Machine Learning Engineer Course
Day 11

- Logistic Regression -

DIVE INTO CODE

Thursday, May 20 2021

DIOP Mouhamed

Agenda

1. Check-in
2. How to proceed
3. Quick Review
4. Logistic Regression
5. Logistic Regression Class of Scikit-learn
6. Assignment
7. Logistic Regression – Sample Code
8. Check-out

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What do you want to know the most right now?
(Anything is fine.)

Today's Objective

How to solve problems
“Scratch Logistic Regression”

[Problem1] Hypothetical function
[Problem2] Steepest descent
[Problem3] Estimattion
[Problem4] Objective function
[Problem5] Learning and estimation
[Problem 6] Plot of learning curve
[Problem 7] Visualization of decision area
[Problem 8] (Advance assignment) Saving weights

Today's Objective

Understanding Statistical Models

- Understand logistic regression through scratching

Learn the difference between classification problems and non-classification problems

- Learn the basics of classification problems

Quick Review (Scratch Linear Regression)

Setting up a linear regression problem!

(1) Formulate an equation (hypothetical function) to derive the predicted value

(2) Calculate the error between the target variable and the predicted value.

(3) Set up a problem to minimize this error and formulate an equation (objective function).

(4)Find the optimal solution of the objective function in an exploratory manner (steepest descent method).

(5)When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

What is logistic regression?

Logistic regression is a type of classification model that follows a Bernoulli / binomial distribution (1).

The output of logistic regression is the "probability of an event occurring," and this output is used to classify the classes.

This probability is derived by "transforming by a function" (2) a linear combination that regresses to the mean, hence the name "regression," even though logistic regression is used for the task of classification problems.

Specifically, the output of logistic regression is the result of converting the output of linear regression into the normal probability of binomial distribution by passing it through the sigmoid function to be introduced later.
Therefore, logistic regression is no different from linear regression in that it attempts to explain one variable by the other variable (with weights applied to each factor).

What is logistic regression?

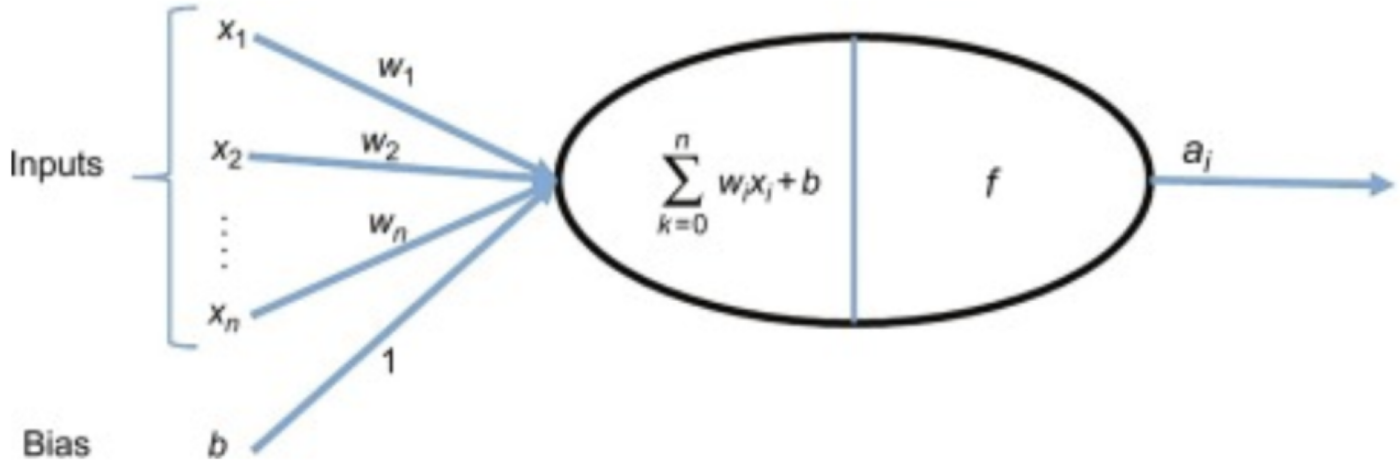
(1) A Binomial Distribution is a mutually independent Bernoulli trial (a trial in which only two outcomes are possible, e.g., "If I throw a dice, is it 1 or not? For example, if you want to know "If I throw a dice, will it be 1 or something else? is a Bernoulli trial, but thinking about "which number will come up when I throw the dice? It is a discrete probability distribution that expresses how many times a certain event occurs (the number of successes) when "n" trials are independently conducted (which is not Bernoulli trial). It has the number of trials and the success probability as parameters. When n is sufficiently large, the binomial distribution approaches a normal distribution with mean np and variance np(1-p), where p is the success probability. The binomial distribution with one trial is equal to the Bernoulli distribution.
https://to-kei.net/distribution/binomial-distribution/#i

(2) Use the sigmoid function to relate the probability to the linear combination (θ0 + θ1x). In general, when the objective variable does not follow a normal distribution, the assumed function is not μ= θ0 + θ1x (μ: mean value), but G (link function) such as G(μ) = θ0 + θ1x, which can be used to relate the model to the data. Such a model is called a generalized linear model (GLM).

What is logistic regression?

