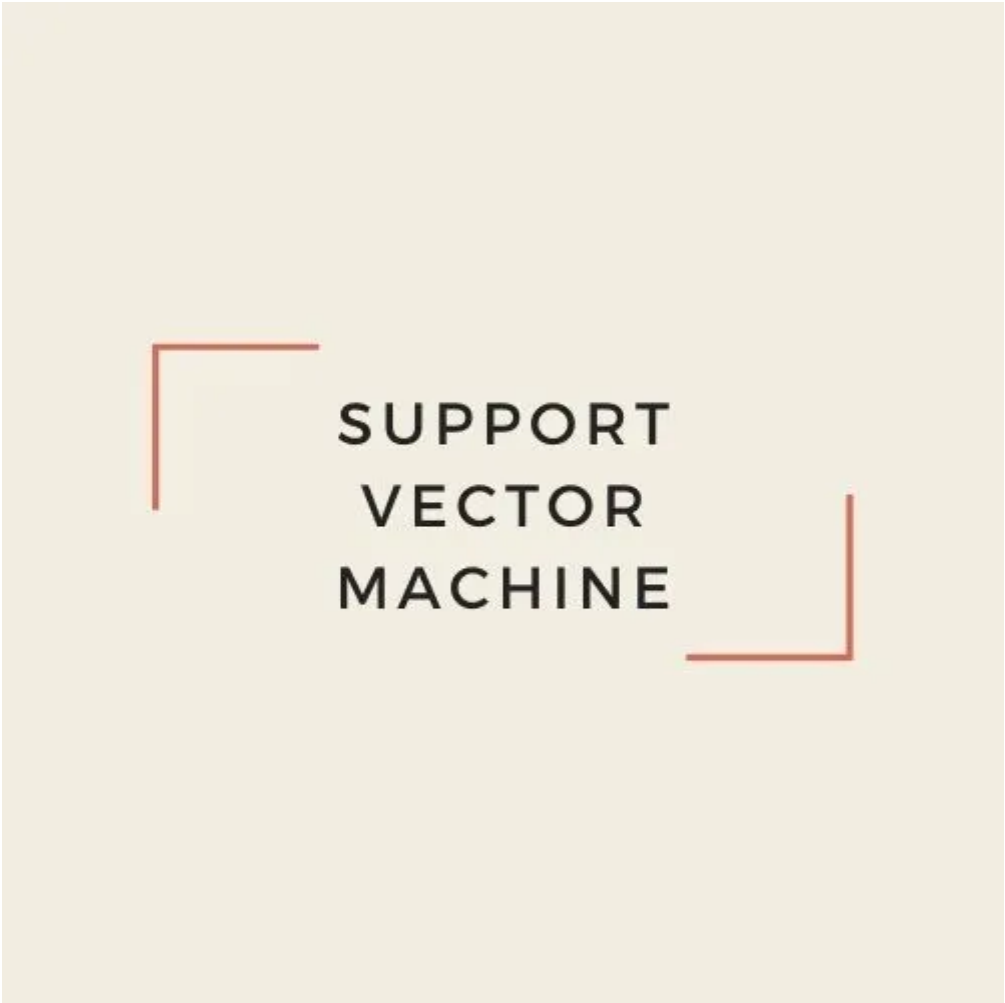
Support Vector Machine (SVM) and Kernels Trick

Indriani Sitorus · Follow
Published in Analytics Vidhya
8 min read · Aug 27, 2020



My Own SVM Design using Canva

What is SVM?

Support Vector Machine (SVM) is a type of algorithm for classification and regression in supervised learning contained in machine learning, also known as

support vector networks. SVM is more commonly used in classification problems rather than regression.

The SVM algorithm first introduced by Vapnik with colleagues Bernhard Bose and Isabelle Guyon in 1992 as a harmonious series of superior concepts in pattern recognition.

SVM works by using Structural Risk Minimization (SRM) principle which aims to obtain the best hyperplane line that divides data into two class in the input space.

At first SVM works linearly, but then SVM was developed again so that it can work non-linearly by looking for the hyperplane that is used to calculate the distance (margin) between data classes. In SVM application can be applied in linearly and non-linearly classification.

The SVM method is divided into two types based on its characteristics, namely linear SVM and non-linear SVM. Linear SVM is to classify data that can be separated linearly in two classes using soft margins. Linear classification is generally applied to datasets that have lower dimensions, that is, where the dataset has few features to classify. Meanwhile, Non-linear SVM is using the kernel concept in a high-dimensional workspace. The kernel concept is a function used by modifying the SVM algorithm to solve non-linear problems.

