Machine Learning Engineer Course
Day 12

- Support Vector Machines (SVM) -

DIVE INTO CODE

Thursday May 27, 2021

DIOP Mouhamed

Agenda

1. Check-in
2. How to proceed
3. Quick Review
4. Support Vector Machines (SVM)
5. SVM module of Scikit-learn
6. Assignment
7. SVM– Sample Code
8. Check-out

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What is your main purpose in this course (your goal) ?

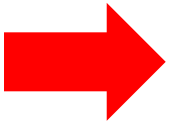
(Anything is fine.)

How to proceed - Objective

What is the purpose?

1. Understanding Statistical Models
Understanding SVM through Scratch

2. Getting familiar with linear models and different methods



Let’s learn the basics of SVM

Quick Review (Scratch 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.

Targeted audience for this assignment

①Those who can write code for learning and estimation using scikit-learn's classification models.

②Know the calculation process of classification models (assumption function, objective function, steepest descent method) (1)

(1) Those who have solved sprint logistic regression scratch

What is SVM?

Support Vector Machines are a type of learning machine that uses kernels to extend linear discriminant machines into the nonlinear domain. As such they can be used to discriminate, or tell the difference, between two classes of data. In laymans terms, they draw a line or plane between the two sets of data and whatever is on one side is of class A and whatever is on the other side is on class B.

SVM – The Flow

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

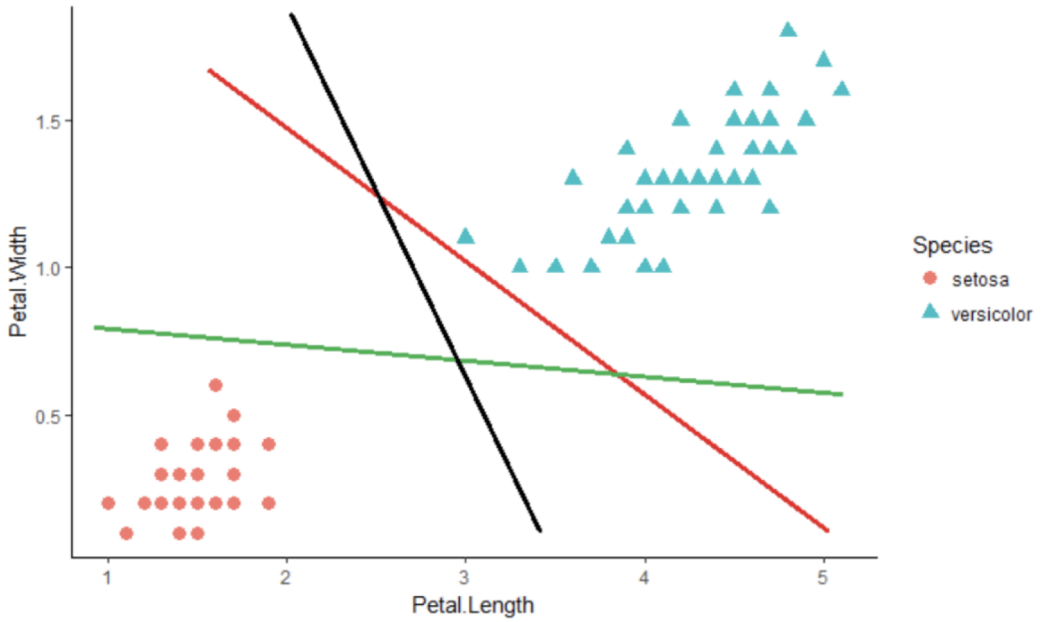
④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

What is SVM?

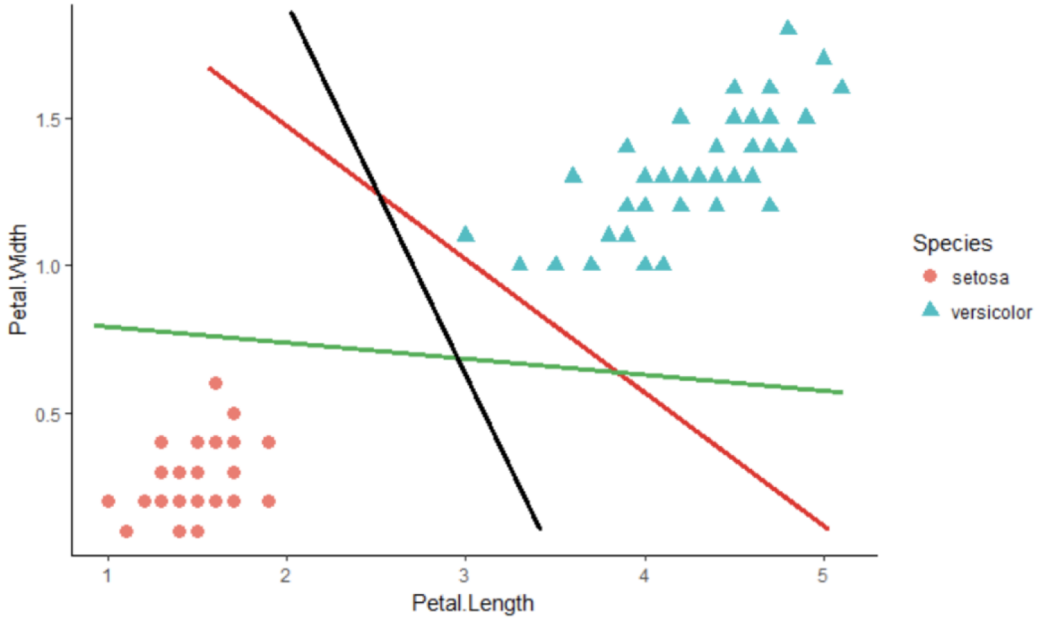
Iris Data

Let’s say you have an iris dataset here. The data points are color-coded for each class in advance. let’s select a feature X1 (petal length) and a feature X2 (petal width) and plot the relationship between the two variables.