On the topic of logistic regression

Logistic regression has the same structure as a neural network without hidden layers.
Since logistic regression has the same structure as neural networks without hidden layers, understanding logistic regression is a bridge to deep learning.



What is logistic regression?

In the logistic regression, the following assumptions are made.

①The explanatory variable (x) is a continuous or discrete value, and the objective variable (y) is a discrete value.

②The prediction ($\hat{y}$) follows a Bernoulli distribution / binomial distribution (mean np, variance np(1-p)).

How to learn logistic regression?

Know the problem setup for logistic regression

(1) Formulate an equation (hypothetical function) that derives the predicted value using the sigmoid function

(2) Formulate an equation (likelihood function) that maximizes the simultaneous probability

(3) Re-set the problem to be minimized and formulate the equation (objective function).

(4) Find the optimal solution of the objective function in an exploratory manner (steepest descent method)

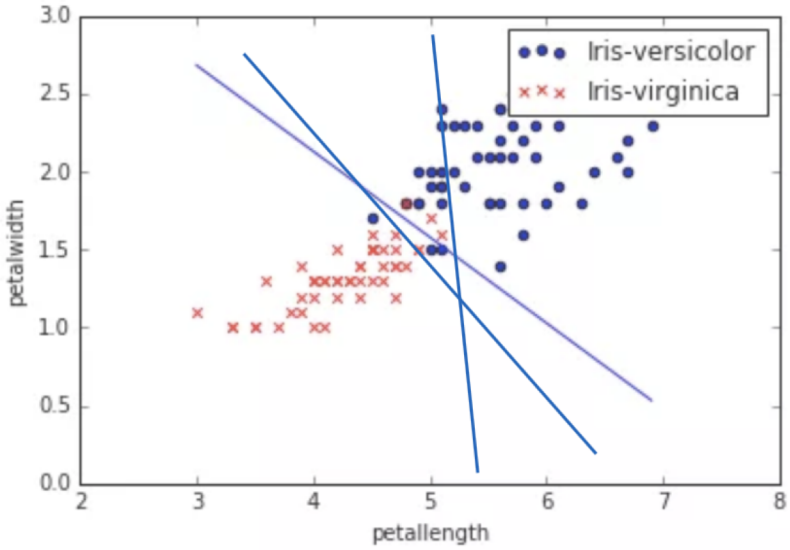
(5) When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

How to learn logistic regression?

Iris data

Let's say we have an Iris data set. The data points are pre-colored by class. Let's choose a feature X1 (petal length) and a feature X2 (petal width) and plot the relationship between the two variables.

It would be nice if we could draw a decision boundary to classify the Iris-versicolor and Iris-virginica classes, but how do we draw the decision boundary in the first place? Let's try to use the straight line (linear combination) of linear regression.



Sigmoid function

Using the function sigmoid in the equation below, we can convert the output of the linear combination to a probability

sigmoid function: $\sigma(z) = \frac{1}{1+\exp(-z)}$

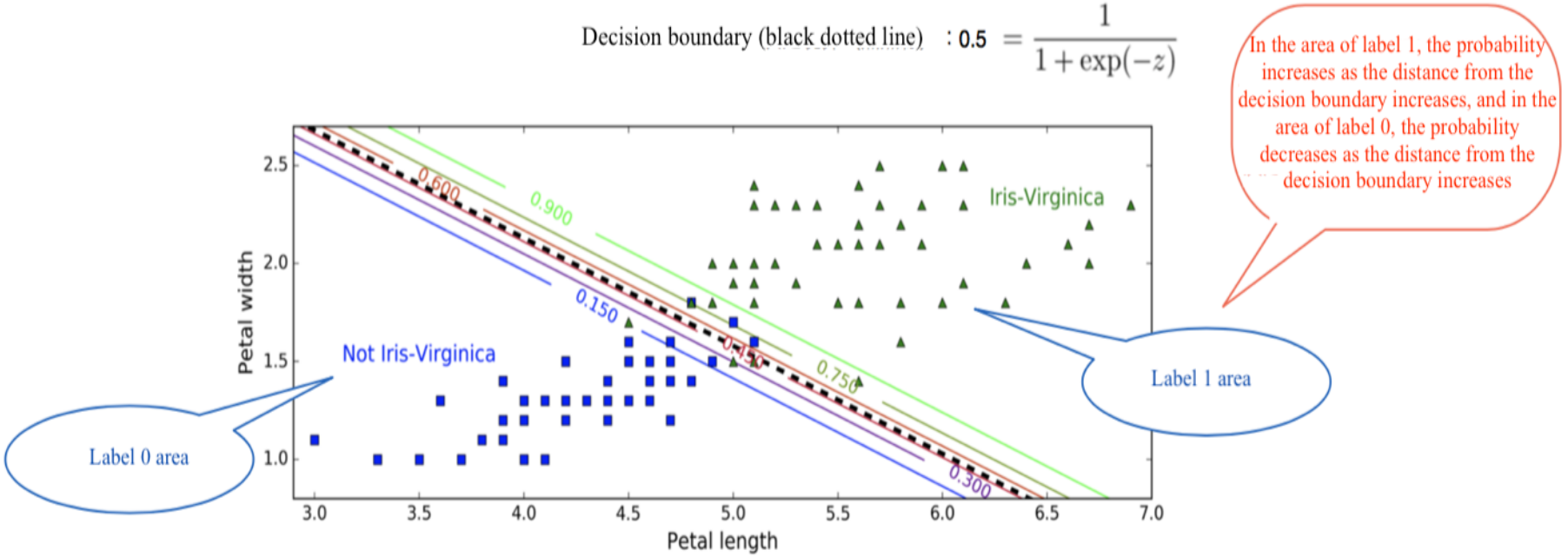
A function that has a threshold (0.5) and returns the probability of a particular classification.

