Lecture 4: Logistic Regression

Beidi Qiang

SIUE

Introduction

The linear regression model that we discussed in STAT560 assumes that the response variable is numerical that takes values over the real line. But in many situations, we require a probability estimate (a value between 0 and 1) as output.

- Predicting probability of Heads for bent coins. You might use features like angle of bend, coin mass, etc.
- An online banking service must be able to determine the probability a transaction being performed on the site is fraudulent, on the basis of features such as the IP address, past transaction history, and so forth.

The Logistic Model

Consider a binary response Y. Using the generic 0/1 coding for the response. We want to model the relationship between $p(X) = Pr(Y = 1|X)$ and the p-dimensional features $X = (X_1, \cdots, X_p)$.

- We can't use a simple linear model: $p(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, since there is no guarantee the predicted value will be a proper probability, i.e. in between 0 and 1
- We must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X.
- We use the following logistic (or sigmoid) function:

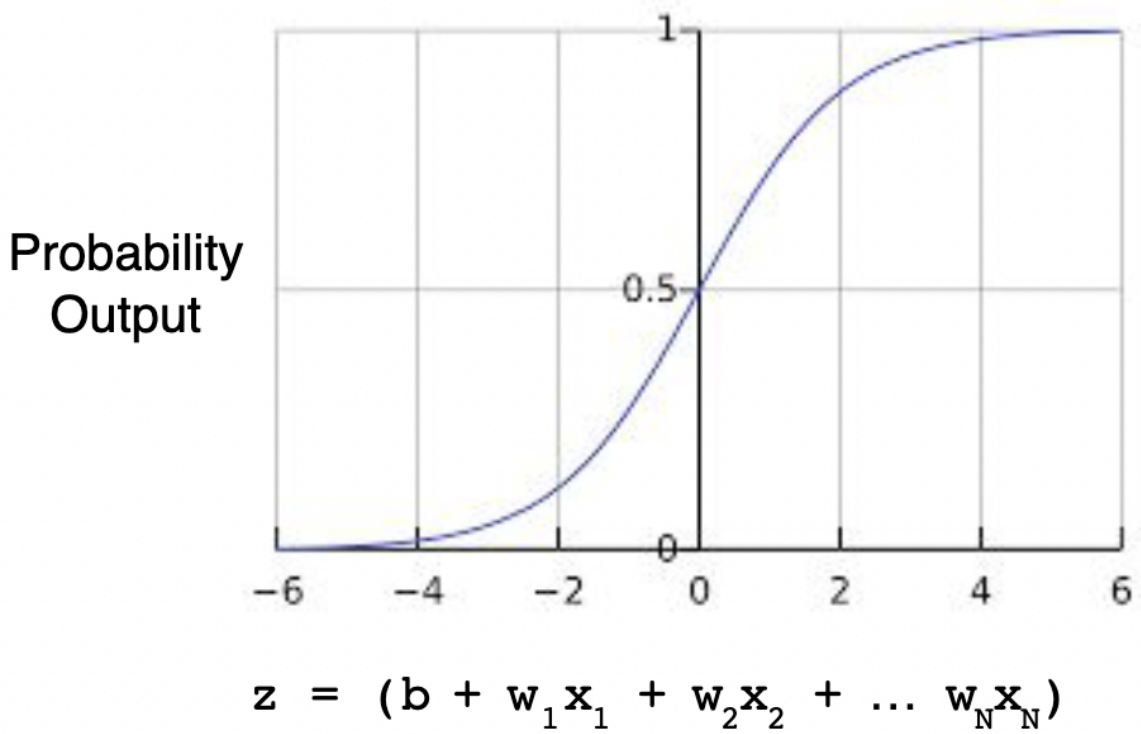
$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}.$$

The Sigmoid Function

The sigmoid function,

$$y = \frac{1}{1 + e^{-z}},$$

yields the following plot:



More on Logistic Model

- After a bit of manipulation of sigmoid function, we get

$$\log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The β values are the model's learned weights, sometimes also denoted as w.
- The quantity $p(X)/[1-p(X)]$ is called the odds, and can only be non-negative.
- Values of the odds close to 0, indicate low probabilities of $Y = 1$. Odds close to ∞, indicate high probabilities of $Y = 1$.
- The quantity $\log\left(p(X)/[1-p(X)]\right)$ is called the log-odds or logit.
- Logistic regression is just a linear regression on the log-odds.

Estimating the Model Weights

β's in the logistic regression model are unknown, and must be estimated based on the training data (labeled examples),