What is SVM?

It would be nice to be able to draw an identification boundary (boundary line) for classifying the Iris-setosa and iris-versicolor classes this time as well, but SVM doesn't output the probability of belonging to the class for all data points like logistic regression. Then, what kind of mechanism will draw the identification boundary ?



Assumption function

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

Assumption function

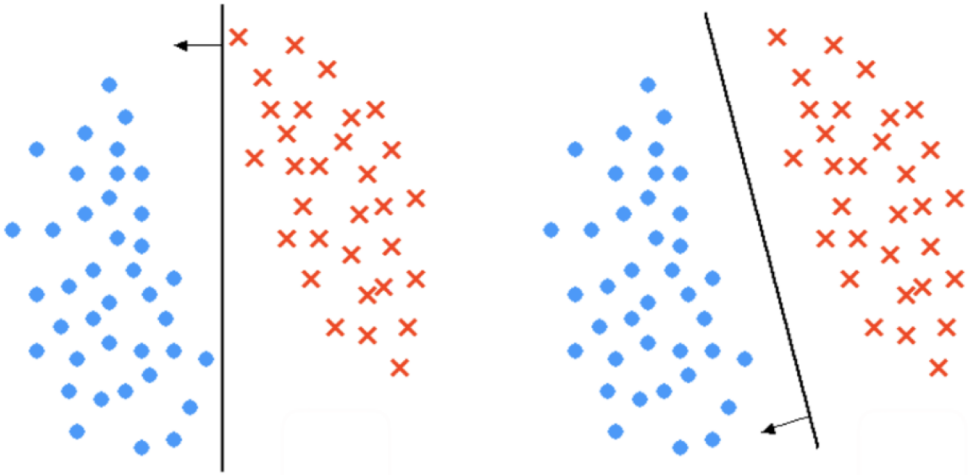
Hypothetical function

This time as well, the following hypothetical function is assumed in order to draw the decision boundary.

Output obtained by inserting a linear function into the sign function:

Linear combination : $w^T x$

$$y = \mathrm{sign}(w^T x)$$



$$y \in \{+1(\bullet), -1(\times)\}$$

Assumption function

sign function returns output similar to the following:

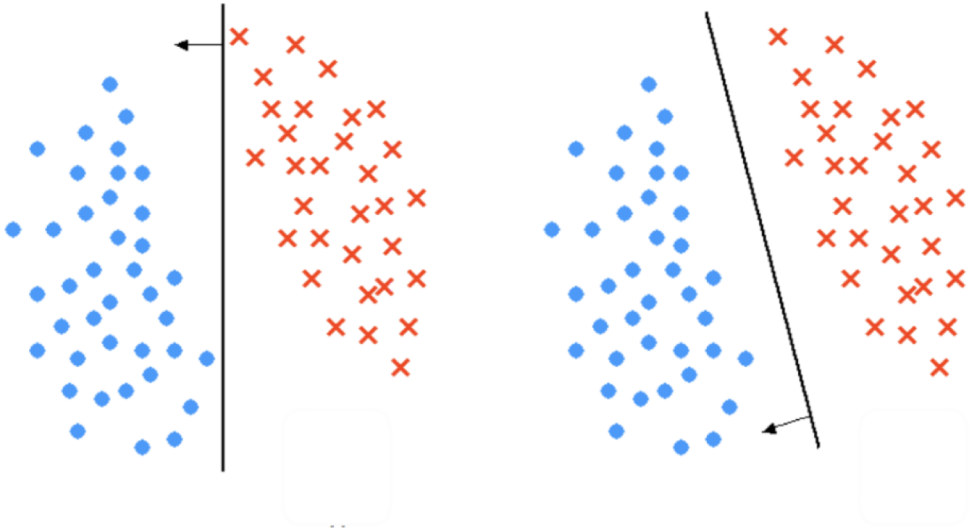
$$\text{sign}(w^T x) = \begin{cases} +1 & \text{if } (w^T x) \geqq 0 \\ -1 & \text{otherwise} \end{cases}$$

since y ∈ {-1, 1}, the conditions can be collectively written as

$$y\,(w^T x) \geqq 0$$

(w x) is positive when y = 1, so it holds when y= -1, (w x) is negative, so it holds

*** It is possible to estimate with this hypothetical function only after solving the dual problem later and obtaining a sparse solution (support vector)**