Substituting the linear combination (θ0 + θ1xi) for z in the equation returns the probability of belonging to a certain class. In other words, it returns the probability that the i-th xi is in class 1.
Then, the point where the probability becomes 0.5 can be regarded as the decision boundary. Rather than drawing a straight line to separate them, it can be considered that a straight line is drawn at the 0.5 point as a result of assigning a probability to the coordinates of the region with 0.5 as the threshold.

Sigmoid function

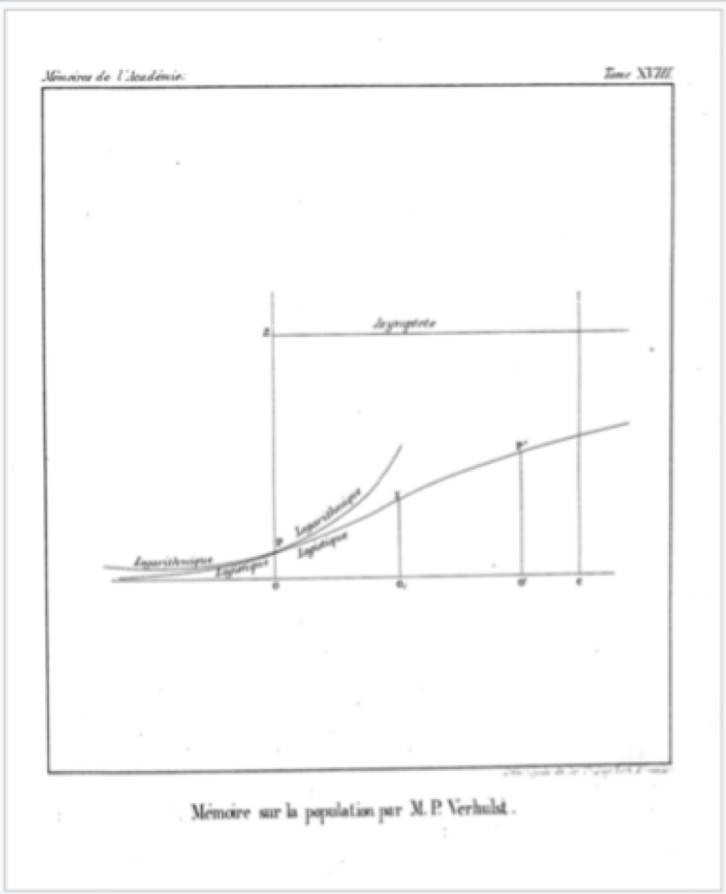
sigmoid function: $\sigma(z) = \dfrac{1}{1+\exp(-z)}$

Decision boundary (black dotted line) : 0.5 $= \dfrac{1}{1+\exp(-z)}$



Sigmoid function

What is the sigmoid function?



Original image of a logistic curve, contrasted with a logarithmic curve

Also known as the logistic sigmoid function.
It was originally developed to model the growth of a population.
It explains that the initial stage of growth is almost exponential (geometric series), then when saturation begins, growth slows down to linear (arithmetic series), and when the population matures, growth stops.

※Wikipedia 「Logistic function」
https://en.wikipedia.org/wiki/Logistic_function#In_ecology:_modeling_population_growth

Sigmoid function

Exponential and sigmoid functions

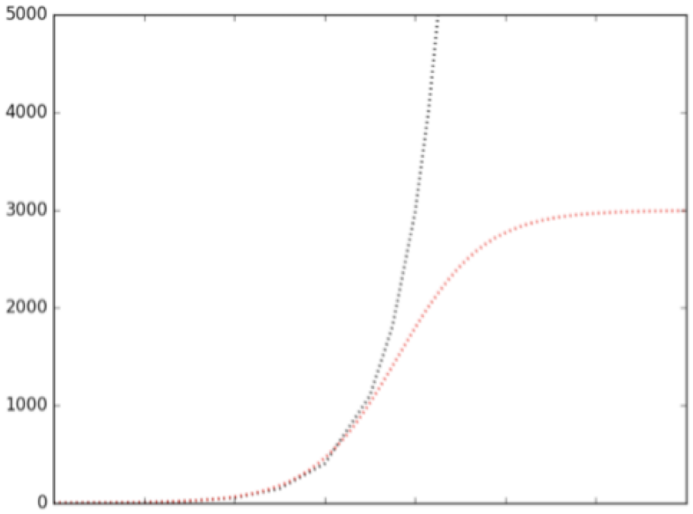


Figure: Exponentiation curve (black) and sigmoid curve (red)

The solution to the logistic equation in the continuous reproduction model of an organism shows exponential growth in the absence of resource constraints and sigmoidal growth in the presence of resource constraints.