$$(x_1, y_1), \cdots, (x_n, y_n).$$

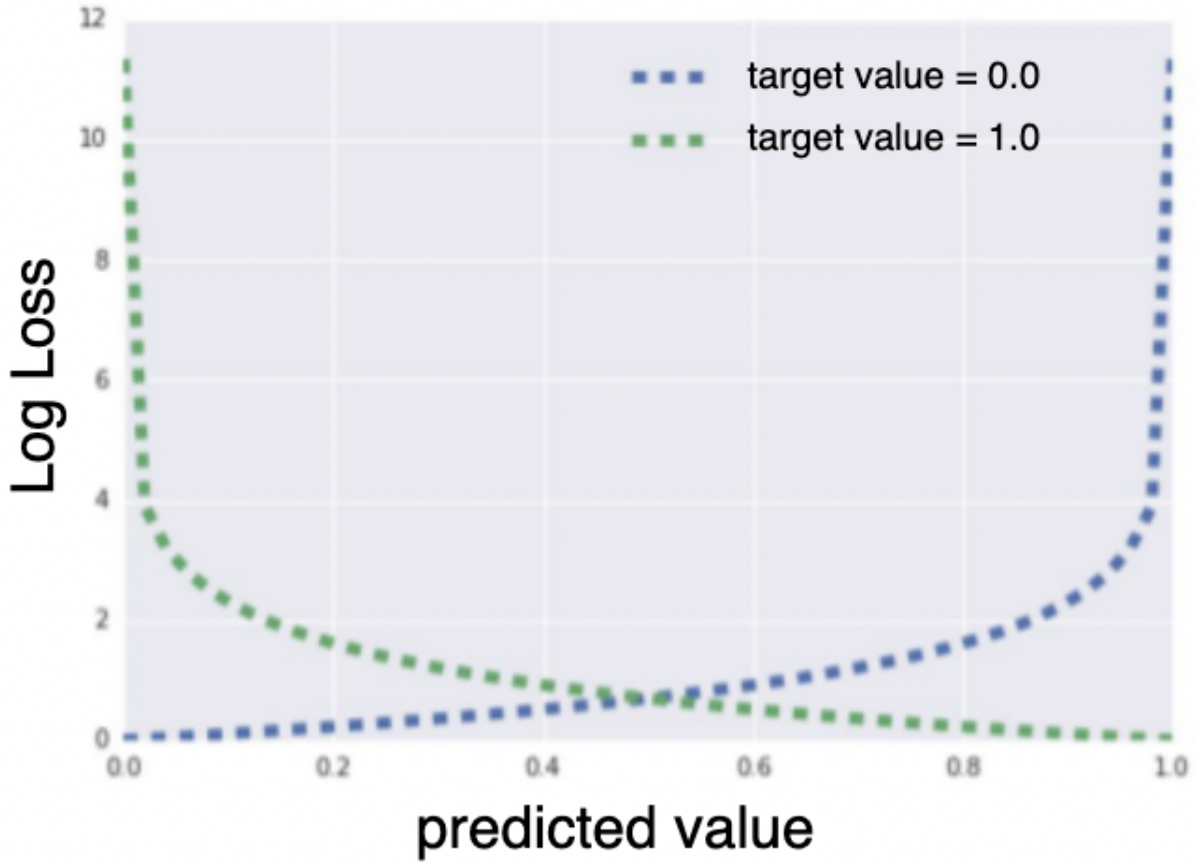
We determine β values by minimizing the loss, defined as follows:

$$logloss = \sum_{i=1}^{n} -y_i \log p(x_i) - (1 - y_i) \log(1 - p(x_i)),$$

$$\text{where } p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}}.$$

- We seek estimates of β such that the predicted probability $p(x_i)$ close to one for all individuals with $y = 1$, and close to zero for all individuals with $y = 0$.

The log loss


target value = 0.0
target value = 1.0
Log Loss
predicted value
0
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
1.0

Example: Default data

The data deals with whether an individual will default on his or her credit card payment, on the basis of annual income, monthly credit card balance and student status. Logistic model out put from R is given below:

```
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = Default)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Coefficient Interpretation

- In logistic regression, the relationship between $p(X)$ and X is not a straight line.
- Increasing X_1 by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.
- If β_1 is positive then increasing X_1 will be associated with increasing $p(X)$, and vice versa.
- The rate of change in probability of $Y = 1$ per unit change in X is not a constant value. It depends on the current value of X since the relationship is not linear.

Significance of Coefficients

Many aspects of the logistic regression output are similar to the linear regression output.

- We measure the accuracy of the coefficient estimates by computing their standard errors (details omitted).
- The z-statistic $= \hat{\beta}/SE(\hat{\beta})$ plays the same role as the t-statistic in the linear regression. A large (absolute) value of the z-statistic indicates evidence against $\beta = 0$.
- $\beta_1 = 0$ in logistic regression implies the probability of $Y = 1$ does not depend on X.
- If p-value associated with β_1 is small enough, we conclude that there is indeed an association between X_1 and probability of $Y = 1$.

Making Prediction

Once the coefficients have been estimated, it is a simple matter to compute the probability of $Y = 1$ for any given $X = x$.

$$\hat{p}(Y = 1|X = x) = \hat{p}(x) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)}}.$$

In example, for a non-student user with balance \$1000 and income \$50000, the predicted probability of default is

$$\hat{p}(default) = \frac{1}{1 + e^{-(-10.87 + 0.005737 * 1000 + 0.000003033 * 50000)}} = 0.0068.$$

Categorical Features

If X is a categorical feature, we creates dummy variables and treat it the same way as in the linear regression, for example,

$$\hat{p}(Y = 1|X = 1) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1)}},$$

$$\hat{p}(Y = 1|X = 0) = \frac{1}{1 + e^{-\hat{\beta}_0}}.$$

For example, being a student, versus a non-student the log odds of default decreases by 0.6468

Logistic Regression for Multi-level Response

- With a response of more than 2 levels (say k levels), we need to model the probability of each of $k-1$ levels.
- Set of models

$$\log\left[\frac{P(Y=1|X)}{P(Y=k|X)}\right]=\beta_1 X,\cdots,\log\left[\frac{P(Y=k-1|X)}{P(Y=k|X)}\right]=\beta_k X.$$

- This is called multinomial logit model.
- In practice multinomial logit model is not used all that often. We use discriminant analysis (next lecture) instead for multiple-class classification.