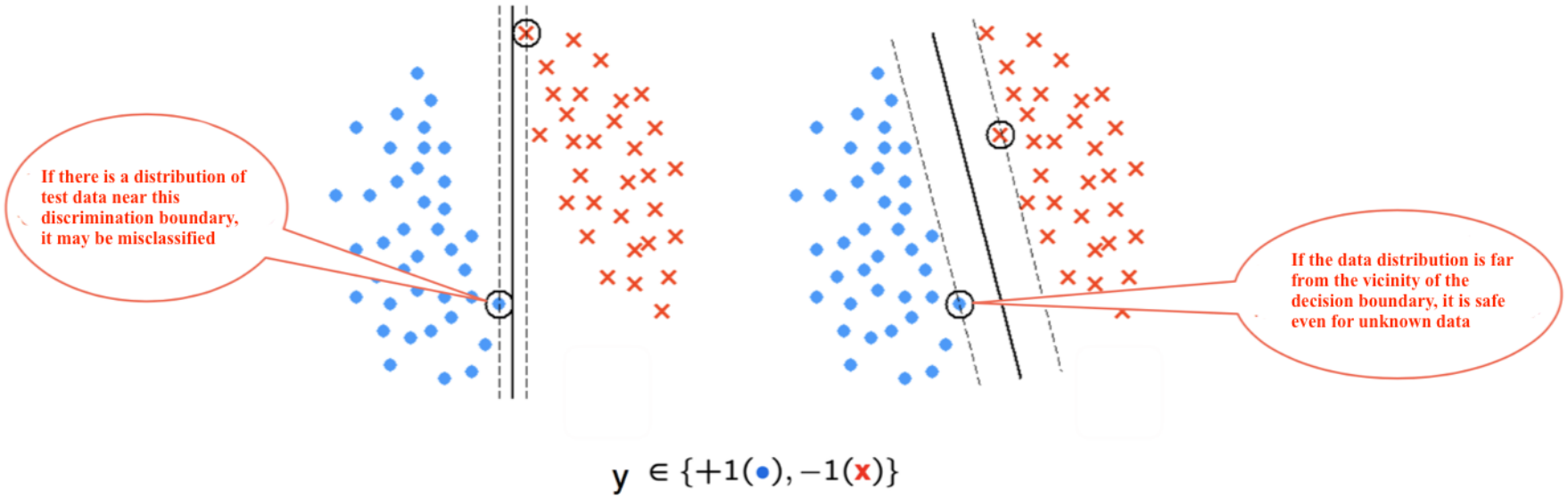


$$y \in \{+1(\bullet), -1(\times)\}$$

Assumption function

Distance from the decision boundary

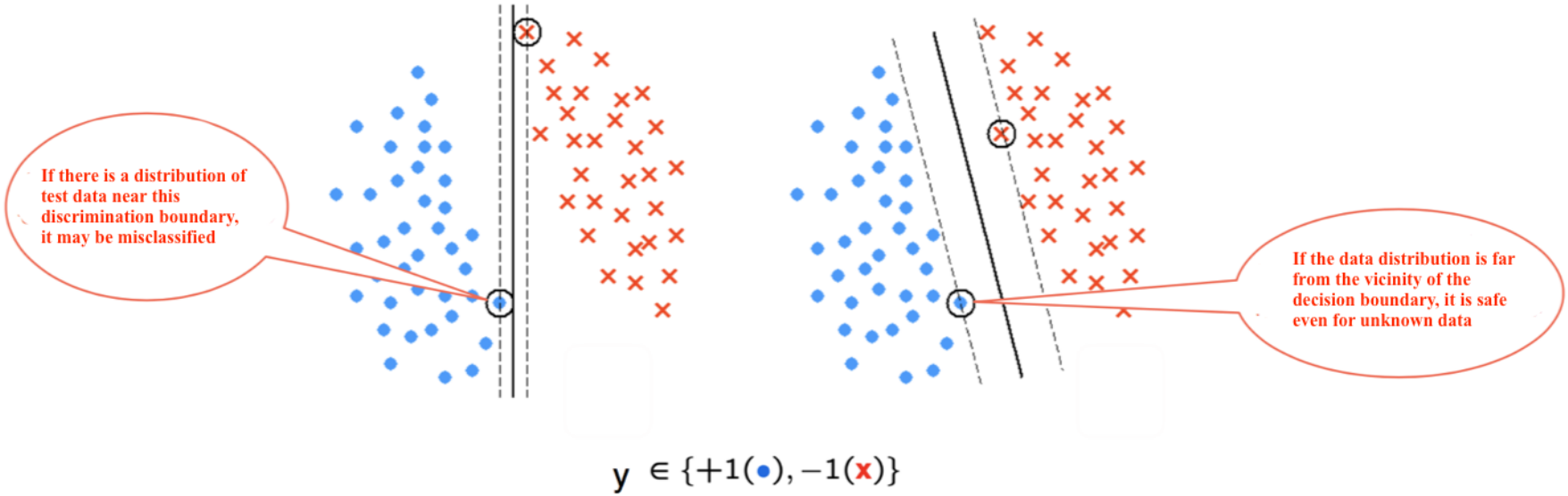
Let's start with the assumption that we use the hypothetical function above and divide it into 1 and -1 class regions. So how should we control the parameters and position the identification boundaries ? The answer to this questions is to find the decision boundary that maximizes the distance (margin) to the closest sample.