The solution to the logistic equation in the continuous breeding model of the organism shows an exponential increase when there is no resource constraint and a sigmoidal increase when there is a resource constraint.
※ note From "Why the Singularity Can't Be Done".

It seems that the sigmoid function can be viewed as an exponential curve up to a certain point, but then stabilizes or decreases at a certain point.

http://skeptics.hatenadiary.jp/entry/2017/06/27/232828

How to learn logistic regression?

Know the problem setup for logistic regression

(1) Formulate an equation (hypothetical function) that derives the predicted value using the sigmoid function

(2) Formulate an equation (likelihood function) that maximizes the simultaneous probability

(3) Re-set the problem to be minimized and formulate the equation (objective function).

(4) Find the optimal solution of the objective function in an exploratory manner (steepest descent method)

(5) When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

Hypothetical function

The assumed function this time

Output of the sigmoid function with the linear combination input:

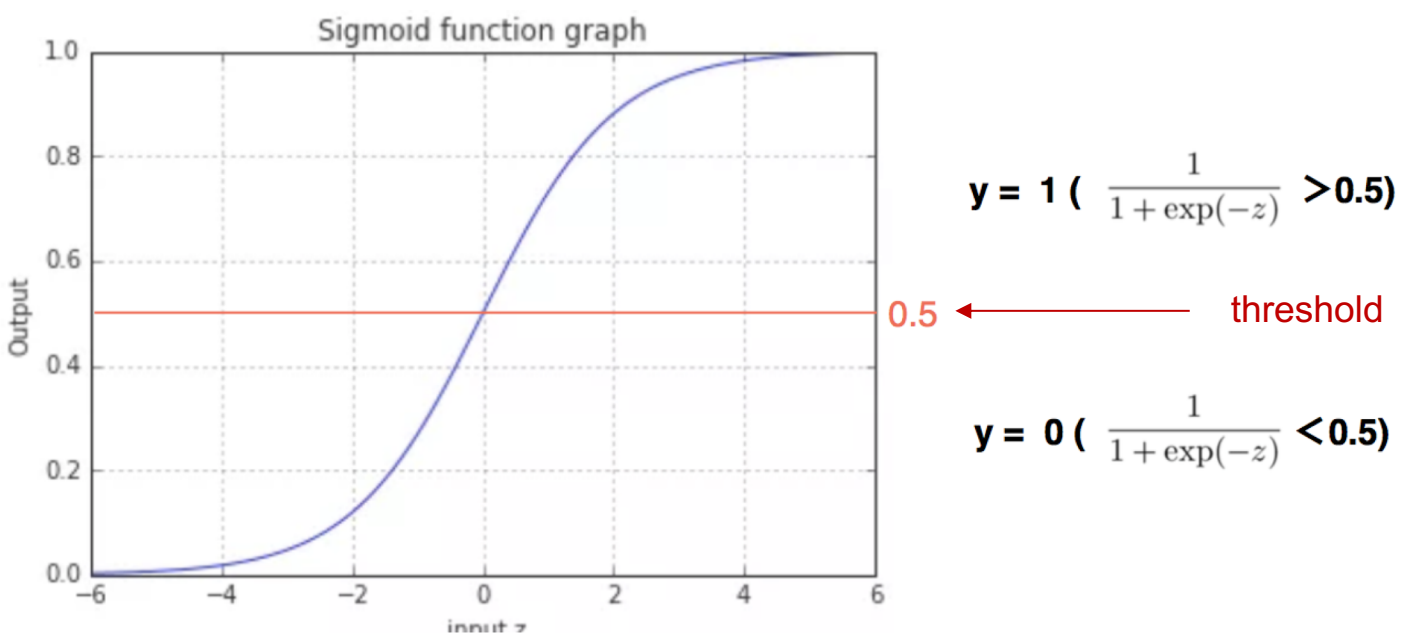
z: linear combination (θ0 + θ1x) (assumed function for linear regression)

$$\hat{y} = \sigma(z) = \frac{1}{1+\exp(-z)}$$

Hypothetical function

The assumed function this time

No matter what value (θ0 + θ1x) takes, its output will be within the output range 0 ≤ y ≤ 1 of sigmoid.
Thanks to this property, we can predict the correct label = {0, 1} for binary classification. This is because, when one class is represented by y = 1 and the other class by y = 0, the output probability can be expressed as
This is because when one class is represented by y = 1 and the other by y = 0, the output probability can be regarded as "the probability that class 1 occurs.