The SVM concept is called an attempt to find the best hyperplane that will divide data into two classes in the input space. The main objective of the training process on the SVM concept is to find the location of the hyperplane. SVM method uses the dot product function. The hyperplane is the line used to separate the dataset. Hyperplane can be a line in two dimensions and can be a flat plane in multiple planes. Illustration of determining the best hyperplane in the SVM algorithm. Following is the illustration of the best hyperplane in SVM.

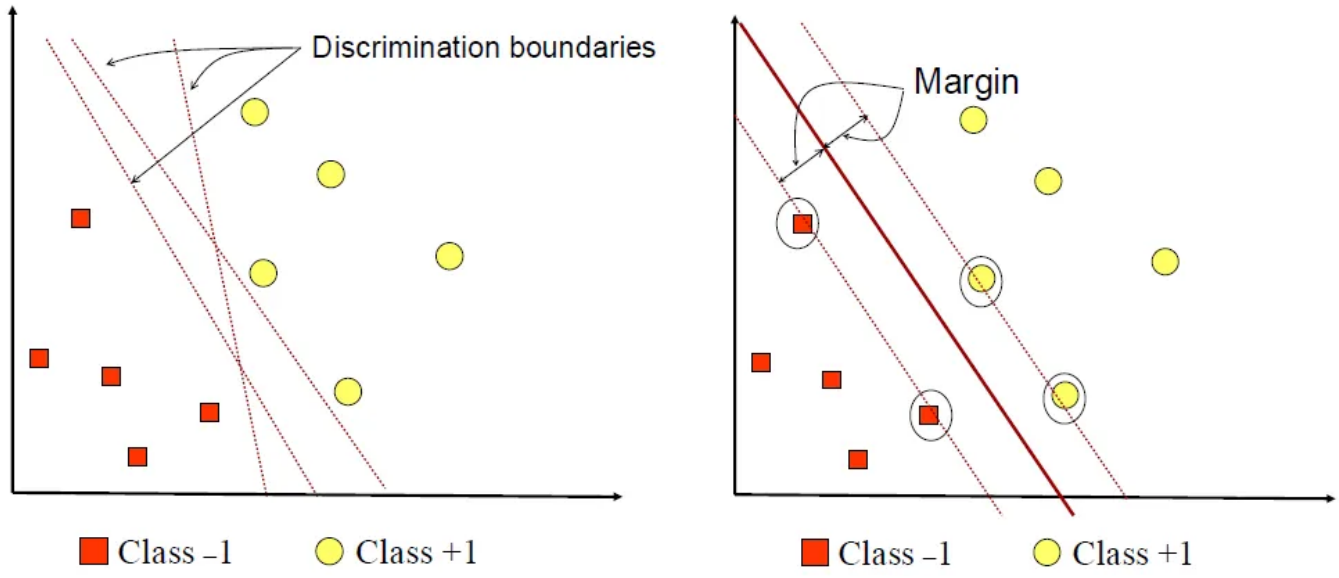


Illustration of Best Hyperplane Determination on SVM

The hyperplane can be obtained by measuring the hyperplane margin, which is the distance between the hyperplane and the closest point of each data class. The closest point that separates the hyperplane is called the support vector.

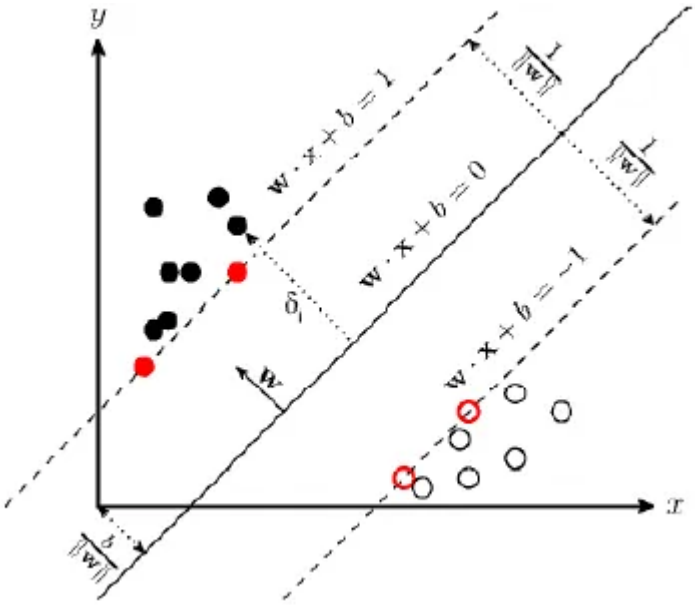
In the figure above, there is a yellow circle data which is data in class +1 and and the red box data is data in class -1. The yellow circle data is a member of class +1, while the red box data is a member of class -1. The best hyperplane that can be seen in the red line is in the middle of the positive hyperplane and the negative hyperplane. Meanwhile, the support vector is a yellow circle and a circled red box. Now I will describe part of the type of SVM. Check it out!

