Logistic Regression

24 Feb 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

Recap and Today

- VIF
- Logistic Regression - More discussion
- Softmax
- Logistic Regression - Even more discussion
- Softmax - More discussion
- Implementation of Logistic Regression/Softmax in Python
- Gradient Descent

Logistic Regression - Main Ideas

Let
$w = [\alpha, \beta_1, \beta_2, ..., \beta_k]$
$x = [1, x_1, x_2, x_3, ..., x_k]$

Logistic Regression - Main Ideas

Let
$w = [\alpha, \beta_1, \beta_2, ..., \beta_k]$
$x = [1, x_1, x_2, x_3, ..., x_k]$

$$\Pr[\mathit{Yes} \text{ given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$

$$\Pr[\mathit{No} \text{ given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Logistic Regression - Main Ideas

Let
$w = [\alpha, \beta_1, \beta_2, ..., \beta_k]$
$x = [1, x_1, x_2, x_3, ..., x_k]$

$$\Pr[\mathit{Yes} \text{ given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$
$$\Pr[\mathit{No} \text{ given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Replace Yes with $+1$ and No with -1

Logistic Regression - Main Ideas

Let
$w = [\alpha, \beta_1, \beta_2, ..., \beta_k]$
$x = [1, x_1, x_2, x_3, ..., x_k]$
$y = 1$, if Yes, $y = -1$ if No

Logistic Regression - Main Ideas

Let
$w = [\alpha, \beta_1, \beta_2, ..., \beta_k]$
$x = [1, x_1, x_2, x_3, ..., x_k]$
$y = 1$, if Yes, $y = -1$ if No

$$\Pr[+1 \text{ given } x] = \frac{1}{1 + \exp(-w \cdot x)}$$
$$\Pr[-1 \text{ given } x] = \frac{1}{1 + \exp(+w \cdot x)}$$

Or more compactly

$$\Pr[y \text{ given } x] = \frac{1}{1 + \exp(-y \times w \cdot x)}$$

Maximum Likelihood Principle

Probability of observing the dataset as per the model is

$$\prod_i \Pr\left[y_i \text{ given } x_i\right] = \prod_i \frac{1}{1 + \exp\left(-y_i w \cdot x_i\right)}$$

We are looking for w which will maximize this, or alternatively w which will minimize the expression below -

$$\min_w \sum_i \log\left(1 + \exp\left(-y_i w \cdot x_i\right)\right)$$

Logistic Regression

A salesperson has visited 1000 customers to sell a book in a city. He has the data of several attributes of these customers (Income, education level, interest in reading, etc.). Finally, he has also maintained the record of who purchased the book or not.
He uses logistic regression model to build a model for this data. Assume that the cost of the book is Rs. 500, selling price of the book is Rs. 3000, average cost of visiting a customer is Rs. 200.

The salesperson gets a new list of 10000 customers and the attributes used in building the model. For a person, the probability of purchasing the book comes out as 0.6. Should the salesperson visit this customer?

$$2300 \times 0.6 + (-200) \times 0.4 = 1300$$

Logistic Regression

$$2300\, p + (1-p)(-200) > 0 \quad \Big|\Big| \quad \boxed{p > 0.08}$$

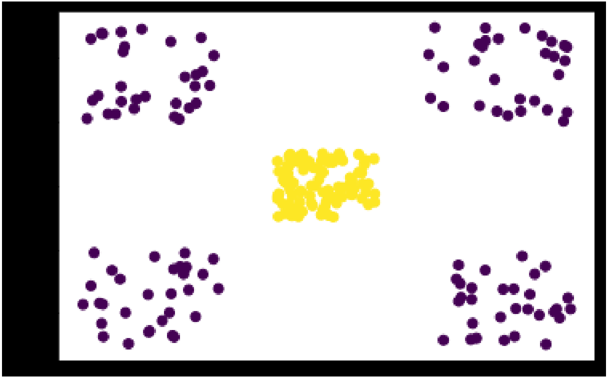
A salesperson has visited 1000 customers to sell a book in a city. He has the data of several attributes of these customers (Income, education level, interest in reading, etc.). Finally, he has also maintained the record of who purchased the book or not.
He uses logistic regression model to build a model for this data. Assume that the cost of the book is Rs. 500, selling price of the book is Rs. 3000, average cost of visiting a customer is Rs. 200.

The salesperson gets a new list of 10000 customers and the attributes used in building the model. For a person, the probability of purchasing the book comes out as 0.6. Should the salesperson visit this customer?
What if the probability comes out as 0.4?