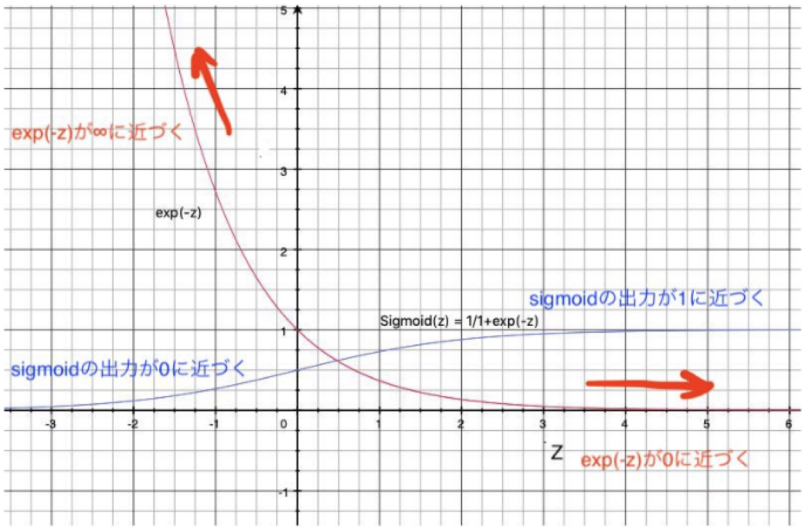
Hypothetical function

Relationship between input and output values of sigmoid function

Let's see how the output result from the sigmoid changes when the value of the input linear combination is increased or decreased.

When input(z):(θ0 +θ1x)>0
exp(-(θ0 + θ1x)) approaches 0
Output: sigmoid(θ0 + θ1x)
→ 1 (approaching 1)

When input(z):(θ0 + θ1x) < 0
exp(-(θ0 + θ1x)) approaches ∞.
Output: sigmoid(θ0 + θ1x)
→ 0 (close to 0)



Hypothetical function

Relationship between sigmoid function and linear combination

The function that relates the predicted value $\hat{y}$ to the linear combination is called logit.

The relationship can be expressed as follows.

$$\hat{y} = \frac{1}{1 + \exp(-\theta^T x)}$$

$$\text{logit}(\hat{y}) = \theta^T x$$

$$= \log \frac{\hat{y}}{1 - \hat{y}}$$

$$\log \frac{\hat{y}}{1 - \hat{y}} = \theta^T x$$

$$\Leftrightarrow \quad \frac{\hat{y}}{1 - \hat{y}} = \exp(\theta^T x)$$

$$\Leftrightarrow \quad \frac{1 - \hat{y}}{\hat{y}} = \frac{1}{\exp(\theta^T x)}$$

$$\Leftrightarrow \quad \frac{1}{\hat{y}} - 1 = \frac{1}{\exp(\theta^T \cdot x)}$$

$$\Leftrightarrow \quad \frac{1}{\hat{y}} - 1 = \frac{1}{\exp(\theta^T x)}$$

$$\Leftrightarrow \quad \frac{1}{\hat{y}} = \frac{1 + \exp(\theta^T x)}{\exp(\theta^T x)}$$

$$\Leftrightarrow \quad \hat{y} = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)}$$

$$\Leftrightarrow \quad \hat{y} = \frac{1}{1 + \exp(-\theta^T x)}$$

Hypothetical function

By transforming the equation between the logit function and the linear combination, we can derive the seemingly complex relationship between the sigmoid function and the linear combination.
The logit function is also called the log odds, and these log odds evaluate the linear combination.

$$\frac{y}{1-y}$$: Probability that an event will occur / Probability that an event will not occur

When y = 0, y / 1-y becomes 0 and the log odds are minus ∞

When y = 0.5, the log odds are 0.

How to learn logistic regression?

Know the problem setup for logistic regression

(1) Formulate an equation (hypothetical function) that derives the predicted value using the sigmoid function

(2) Formulate an equation (likelihood function) that maximizes the simultaneous probability

(3) Re-set the problem to be minimized and formulate the equation (objective function).

(4) Find the optimal solution of the objective function in an exploratory manner (steepest descent method)

(5) When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

Likelihood function

Now, if the probability that yi = 1 is pi, and the probability that yi = 0 is 1 - pi, we can write the expression for likelihood (1) as follows

$$l_i = p_i^{\,y_i}\left(1-p_i\right)^{1-y_i}$$

Replacing (1) in Eq. (1) with (2), which is a linear combination of the sigmoid function, we obtain the following equation

We want to find the simultaneous probability by considering each of the obtained N (number of samples) realizations (li) as an independent probability. This can be expressed as a product.

$$l_i = \sigma\left(\theta^T \mathbf{x}_i\right)^{y_i}\left(1-\sigma\left(\theta^T \mathbf{x}_i\right)\right)^{1-y_i} \qquad L = l_1 \times l_2 \times \cdots \times l_i = \prod_{i=1}^{N} l_i$$

The function representing the simultaneous probability is called the likelihood function.

(1) The likelihood is the probability (density) multiplied by the number of samples, but what it means is "the probability of the event", but it is not, it indicates "how well the probability distribution by the assumed parameters fits the observed data (y) by people".

(1)→ $p_i^{\,y_i}$ **(2)→** $\sigma\left(\theta^T \mathbf{x}_i\right)^{y_i}$ (C：クラス) $P\left(C = y_i \mid \mathbf{x}_i; \theta\right)$

Likelihood function

The problem of maximizing the likelihood (maximum likelihood estimation method) → The problem of minimizing the loss

$$
\begin{aligned}
L(\theta) &= -\prod_{i=1}^{N} P\left(C = y_i \middle| \mathbf{x}_i; \theta\right) \\
\log L(\theta) &= -\sum_{i=1}^{N} \log P\left(C = y_i \middle| \mathbf{x}_i; \theta\right) \\
&= -\sum_{i=1}^{N} \log \sigma\left(\theta^T \mathbf{x}_i\right)^{y_i} \left(1 - \sigma\left(\theta^T \mathbf{x}_i\right)\right)^{1-y_i} \\
&= -\sum_{i=1}^{N} y_i \log \sigma\left(\theta^T \mathbf{x}_i\right) + (1 - y_i) \log\left(1 - \sigma\left(\theta^T \mathbf{x}_i\right)\right)
\end{aligned}
$$

The conditional probability and log-likelihood functions introduced so far are shown in the left equation. What is different from the previous equations is that there is a minus at the beginning. The conditional probability and the log likelihood function are intended to be maximized, but here they are reversed by adding a minus sign, making it a minimization problem. The last expression derived by transforming the equation is the objective function of this study.