Linier SVM

Linear classification is generally used on datasets with lower dimensions. The lower dimension of a dataset means that it has fewer features to classify. Hyperplane in both images can be obtained by measuring the distance (margin) between the hyperplane and the closest point in each class.

Examples of cases belonging to the linear classification are to determine whether age and dietary factors affect human health. Where in this case there are only two features that are factors that affect human health, namely the age factor as feature x and thqe food factor as feature y. The following is a visualization of the linear SVM case.

Linear SVM is one of the working principles of SVM which is used for data that can be separated linearly as in the figure below.



Visualization of Linier SVM

The data available in SVM is symbolized by the notation (xi) ∈ R^d and the label of each class, namely class +1 and class -1 which are assumed to be perfectly separated by a hyperplane with d dimension given the notation yi ∈ {-1, + 1} , where i = 1,2,..., *l*; where l is a lot of data. So that the definition of the hyperplane equation is obtained as follows:

f (x) = w ^ T.x + b or w.x + b = 0

So that according to a hyperplane equation is obtained in the linear SVM for positive class:

w. (xi) + b ≤ + 1

Whereas for the negative class hyperplane equation in the linear SVM are:

w. (xi) + b ≥ - 1

Information:

w = weight (weight vector)

x = matrix input value (feature)

b = bias

To calculate the largest margin value, it is done by optimizing the distance value between the hyperplane and the closest point in each class. Quadratic Programming (QP) is used as a formula to find the minimal point of an equation with equation constraints:

τ (w) = 1/2 ||w|| ^ 2

yi ((xi) .w + b)-1≥0

The above problems can be solved with various computational techniques, one of which is to use the Langrange Multiplier equation as follows:
L (w, b, α) = 1/2 ||w || ^ 2-∑_ (i = 1) ^ *l* αi (yi ((xi) .w + b) -1)

With,
i = 1,2,..., *l*

Langrage multipliers = *αi* which has a zero or positive value (αi≥0), where i = 1,2,..., *l*. So that the Langrange multiplier equation is modified and only contains αi as follows:
∑_ (i = 1) ^ *l* αi-1/2 ∑_ (i, j = 1) ^ *l* αi αj yi yj (xi), xj

With,
αi ≥ 0; (i = 1,2,..., *l*); ∑_ (i = 1) ^ *l* αi yi = 0

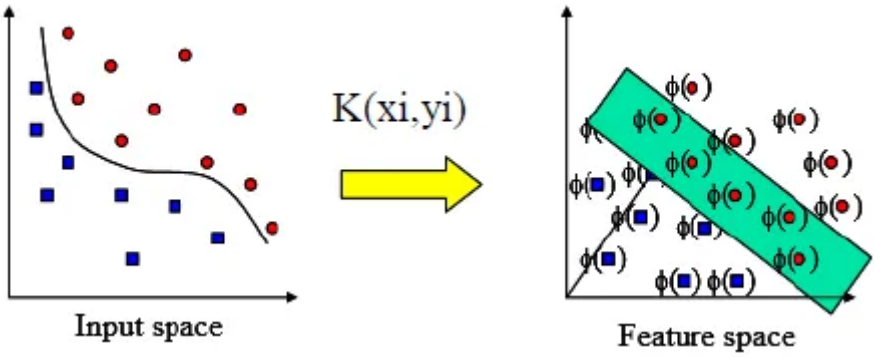
Most of the above calculations obtained a positive αi, where the data correlated with positive αi is called the support vector. So that the following equation is used to determine the results of the new data classification:
Class = sign f (x)

Non-Linier SVM

Non-Linear SVM is another working principle on SVM that is used for data that cannot be separated linearly because of its high dimensions. Non-linear classification is carried out using the kernel concept. The kernel concept in the non-linear case plays a role in determining the classification limits used as a model.

