\pi e^{-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_0)^2}{2\sigma^2}}}{1-\pi+\pi e^{-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_0)^2}{2\sigma^2}}} \\ &= \frac{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_0)^2}{2\sigma^2}+\log\frac{\pi}{1-\pi}}}{1+e^{-\frac{(x-\mu_1)^2}{2\sigma^2}+\frac{(x-\mu_0)^2}{2\sigma^2}+\log\frac{\pi}{1-\pi}}} \\ &= \frac{1}{1+e^{\frac{(x-\mu_1)^2}{2\sigma^2}-\frac{(x-\mu_0)^2}{2\sigma^2}-\log\frac{\pi}{1-\pi}}} \\ &= \frac{1}{1+e^{\frac{(\mu_0-\mu_1)x}{\sigma^2}-\frac{(\mu_1^2-\mu_0^2)}{\sigma^2}-\log\frac{\pi}{1-\pi}}} \\ &= \frac{1}{1+e^{-\beta x-\gamma}} \end{aligned}$$

Where $\beta = \frac{(\mu_1-\mu_0)}{\sigma^2}$ and $\gamma = \frac{(\mu_0^2-\mu_1^2)}{2\sigma^2} + \log\frac{\pi}{1-\pi}$