※See below for formulas for the logarithm of the product and logarithm of the power used in the formula transformation. https://sci-pursuit.com/math/logarithm-formulae-and-calculation.html

How to learn logistic regression?

Know the problem setup for logistic regression

(1) Formulate an equation (hypothetical function) that derives the predicted value using the sigmoid function

(2) Formulate an equation (likelihood function) that maximizes the simultaneous probability

(3) Re-set the problem to be minimized and formulate the equation (objective function).

(4) Find the optimal solution of the objective function in an exploratory manner (steepest descent method)

(5) When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

Objective function

The objective function this time

Cross entropy loss function + regularization term

The objective function derived earlier from the equation transformation of the log-likelihood function with minus, divided by the number of samples, is the objective function this time. This expression is called the cross-entropy loss function. In machine learning, the cross-entropy loss function is often used as the objective function when determining which of the k classes is applicable.

Divide the error by the number of samples

Regularization term

$$h_\theta(x) = \theta^T x$$

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2N}\sum_{j=1}^{m}\theta_j^2$$

Objective function

What do you want to achieve with this objective function?

We can learn to make the difference between the true probability distribution y and the predicted probability distribution log smaller, that is, we can learn to make the shape of the predicted probability distribution closer to y.

$$h_\theta(x) = \theta^T x$$

① Returns 0 if y = 0　　② Returns 0 if y = 1

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2N}\sum_{j=1}^{m}\theta_j^2$$

Objective function

Check the meaning of the equation for the cross entropy loss function. If the class y of the data x in the first part (①) is 0, then the calculation of the entire (①) is 0, and only the second part (②) remains. Conversely, if the class y of the data x in ② is 1, the calculation of the entire ② is 0, and ① remains. The goal is to find a z that is small enough as a whole by adding minus values for either ① or ②.

$$h_\theta(x) = \theta^T x$$

① Returns 0 if y = 0　　② Returns 0 if y = 1

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2N}\sum_{j=1}^{m}\theta_j^2$$

On regularization terms

What is a regularization term?
Cross Entropy Loss Function + Regularization Term

Regularization is a technique that introduces
additional information to avoid overfitting the dataset
and to reduce generalization error.

In the field of machine learning, it is called "weight decay," because the weights that are judged to be not necessary for fitting the model gradually decay and become closer to zero. The regularization term can also be viewed as a "norm-based penalty," but different types of norms are used for different purposes.
The two most common types of regularization in machine learning are L1 regularization (p=1), which uses the L1 norm, and L2 regularization (p=2), which uses the L2 norm.
The L1 regularization is characterized by feature selection by setting weights smaller than a in the regularization parameter to 0, and setting the other weights closer to 0 by a. The L2 regularization is a system where weights are set closer to 0 according to their relative importance, and the larger the weight, the more it is controlled and leveled overall. In both cases, the regularization parameter is a positive constant, and the larger it is, the more effective the regularization becomes, but if it is too large, there will be under-fitting.

(1) http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.642.3159&rep=rep1&type=pdf

On regularization terms

When p = 1, the sum of the absolute values of weights makes the diamond-shaped region a constraint on the objective function