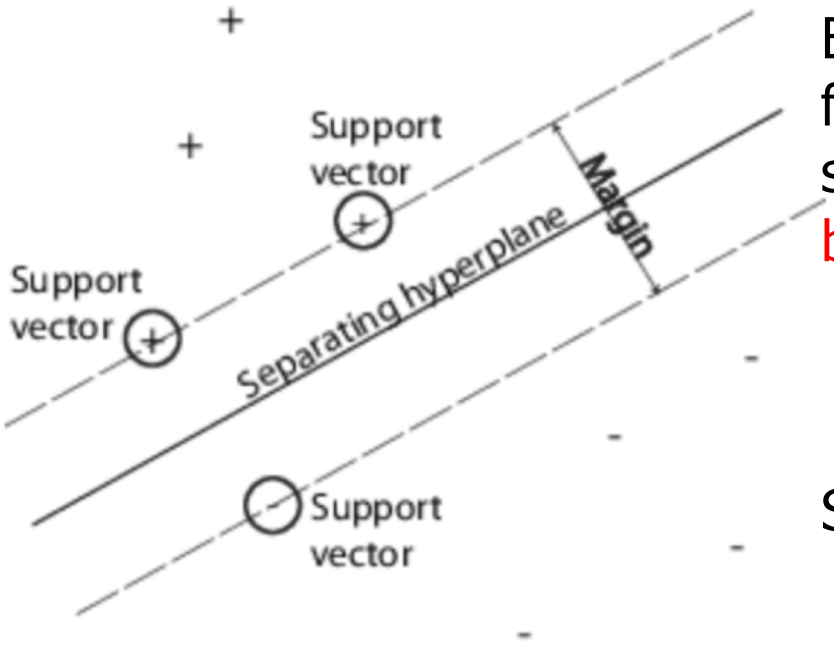
Assumption function

Therefore, it is necessary to calculate the distance from the discrimination boundary. Assuming that the decision boundary is $y\,(w^T x) = 0$, the distance from the decision boundary to each sample (x_i) is given by $\frac{y_i\,(w^T x_i)}{\|w\|}$ (1)

(1)from the Formula of the distance between a point and a line



Assumption function



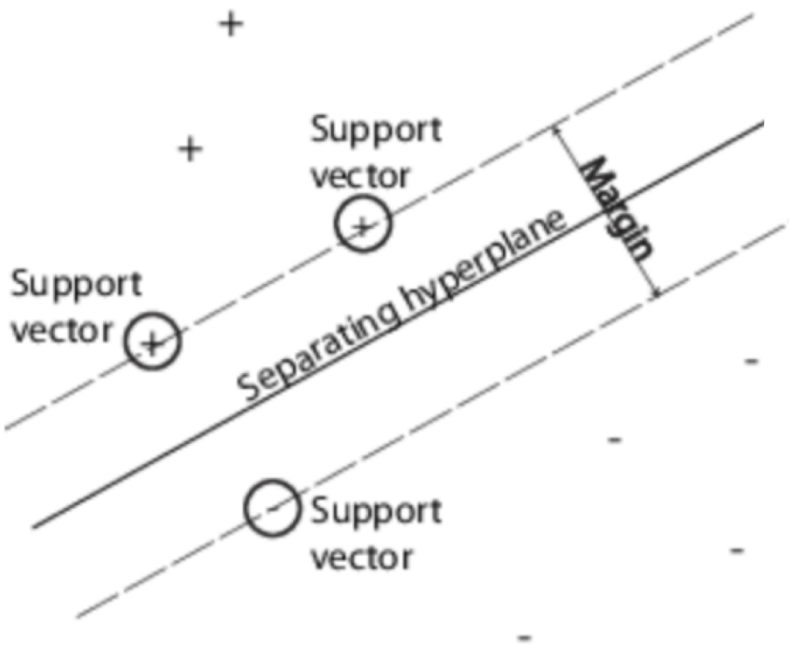
Aim to maximize margin

Before maximizing the margin, we first want to find xi, which is the smallest distance from the decision boundary. The formula is as follows:

$$\min_i \frac{y_i \left(\mathrm{w}^T x_i\right)}{\|\mathrm{w}\|}$$

Such x_i is called a support vector

Assumption function



Since w doesn't depend on i, we can put 1/||w|| in front

$$\frac{1}{\|w\|} \min_i \left(y_i \left(w^T x_i\right)\right)$$

Furthermore, in order to find w that maximizes the positive separation (entire equation) for such x, the following equation should be optimized.

$$\arg\max_w \left\{ \frac{1}{\|w\|} \min_i \left(y_i \left(w^T x_i\right)\right) \right\}$$

for that x, find the one with the most full key (maximum margin)

this becomes smaller, that is, find the support vector

Here, even w, which is a constant such as kw for all values represented by the vector , is the most convincing (since w exists in both parameters, it is offset even if it is multiplied by k). Therefore, the constant may be increased so that the normal separation from the parameter boundary to the subject is obtained.

$$y_i \left(w'^T x_i\right) = 1$$

$$\arg\max_{w'} \left\{ \frac{1}{\|w'\|} \min_i \left(y_i \left(w'^T x_i\right)\right) \right\}$$

Objective function

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

Objective function

Set up an objective function

Scaling $y_i \left(w'^T x_i\right) = 1$ to set the objective function eliminates the middle $\min_i \left(y_i \left(w'^T x_i\right)\right)$ and replacing it with the variable w'= w results in the problem of maximizing the following

$$\arg\max_w \frac{1}{\|w\|}$$

This is a problem that maximizes $\|w\|^2$, but reformulates it as a problem that minimizes $\|w\|^{-1}$;

$$\arg\min_w \|w\|^2$$

In addition, a coefficient of ½ is added to facilitate the derivative performed later

$$\arg\min_w \frac{1}{2}\|w\|^2$$

Since we scaled to $y_i \left(w'^T x_i\right) = 1$ above, the following constraints are satisfied for all i (i ∈ N)

$$y_i \left(w^T x_i\right) \geqq 1$$

Objective function

Describe the objective function and constraints that have been established so far.