Non-Linear SVM applies the function of the kernel concept to a space that has high dimensions. What is meant by high dimension is that the dataset has more than two features to classify. For example, non-linear classification cases, namely factors that affect human health, consist of age factors, dietary factors, exercise factors, heredity, disease history and stress levels.

In this example, the kernel concept serves to determine the classification boundaries used as a model. Visualization of the non-Linear SVM case can be seen in the following Figure.



Visulization of Non-linier SVM

The accuracy of the model generated by the process in the SVM algorithm is very dependent on the parameters and kernel functions used. In the use of kernel functions in non-linear SVM is something that needs to be considered because the performance of SVM depends on the choice of kernel function.

Non-linier SVM is implemented in practice using a kernel, so it can separate data with the kernel function it called kernel trick.

The Kernel Trick

SVM can work well in non-linear data cases using kernel trick. The function of the kernel trick is to map the low-dimensional input space and tranforms into a higher dimensional space.

- **Radial Basis Function Kernel (RBF)**

The RBF kernel is the most widely used kernel concept to solve the problem of classifying datasets that cannot be separated linearly. This kernel is known to have good performance with certain parameters, and the results of the training have a small error value compared to other kernels. The equation formula for the RBF kernel function is:

```
K(x,xi) = exp(-gamma * sum((x - xi^2))
```

The Gaussian kernel RBF has two parameters, namely gamma and sigma. The gamma parameter has a default value, which is $\gamma = 1 / (2\sigma) \wedge 2$. When gamma is high, the points around the data are likely to be considered in the calculation. The sigma parameter is used to find the optimal value for each dataset.

In the RBF kernel function equation, ||xi-x || is the Euclidean Distance between x1 and x2 in two different feature spaces and σ (sigma) is the RBF kernel parameter that determines the kernel weight. In SVM, sigma parameters need to be adjusted to provide accurate classification results. The default value of the sigma parameter is σ = 1.

- **Polynomial Kernel**

A Polynomial Kernel is more generalized form of the linear kernel. In machine learning, the polynomial kernel is a kernel function suitable for use in support vector machines (SVM) and other kernelizations, where the kernel represents the similarity of the training sample vectors in a feature space. Polynomial kernels are also suitable for solving classification problems on normalized training datasets. The equation for the polynomial kernel function is:

```
K(x,xi) = 1 + sum(x * xi)^d
```

This kernel is used when data cannot be separated linearly.
The polynomial kernel has a degree parameter (d) which functions to find the optimal value in each dataset. The d parameter is the degree of the polynomial kernel function with a default value of d = 2. The greater the d value, the resulting system accuracy will be fluctuating and less stable. This happens because the higher the d parameter value, the more curved the resulting hyperplane line.

- **Sigmoid Kernel**

The concept of the sigmoid kernel is a development of an artificial neural network (ANN) with the equation for the kernel function is:

```
K(x,xi) = tanh (α xi.xj + β)
```

The Sigmoid kernel has been proposed theoretically for a Support Vector Machine (SVM) because it originates from a neural network, but until now it has not been widely used in practice.

The sigmoid kernel is widely applied in neural networks for classification processes. The SVM classification with the sigmoid kernel has a complex structure and it is difficult for humans to interpret and understand how the sigmoid kernel makes classification decisions. Interest in these kernels stems from their success in

classifying with the neural netwotk and logistic regression, specific properties, linearity and cumulative distribution.

The sigmoid kernel is generally problematic or invalid because it is difficult to have positive parameters. The sigmoid function is now not widely used in research because it has a major drawback, namely that the output value range of the sigmoid function is not centered on zero. This causes the backpropagation process to occur which is not ideal, so that the weight of the ANN is not evenly distributed between positive and negative values and tends to approach the extreme values 0 and 1.

- **Linear Kernel**

A linear kernel can be used as normal dot product any two given observations. The equation for the kernel function is:

```
K(x, xi) = sum(x * xi)
```

Finally, that's it. Hopefully, this section helpful in understanding the concept of SVM and kernel trick for you guys. You may give some comments, thoughts, feedback or suggestions below. Keep learning and stay tuned for more!

Python | Machine Learning | Artificial Intelligence | Image Preprocessing | Svm

Follow

Written by Indriani Sitorus

36 Followers · Writer for Analytics Vidhya

Software Quality Assurance