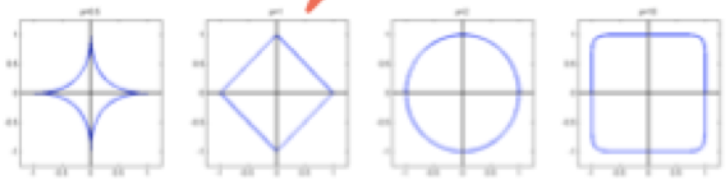
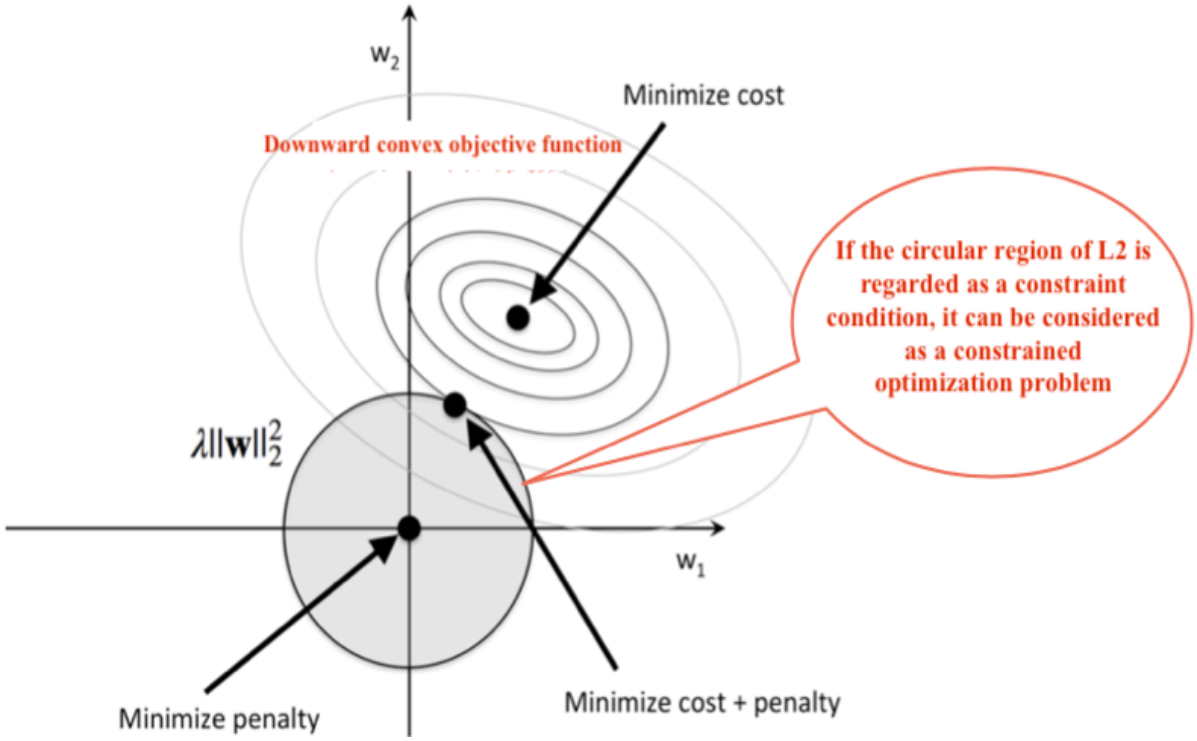


Figure 1: Unit circles for several Minkowski-p-norms $\|\mathbf{x}\|_p$: from left to right $p = 0.5$, $p = 1$ (Manhatten), $p = 2$ (Euclidean), $p = 10$.

On regularization terms

What does it mean to add a regularization term?
Why do we use a regularization term in the objective function of logistic regression?



http://robonchu.hatenablog.com/entry/2017/10/15/112724

How to learn logistic regression?

Know the problem setup for logistic regression

(1) Formulate an equation (hypothetical function) that derives the predicted value using the sigmoid function

(2) Formulate an equation (likelihood function) that maximizes the simultaneous probability

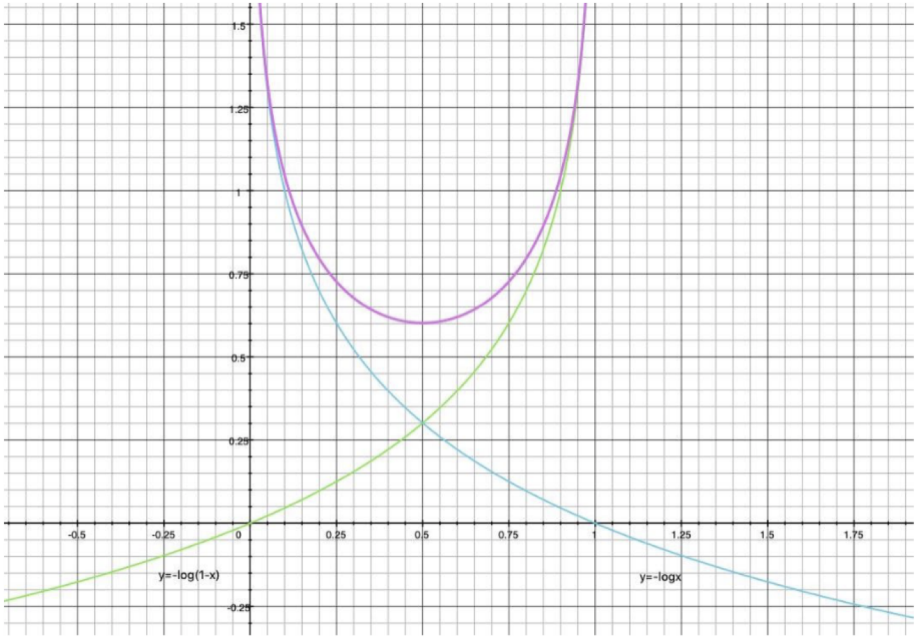
(3) Re-set the problem to be minimized and formulate the equation (objective function).

(4) Find the optimal solution of the objective function in an exploratory manner (steepest descent method)

(5) When the optimal solution is reached, the optimal parameters of the assumed function are obtained.

Least-squares method

Can the optimal solution be found by the steepest descent method?