$$\arg\min_{w} \frac{1}{2}\|w\|^2$$
$$\text{subject to } y_i\left(w^T x_i\right) \geqq 1$$

A problem that minimizes a quadratic function under such a linear inequality or a constraint by a linear equation is called a quadratic programming problem. The quadratic programming problem is guaranteed to be the optimal solution for the local solution by solving the **lagrange undetermined multiplier** method (with **KKT conditions**), which is also the quadratic programming problem. This is the replacement of the theorem based on the dual problem. L (Lagrange function), which is a combination of the above objective function and inequality constraints, is defined as follows:

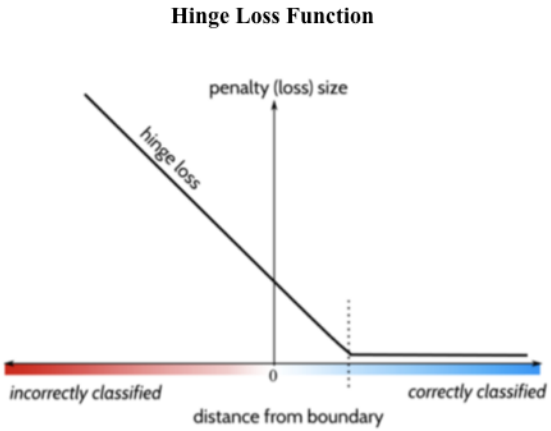
$$L(w,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i \left\{y_i\left(w^T x_i\right) - 1\right\}$$

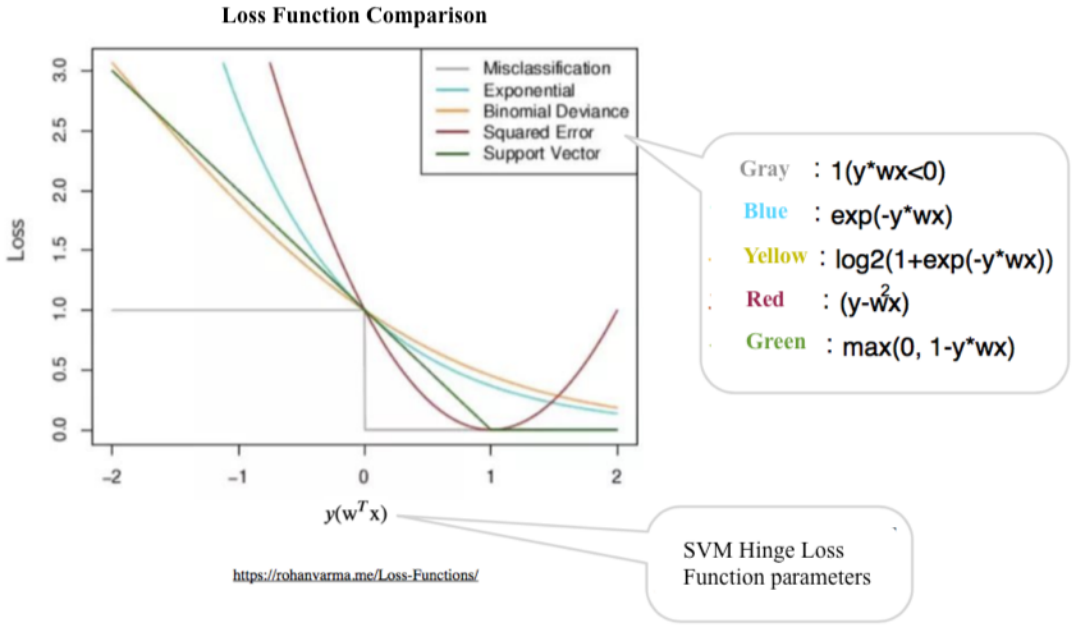
The lambda in the formulas is called the **Lagrange multiplier.** Replacing w in this equation gives a dual representation to be optimized. If the main problem is to minimize the objective function, the dual problem is to maximize the Lagrange function.

Objective function

About the constraints

The above-mentioned SVM constraint $y_i \left(w^T x_i\right) \geqq 1$ can be formulated as a loss function. This is called the hinge loss function. The hinge loss function of SVM decreases linearly until $y_i \left(w^T x_i\right)$ becomes 1. A function whose output is constant at 0 when it is 1 or more. A regularization term may be added after it.





