Estimating the Logistic Model

Pavlos Protopapas, Ignacio Becker

Lecture Outline

- What is classification
- Classification: Why not Linear Regression?
- Binary Response & Logistic Regression
- Estimating the Simple Logistic Model
- Classification using the Logistic Model
- Multiple Logistic Regression
- Extending the Logistic Model
- Classification Boundaries

Estimating the Simple Logistic Model

Estimation in Logistic Regression

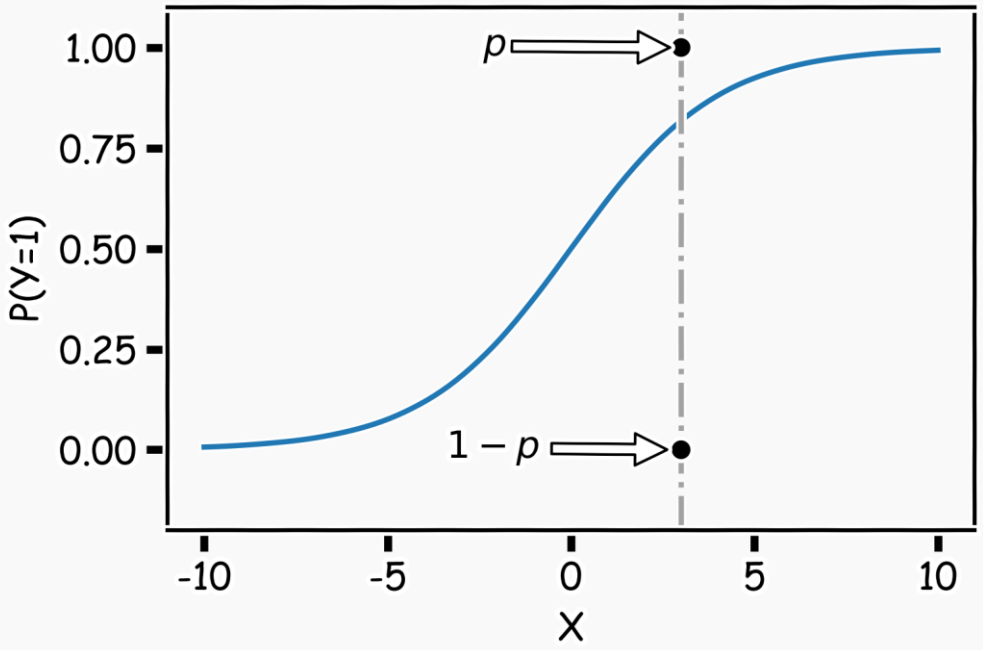
Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

Questions:

- In linear regression what loss function was used to determine the parameter estimates?
- What was the probabilistic perspective on linear regression?

Logistic Regression also has a likelihood-based approach to estimating parameter coefficients.

Estimation in Logistic Regression



Probability $Y = 1$: p

Probability $Y = 0$: $1 - p$

$$P(Y = y) = p^{y}(1 - p)^{(1-y)}$$

where:

$p = P(Y = 1|X = x)$ and therefore p depends on X.

Thus, not every p is the same for each individual measurement.

Likelihood

The likelihood of a single observation for *p* given *x and y* is:

$$L(p_i|Y_i) = P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Given the observations are independent, what is the likelihood function for *p*?

$$L(p|Y) = \prod_i P(Y_i = y_i) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i}$$

$$l(p|Y) = -\log L(p|Y) = -\sum_i y_i \log p_i + (1-y_i)\log(1-p_i)$$

Loss Function

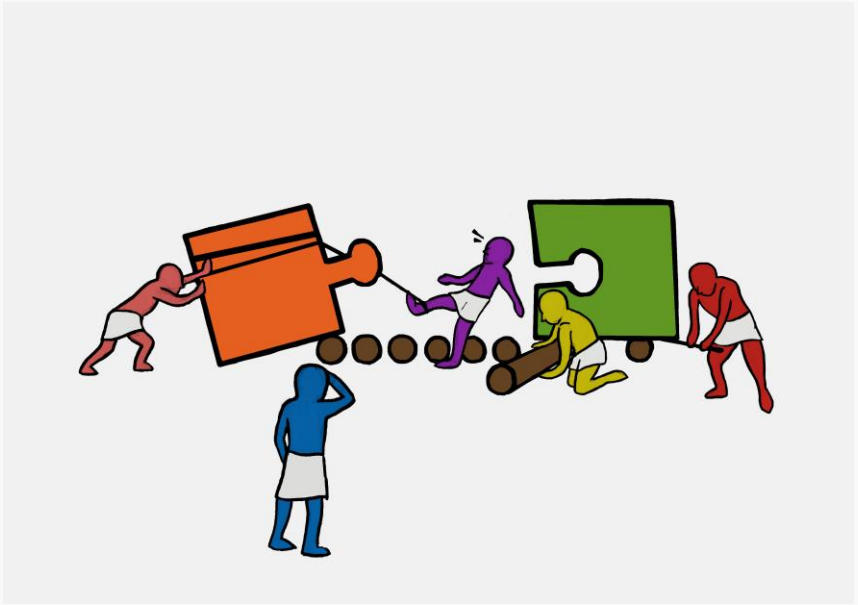
$$l(p|Y) = -\sum_i \left[y_i \log \frac{1}{1+e^{-\beta X_i}} + (1-y_i) \log \left(1 - \frac{1}{1+e^{-\beta X_i}} \right) \right]_i$$

How do we minimize this?

Differentiate, equate to zero and solve for it!

But jeeze does this look messy?! It will not necessarily have a closed form solution.

So how do we determine the parameter estimates? Through an iterative approach such as Gradient Descent.



Exercise B.1 - Simple Logistic Regression

The aim of this exercise is to try to come up with a simple logistic regression model using the `sklearn` package.

Dataset Description:

The dataset used here is called the Iris dataset. This dataset has several features such as sepal length, sepal width based on which we predict which of the iris species that particular flower belongs to.

Instructions:

1. Read the `IRIS.csv` file into a pandas dataframe.
2. Assign the predictor and response variables. Remember the aim is to predict the iris `species`
3. Standardise the predictor variable.
4. Split the dataset into train and validation sets, with 80% of the data for training
5. Fit a logistic regression model to the dataset
6. Compute and print the train and validation accuracy
7. Perform 10 fold cross-validation. Compute and print the accuracy