1 Logistic Regression

1.1 Maximum Likelihood

Suppose that there is a coin which you think is biased. You flip the coin 10 times and get 7 heads. Based on this *training data* set, what is the probability that the next flip will show heads? Can you justify your answer? To provide a rigorous argument for the answer, we can use *maximum likelihood* estimation.

Let y be the (unknown) probability that the biased coin shows heads. Since there are only two possible results, heads or tails, the experiment of flipping the coin is a *Bernoulli trial*. We introduce a *Bernoulli random variable* T such that

$$T = \begin{cases} 1 & \text{if the coin shows heads,} \\ 0 & \text{otherwise.} \end{cases}$$

with probabilities

$$P(T = 1|y) = y \quad \text{and} \quad P(T = 0|y) = 1 - y, \quad \text{respectively.}$$

The probability mass function $\mathrm{Ber}(t|y)$ of T can be written in a compact way

$$\mathrm{Ber}(t|y) = y^t(1 - y)^{1-t}, \quad t = 0, 1.$$

Suppose that we have a data set $\mathcal{D} = \{t_1, \ldots, t_N\}$ of observed values of $t \in \{0, 1\}$ from N coin flips. Let $p(\mathcal{D}|y)$ denote the probability of getting the set $\mathcal{D}$, given that y is the probability of the biased coin. Since each flip is independent, the probability $p(\mathcal{D}|y)$ is equal to the product of the probabilities of individual flip results, i.e. it is equal to

$$p(\mathcal{D}|y) = \prod_{n=1}^{N} \mathrm{Ber}(t_n|y) = \prod_{n=1}^{N} y^{t_n}(1 - y)^{1-t_n}.$$

This is a function of y; as y varies, the function $p(\mathcal{D}|y)$ produces probabilities of the set $\mathcal{D}$ determined by y. When $p(\mathcal{D}|y)$ is maximized at a value y_{ML}, we can say that y_{ML} has maximum likelihood based on the data st $\mathcal{D}$. This leads to

Task: Estimate a value y_{ML} that maximizes $p(\mathcal{D}|y)$.

This is a Calculus problem. We will have to take the derivative of the function. Before taking the derivative, we notice that the expression of $p(\mathcal{D}|y)$ is a product of exponentials. In such a situation, logarithmic differentiation can be used to ease the process of differentiation, and the function can be replace by

$$\ln p(\mathcal{D}|y) = \sum_{n=1}^{N} \{t_n \ln y + (1 - t_n) \ln(1 - y)\}.$$

Now we take the derivative of the function with respect to y, set the derivative equal to 0 to obtain a critical value, and check if the critical value y_{ML} indeed maximizes the function. The result is

$$y_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} t_n,$$

which is the *sample mean*. This justifies our guess for a biased coin in a rigorous way.

Exercises: Provide details to obtain y_{ML}.

1.2 Logistic Regression

Suppose that there is a certain test (similar to GRE) for undergraduate students who apply for a graduate program. The test score is used for admission decisions as one of the credentials that students provide. The table below shows the test scores of 10 students and the results of admission decisions:

x	:	272	331	295	287	315	266	303	294	317	309
t	:	0	1	1	0	1	0	0	0	1	1

x test score, $t = 1$ accepted, $t = 0$ rejected

There is a student whose test score is 299. Based on this data, what is the probability that the student will be accepted? If the probability is greater than 0.5, we can *classify* the student as *Likely Accepted*; if the probability is less than 0.5, as *Likely Rejected*. This type of problem is called *classification* problem in machine learning.

An answer to this problem would have a procedure to produce probability y from a test score x. A typical way to obtain a number between 0 and 1 (that is to be considered as probability) from an arbitrary real number is to use a function whose graph is an S-shaped curve, called a *sigmoid* function. One of the most common sigmoid functions is the *logistic* function

$$\sigma(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}.$$

Clearly, the logistic function converts $(-\infty, \infty)$ to $(0, 1)$. In particular, the value 0 corresponds to 0.5. In order to apply this function to
a data set $\{x_1, x_2, \ldots, x_N\}$ of test scores, we need to first adjust the values x_n in accordance with the logistic function in such a way that the result will best fit for the given data. Here we make an assumption that the adjustment is done by a linear function of the form $w_1 x_n + w_2$ so that the probability is given by

$$y_n = \sigma(w_1 x_n + w_2). \tag{1}$$

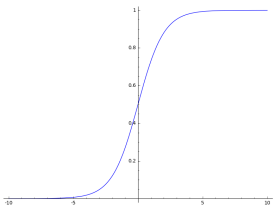

Figure 1: Graph of the Logistic Function

This is the same kind of assumption we made for linear regression.