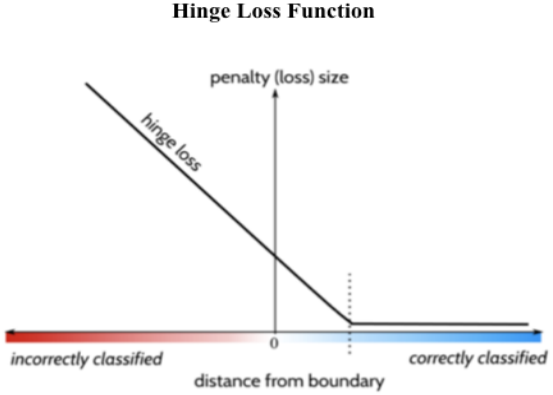
https://rohanvarma.me/Loss-Functions/

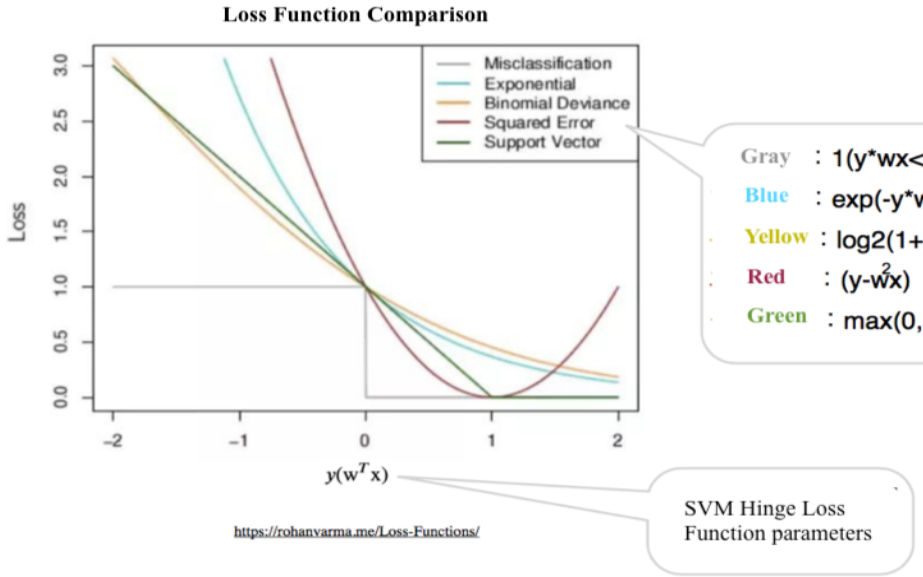
Objective function

Hinge Loss Function

$$hinge((\mathrm{w}^T x), y) = \max\{0, 1 - y(\mathrm{w}^T \mathrm{x})\}$$

No loss penalty is given when $y(\mathrm{w}^T \mathrm{x})$ is greater than or equal to 1 (away from occupational boundaries), fro example, using mean squared differences also penalizes well-identified samples. In the case of the hinge loss function, it doesn't affect the samples further from the support spectrum $y(\mathrm{w}^T \mathrm{x}) \geq 1$, while the closer it is to the identification boundary than the support spectrum, the greater the penalty is imposed





https://rohanvarma.me/Loss-Functions/

Lagrange undetermined multiplier method

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

Lagrange undetermined multiplier method

Lagrange undetermined multiplier method

This is an equal sign constrained optimization method that finds the extremum of a function under constraints. Specifically, it is a method of finding x that minimizes (or maximizes) the objective function f(x) under the constraint that g(x)=0. The following equation is a general form of the condition that the optimization problem must meet

We want to minimize f(x) but maximize L (x, λ)

$$\frac{\partial L}{\partial x} = 0$$

$$\frac{\partial L}{\partial \lambda} = 0$$

$$L(x, \lambda) = f(x) - \lambda^{T} g(x)$$

$$\lambda = (\lambda_1, \ldots, \lambda_m)^{T}, g(x) = (g_1(x), g_2(x), \ldots, g_m(x))^{T}$$

This time the problem is to minimize f(x), but in the case of the problem tomaximize:

$$L(x, \lambda) = f(x) + \lambda^{T} g(x)$$

General form of Lagrange function

1/2 ||w|| wants to be minimized but L(x, λ) as a whole is maximized

$$\longrightarrow L(w, \lambda) = \underbrace{\frac{1}{2}\|w\|^2}_{f(x)} \underbrace{- \sum_{i=1}^{N} \lambda_i \left(y_i \left(w^T x_i \right) - 1 \right)}_{-\lambda g(x)}$$

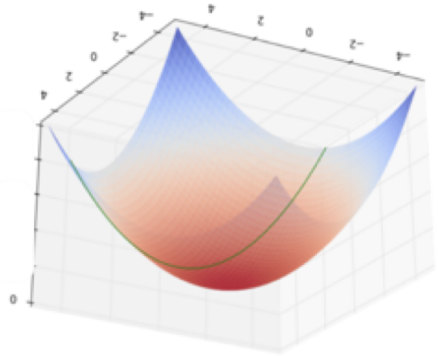
Put the objective function and constraints of the main problem into the form of this Lagrange function

Lagrange undetermined multiplier method

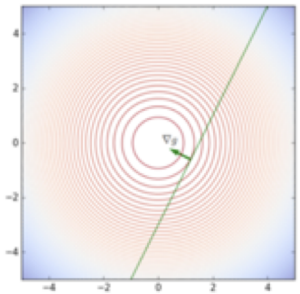
What the Lagrange function shows

Let's look at the geometrical relationships and understand them intuitively. When f(x) is represented by a downwardly convex three-dimensional curved surface as shown in the figure below, the curve (contour line) on the curved surface represents f(x)=z.
Furthermore, g(x) = 0, can also be drawn with a curve (green line)

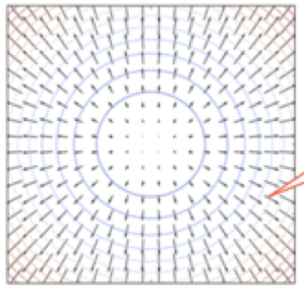
The goal of this optimization problem is to find the point where f(x) is the smallest at g(x) superior or equal to 0. Therefore, first, find the point where the contour line and g (x) = 0 are in contact. If there is such a point on g(x) = 0 (equation condition), it will be input λ different to 0, which is the support vector. On the other hand, g(x) > 0 (inequality) where λ = 0, which is not a support vector. This satisfies the complementary condition of the KKT conditions



f(x) and g(x)=0



Normal vector of g(x)=0



Contour normal vector

At the point where the contour line and g(x) = 0 are in contact, the gradient vectors ∇ and ∇g are parallel according to the KKT condition (this time, the direction of the vector is also the same). The gradient vector ∇f of the curved surface is the normal spectrum of the contour line. And this normal vector extends radially from the origin.