Logistic Regression - Implementation in Python

Equivalent Cutoff condition with threshold

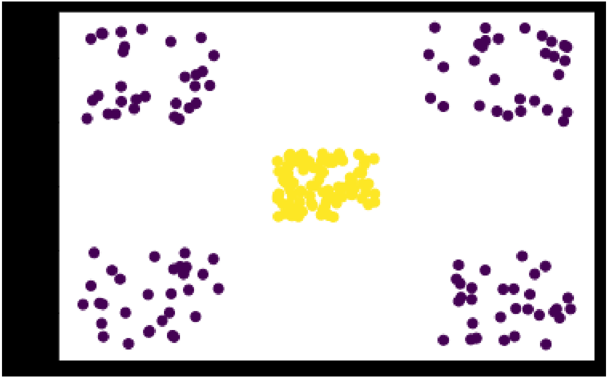
Let us say we keep the threshold as 0.5. What is the decision surface?



Decision Surface is a line (when data is 2-D, else a plane (or hyperplane))

Equivalent Cutoff condition with threshold

Let us say we keep the threshold as 0.5. What is the decision surface?



Decision Surface is a line (when data is 2-D, else a plane (or hyperplane))

For any threshold, logistic regression cannot be a good fit to this dataset because??

Multi-class classification - Model

multiple classes in the data
$y \in \{1, 2, 3, \cdots, r\}$
Instead of a single weight vector w, we consider r weight vectors $w_1, w_2, w_3, \cdots, w_r$.

$$\Pr[\text{ output } i \text{ given } x] = \frac{\exp\left(w_j \cdot x\right)}{\sum_{j=1}^{r} \exp\left(w_j \cdot x\right)}$$

Multi-class classification - Model

multiple classes in the data

$y \in \{1, 2, 3, \cdots, r\}$

Instead of a single weight vector w, we consider r weight vectors $w_1, w_2, w_3, \cdots, w_r$.

$$\Pr[\text{ output } i \text{ given } x] = \frac{\exp\left(w_j \cdot x\right)}{\sum_{j=1}^{r} \exp\left(w_j \cdot x\right)}$$

This is called the Soft-Max function, it converts a given set of numbers to probabilities.

Multi-class classification

S.No	X1	X2	Class
1	0.12	0.11	Green
2	0.14	0.44	Blue
3	0.47	0.50	Red
4	0.43	0.14	Green
5	0.37	0.40	Blue
6	0.13	0.49	Blue
7	0.41	0.13	Blue
8	0.16	0.12	Red
9	0.31	0.38	Green
10	0.22	0.29	Red

Consider a model with

$\alpha_g = 1, \beta_{g1} = 2, \beta_{g2} = 3$

$\alpha_b = 4, \beta_{b1} = -5, \beta_{b2} = 6$

$\alpha_r = 7, \beta_{r1} = 8, \beta_{r2} = -9$

What is the likelihood?

Multi-class classification

Let p_{ij} be the probability that the i^{th} data point is of type j
Let's consider the first data point for which X1 = 0.12, X2 = 0.11
As per the model, the probability of it being green can be computed as -

$$p_{1g} = \frac{e^{1+2*0.12+3*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

Similarly,

$$p_{1b} = \frac{e^{4+(-5)*0.12+6*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

And,

$$p_{1r} = \frac{e^{7+8*0.12+(-9)*0.11}}{e^{1+2*0.12+3*0.11} + e^{4+(-5)*0.12+6*0.11} + e^{7+8*0.12+(-9)*0.11}}$$

Example - Contd.

A similar exercise can be performed on all the data points. We get the following result.

S.No	Green	Blue	Red	Observed Class
1	0.004265	0.051441	0.944294	Green
2	0.029431	0.830513	0.140056	Blue
3	0.047325	0.158705	0.79397	Red
4	0.001005	0.001515	0.99748	Green
5	0.027343	0.136793	0.835864	Blue
6	0.025891	0.910426	0.063683	Blue
7	0.001005	0.001691	0.997304	Blue
8	0.003846	0.036124	0.96003	Red
9	0.028342	0.203231	0.768427	Green
10	0.0173	0.177809	0.804891	Red

Thus, the likelihood of observing the data given the model is -

$p_{1g}.p_{2b}.p_{3r}.p_{4g}.p_{5b}.p_{6b}.p_{7b}.p_{8r}.p_{9g}.p_{10r}$

Softmax for 2 classes

$$P(X = Yes) = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

$$P(X = Yes) = \frac{e^{\alpha_y + \beta_{1y} x_1 + \beta_{2y} x_2 + \ldots + \beta_{ky} x_k}}{\underline{\qquad\qquad} + e^{\alpha_n + \beta_{1n} x_1 + \ldots + \beta_{kn} x_k}}$$

$$P(X = No) = \frac{e^{\alpha_n + \beta_{1n} x_1 + \ldots + \beta_{kn} x_k}}{\underline{\;\checkmark\;} + e^{\alpha_y + \beta_{1y} x_1 + \ldots + \beta_{ky} x_k}}$$

Gradient Descent

Thank you for your attention