Now the question is how to find the best values of w_0 and w_1. Let us consider the likelihood of the given data set. Each score x corresponds to probability y and a student with score x will be accepted with probability y through a Bernoulli trial. It is like flipping a biased coin with probability y. Given a data set $\mathcal{D} = \{(x_n, t_n)\}$, $t_n \in \{0, 1\}$, $n = 1, \ldots, N$, we compute the likelihood of this data set and obtain

$$p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} \mathrm{Ber}(t_n|y_n) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n},$$

where $y_n = \sigma(w_1 x_n + w_2)$ and $\boldsymbol{w} = (w_1, w_2)$. The best choice of $\boldsymbol{w} = (w_1, w_2)$ would make this likelihood the largest possible. Thus our task can be summarized as the following.

Task: Determine $\boldsymbol{w}$ that maximize the likelihood $p(\mathcal{D}|\boldsymbol{w})$.

Test scores make only one feature in credentials for admission decisions. Assume that there are k different features.
Then the n^{th} student provides credentials $\{x_{n1}, x_{n2}, \ldots, x_{nk}\}$, and as in linear regression, the credentials of N student can be arranged into a matrix

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix}.$$

The function in (1) is now generalized to

$$y_n = \sigma(w_1 x_{n1} + w_2 x_{n2} + \cdots + w_k x_{nk} + w_{k+1}),$$

which can be rewritten in vector notations

$$y_n = \sigma(\boldsymbol{x}_n \boldsymbol{w}),$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_k, w_{k+1})^\top$ is a column vector to be determined and $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nk}, 1)$. With $y_n = \sigma(\boldsymbol{x}_n \boldsymbol{w})$, the expression of the likelihood $p(\mathcal{D}|\boldsymbol{w})$ does not change.

To make differentiation easier, we take the logarithm of the likelihood and want to determine $\boldsymbol{w}$ which maximizes

$$\ln p(\mathcal{D}|\boldsymbol{w}) = \sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

or equivalently, $\boldsymbol{w}$ which minimizes

$$E(\boldsymbol{w}) := -\ln p(\mathcal{D}|\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}.$$

This problem is called *logistic regression*.

Unlike linear regression, it is difficult to find closed form formulas for solutions to this problem. Instead, we will use an inductive process to approximate the vector $\boldsymbol{w}$. We will study two standard methods for this purpose, called *Gradient Descent* and *Newton's Method*.

1.3 Applying Gradient Descent and Newton's method to the logistic regression:

$$E(\boldsymbol{w}) = -\ln p(t|\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

where $y_n = \sigma(a_n) = \sigma(\boldsymbol{x}_n^\top \boldsymbol{w})$. Note that $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

- We obtain
$$\nabla E(\boldsymbol{w}) = \left[\sum_{n=1}^{N} (y_n - t_n) x_{nj}\right] = X^\top (\boldsymbol{y} - \mathbf{t}),$$
where
$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} & 1 \\ x_{21} & \cdots & x_{2k} & 1 \\ \vdots & & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nk} & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \text{and } \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}.$$

- For Gradient Descent, we have
$$\boxed{\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \eta X^\top (\boldsymbol{y} - \mathbf{t})},$$
where R and $\boldsymbol{y}$ are determined by $\boldsymbol{w}_i$ in each step.

- We also get

$$\mathbf{H}E = \left[\sum_{n=1}^{N} y_n(1-y_n)x_{ni}x_{nj}\right] = X^\top RX,$$

where $R = \text{diag}(y_n(1-y_n))$.

- For Newton's Method,

$$\boxed{\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \eta(XRX^\top)^{-1}X(\boldsymbol{y}-\mathbf{t})},$$

where R and $\boldsymbol{y}$ are determined by $\boldsymbol{w}_i$ in each step.

1.3.1 Example: Test scores and admission to a graduate school

- Choose $k = 1$. Then $a_n = w_1x_n + w_2$. From the data, we have

$$X^\top = \begin{bmatrix} 272 & 331 & 295 & 287 & 315 & 266 & 303 & 294 & 317 & 309 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and

$$\mathbf{t} = [0,1,1,0,1,0,0,0,1,1]^\top.$$

- Choose $\boldsymbol{w}_0 = (0,0)^\top$. Then $\boldsymbol{w}_i$ converges to

$$(0.1910, -57.2937)^\top.$$

- If $x = 299$, then the probability of admission is

$$y = \sigma(0.1910 * 299 - 57.2937) \approx 0.454.$$