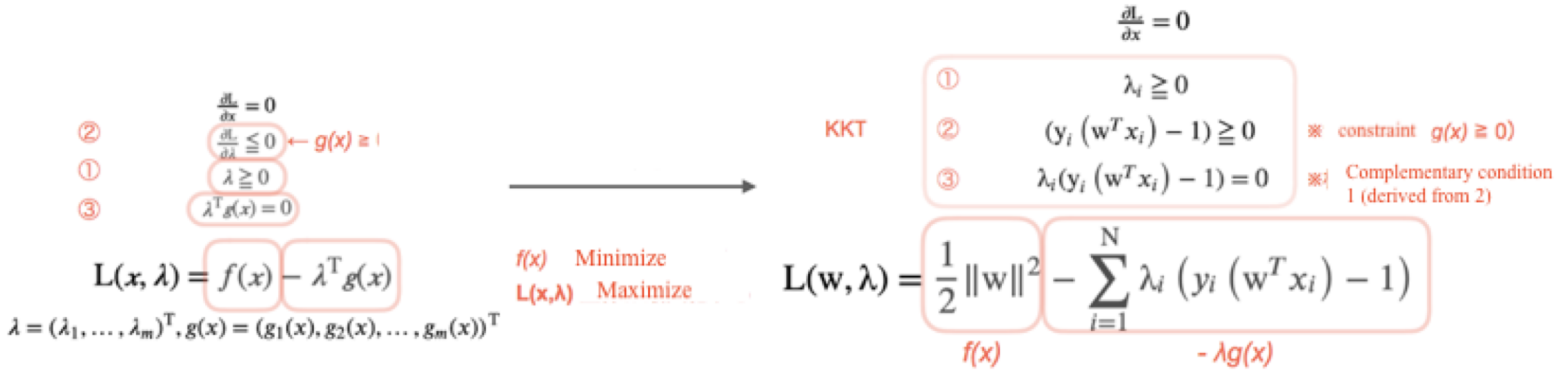
Karush-Kuhn Tucker conditions

KKT condition (Karush-Kuhn Tucker condition)

A requirement that an extremum must meet when solving an optimization problem. The advantage of applying the KKT condition to the Lagrange undetermined multiplier method to find the solution is that the Lagrange undetermined multiplier method is generalized and can handle not only equal signs but also inequality constraints.

Since the objective function of SVM has the inequality sign y_i (w^T x_i)≥1, it is formulated by the lagrange undetermined multiplier method (with KKT conditions) and the dual problem is solved.

Finally, the inequality sign constraint g(x) ≥ 0 finds x that minimizes the objective function f(x)



General form of lagrange function (with KKT condition)

lagrange function (with KKT condition)

Dual representation and gradient method preparation

Dual representation and gradient method preparation

The above-mentioned Lagrange function is expanded to obtain a dual representation. Furthermore, the dual representation is differentiated by the Lagrange multiplier to prepare a gradient for the later gradient method

In the middle of the formula expansion, the formula is differentiated for w, and $w = \sum_{i=1}^{N} \lambda_i y_i x_i$ and substituteed, and the parameter of the Lagrange function is only the N-Grange multiplier that can be used.

$f(x)$ $-\lambda g(x)$

$$
\begin{aligned}
L(w,\lambda) &= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i \left(y_i \left(w^T x_i\right) - 1\right) \\
&= \frac{w^T w}{2} - \sum_{i=1}^{N} \lambda_i \left(y_i \left(w^T x_i\right) - 1\right) \\
&= \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i x_i\right)^T \left(\sum_{i=1}^{N} \lambda_i y_i x_i\right) - \sum_{i=1}^{N} \lambda_i \left\{y_i \left(\left(\sum_{i=1}^{N} \lambda_i y_i x_i\right)^T x_i\right) - 1\right\} \\
&= \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i \left(\sum_{i=1}^{N} \lambda_i y_i x_i\right)^T x_i \\
&= \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i - \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j \\
&= \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j
\end{aligned}
$$

Text

This gradient is obtained by differentiating the objective function of the above function of the equation with an input

$$\frac{\partial L(\lambda)}{\partial \lambda_i} = 1 - y_i \sum_{j=1}^{N} \lambda_j y_j x_i^T x_j$$

This is a dual representation for the main problem (SVM).
(The latter formula is summarized, but the same objective function as the DIVER formula)

Gradient method

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

Gradient method

Maximize the Lagrange function

Update the input with the gradient obtained by differentiating the dual representation with the input (Lagrange multiplier).