The objective function (cross entropy loss function) for which the optimum solution should be searched this time consists of two graphs.

$$-y^{(i)}log(h_\theta(x^{(i)}))$$ (Blue graph)

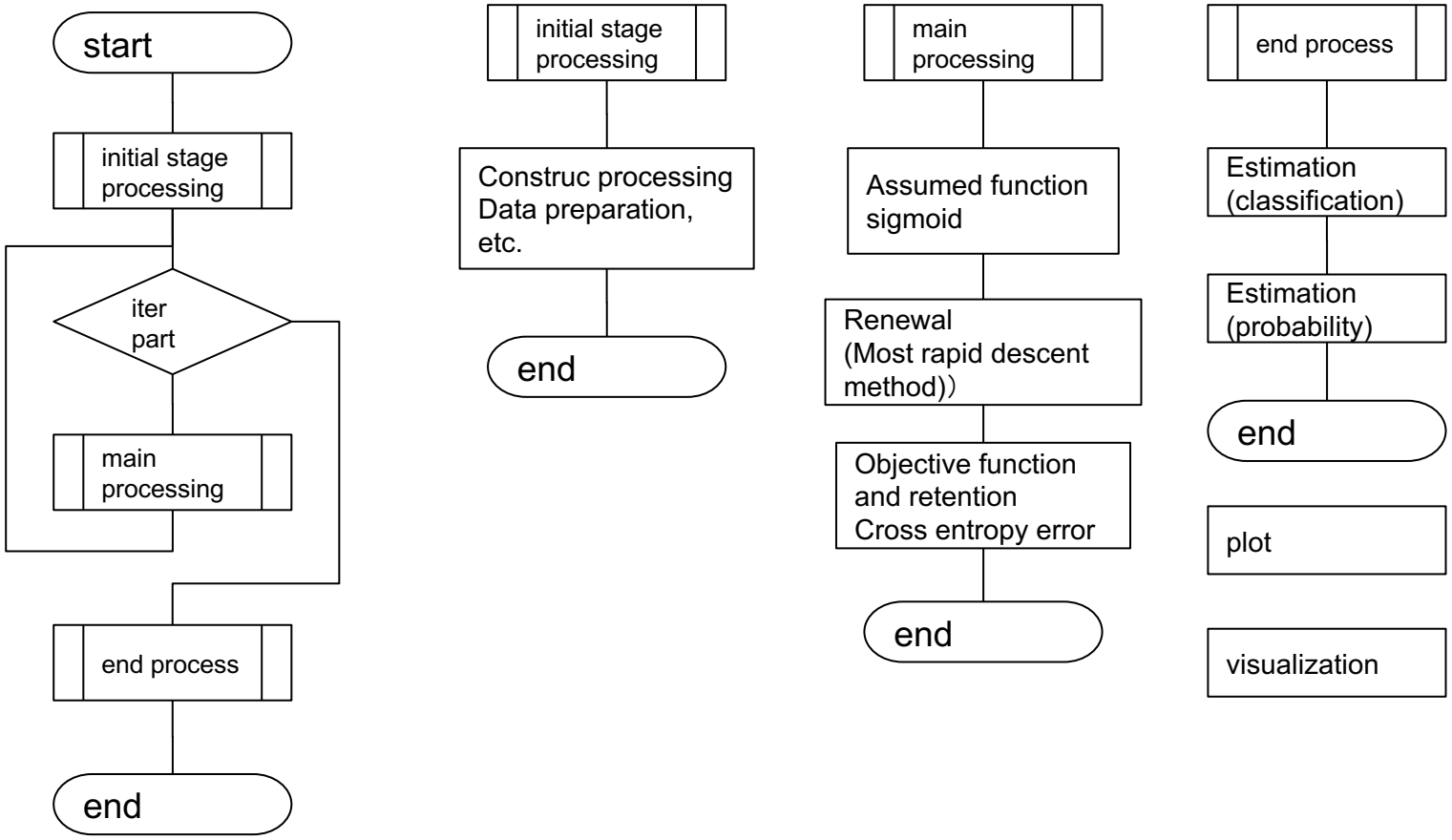
$$-(1-y^{(i)})log(1-h_\theta(x^{(i)}))$$ (Green graph)

As shown in the figure on the left, if you add the convex functions of the blue graph and the green graph, you get a downward convex function, that is, a pink graph.

With this objective function, the optimal solution can be found exploratory by the steepest descent method, and the parameters of the hypothetical function can be updated.

How to learn logistic regression?

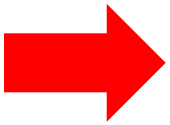
Functions required for scratching logistic regression



Logistic Regression of scikit-learn

Let’s first have a look at the one used until now with the help of the scikit-learn library.

Scikit-learn’s LogisticRegression Class



Let’s see a sample code and get inspirations from it.

Sprint 4 – Scratch Logistic Regression

Explanation about this Sprint is given but please try it on your own first.

Sprint 4 – Scratch Logistic Regression

Please work on your own after class and submit your assignments on DIVER.

Sprint 4 – Sample Code

A Sample Code of this Sprint is given but please try it on your own.

[Sprint 4 – Scratch Logistic Regression]

Please work on your own after class and submit your assignments on DIVER.

ToDo by next class

Next class will be Zoom : Thursday 27 May 2021

🚨ToDo: Scratch SVM
https://diver.diveintocode.jp/curriculums/1647

Check-out

3 minutes Please post the following point to Zoom chat.

Q. Current feelings and reflections
(joy, anger, sorrow, anticipation, nervousness, etc.)

Thank You For Your Attention