This dual representation function is updated by adding a gradient to the input in order to maximize it with respect to the input. When searching for the maximum value of a function in this way, it is called the gradient ascent method

$$\frac{\partial L(\lambda)}{\partial \lambda_i} = 1 - y_i \sum_{j=1}^{N} \lambda_j y_j x_i^T x_j$$

Gradient

$$\lambda_i^{new} = \lambda_i + \alpha(1 - y_i \sum_{j=1}^{N} \lambda_j y_j x_i^T x_j)$$

Learning rate

Support Vector

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

Support vector

Assumption Function

Estimation is performed using the sparse solution () obtained as result of finding the optimum solution and the hypothetical function $f(x) = \text{sign}(\mathrm{w}^T x)$

The w that appears in the hypothetical function is replaced by $\mathrm{w} = \sum_{i=1}^{N} \lambda i y_i x_i$

The sign function $\text{sign}(\mathrm{w}^T x) = \begin{cases} +1 & \text{if } (\mathrm{w}^T x) \geqq 0 \\ -1 & \text{otherwise} \end{cases}$ returns 1 or -1.

$$f(x) = \text{sign}(\mathrm{w}^T x)$$

$$= \mathrm{y}(\mathrm{w}^T x)$$

This equation is obtained by differentiating the Lagrange function with w

Now replace w with $\mathrm{w} = \sum_{i=1}^{N} \lambda i y_i x_i$

Hypothetical function for estimation :

$$f(x) = \sum_{n=1}^{N} \lambda_n y_n x^T s_n$$

Unknown data

Support vector

Number of support vectors

SVM – The Flow

Knowing the SVM Problem Setting

① Formulate an equation (hypothetical function) to derive predictions

② Set up a problem to be minimized and formulate an equation (objective function).

③ Replace the objective function of the main problem with the objective function of the dual problem

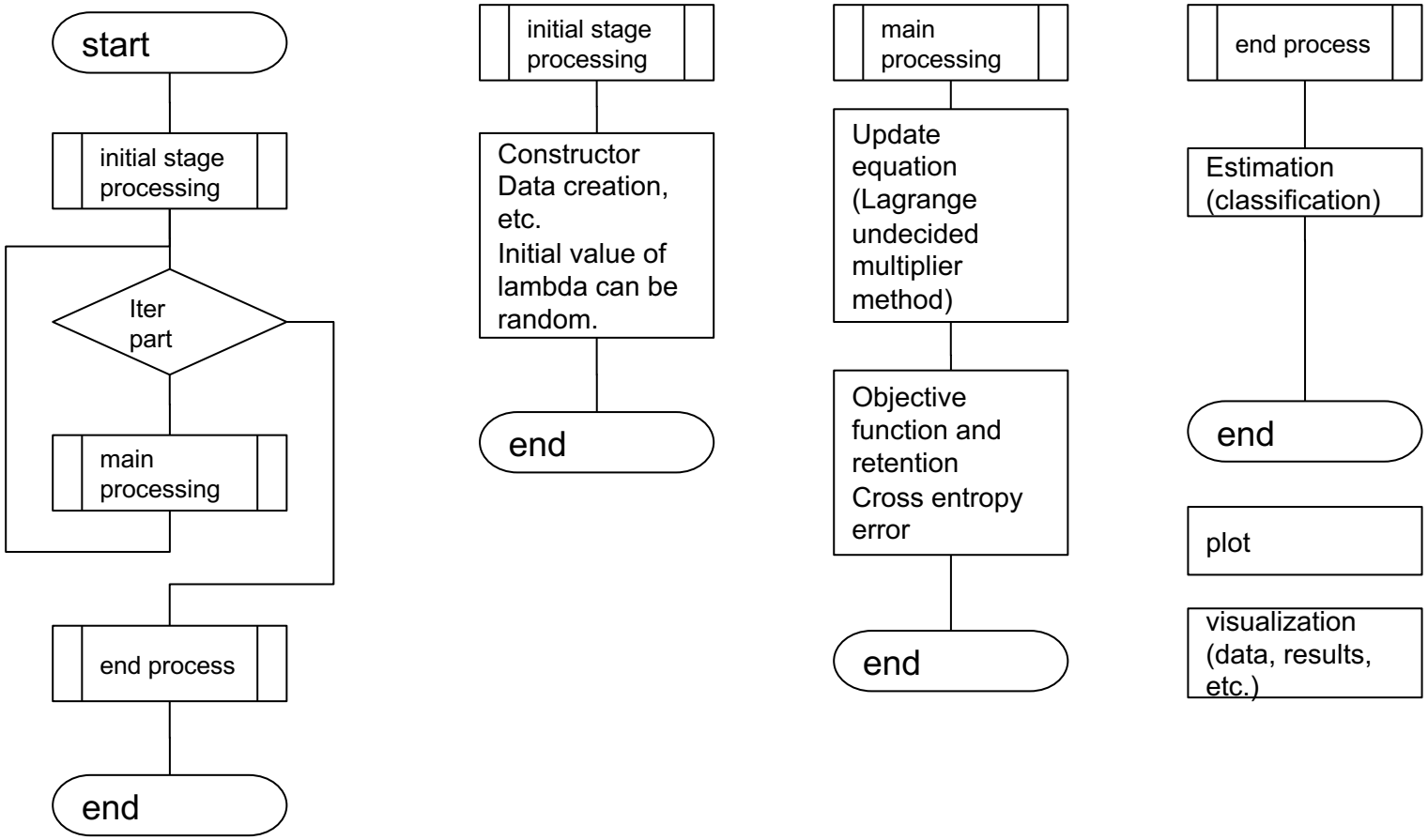
④ Exploratory search for the optimal solution of the objective function (gradient method)

⑤ Estimate using the obtained sparse solution (support vector)

SVM
Completed

How to learn SVM?

Functions required for the Scratch SVM.



SVM 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 SVM module

Sprint 5 – Scratch SVM

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

Sprint 4 – SVM

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

Sprint 5 – Sample Code

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

[Sprint 5 – Scratch SVM]

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

ToDo by next class

Next class will be Zoom : Thursday 3 June 2021

🚨ToDo: Scratch Decision Tree
https://diver.diveintocode.jp/curriculums/1648

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