Summary for Support Vector Machine (SVM)

MossyFighting

March 19, 2020

Contents

1 Introduction

They key idea of SVM is to use a hyperplane (or separator) to classify a dataset that can be linearly seperable or linearly non-separable.
The dimension of hyperplane is less than one compared to its ambient space that is the dimension of the considered object (or point). For instance, if the point is in three dimension coordination, the hyperplane will have two dimension and it is equivalent to a plane. if the point is in two dimension corrdination, the hyperplane will have one dimension that is a simple line.
A dataset can be comprised of many data points, each data point can have n dimensions (features, attributes) depending on problem we are facing.

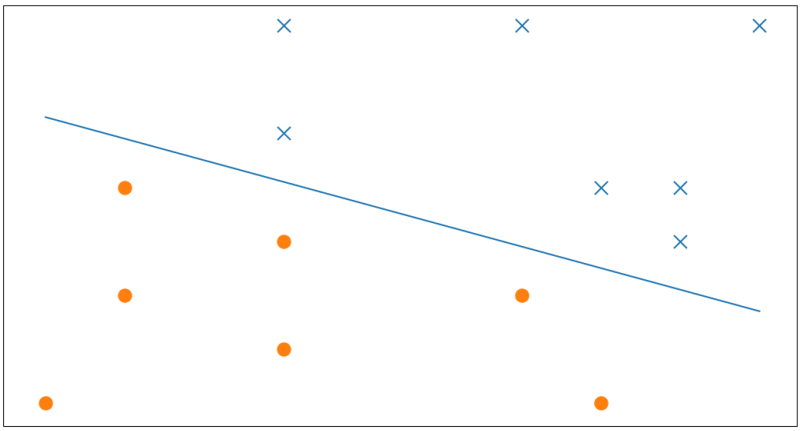

Figure 1: Linearly separatable

Linearly separatable data can be obtained if there is a line (see Fig. 1) that can separate group of red points and group of blue points.
Linearly non-separable dataset, if we can not find any line to completely separate two groups in red and green in dataset see Fig. 2.

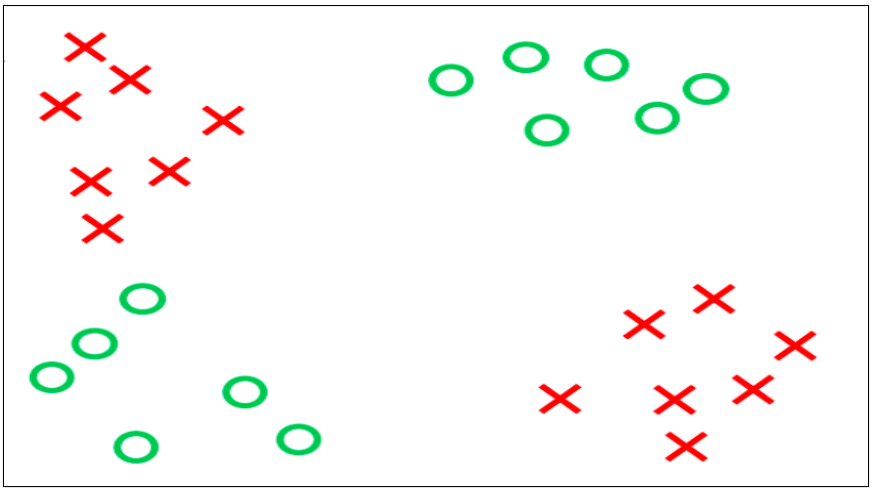

Figure 2: Linearly non-separatable

2 Mathematic notes

2.1 Hyperplane equation

Given a vector $\mathbf{w} = \{w_1, w_2, ..., w_n\}$, a ambient space with n dimension representing by vector $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, and a bias b. Then, the hyperplane is defined as follows:

$$\mathbf{w} \cdot \mathbf{x} + b = 0,$$

any point $\mathbf{x}^{(j)}$ where superscript j inside the bracket refers to index of a point in dataset. Each point has n dimension.
Each point $x^{(j)}$ in dataset is associated to a label $y^{(j)}$ (e. g., blue, red in Fig. 1 or blue, green in Fig. 2). In numeric way, we can assign blue as label ($y^{(j)} = +1$) and red as label ($y^{(j)} = $ -1).
To separate dataset (if they are linearly separable) using hyperplane, a hypothesis function is defined as following:

$$h(x^{(j)}) = sign(\mathbf{w} \cdot \mathbf{x}^{(j)} + b),$$

where $sign()$ function is label (+1) or label (-1) depending on $\mathbf{w} \cdot \mathbf{x} + b$ as follows:

$$h(x^{(j)}) = \begin{cases} +1 & if \quad \mathbf{w} \cdot \mathbf{x}^{(j)} + b \geq 0 \\ -1 & if \quad \mathbf{w} \cdot \mathbf{x}^{(j)} + b < 0 \end{cases}$$

2.2 Optimal hyperplane

It seems that there will be many hyperplanes that are able to separate the dataset. However, if the separating hyperplane is too closed to data points of one class, it has high chance to fail to generalize on the unseen data.
Therefore, our task is to find the best hyperplane $(\mathbf{w}, b)$ among many separating hyperplanes. In order to do that, the metric (e.g., distance) is used.

In order to correctly classify data points $x^{(j)}$ with its label $y^{(j)}$, this expression $D^{(j)}) = y^{(j)}(\mathbf{w} \cdot \mathbf{x}^{(j)} + b)$ must be positive. Otherwise, if $D^{(j)})$ is negative, the data points are incorrectly classified, which is we are not interested in.
To obtain the optimal hyperplane which is equivalent to find $(\mathbf{w}_k, b_k)$, we need to find the minimum distance among $j = \{1, 2, ..., M\}$ in dataset for each hyperplane, and then to maximize all the minimum distances assocciated to $k = \{1, 2, 3, ..., K\}$ hyperplanes. So the best hyperplane should be:

$$\underset{k=1,2,...,K}{maximize} \ (\underset{j=1,2,...,M}{minimize} \ d_k^{(j)}),$$

where, $d^{(j)} = \mathbf{w} \cdot \mathbf{x} + b$,and $F = \underset{j=1,2,...,M}{minimize} \ d^{(j)}$ is defined as **functional margin**.
However, $d^{(j)}) = y^{(j)}(\mathbf{w} \cdot \mathbf{x}^{(j)} + b)$ is highly depending on $(\mathbf{w}_k, b_k)$ in which $\mathbf{w}$ is a vector that is perpendicular to the hyperplane. If we scale up or scale down this vector, the vector direction is unchanged, but this will heavily affect on the distance of data points. Then to make sure everything have the same reference, a normalization will be carried out, and $d^{(j)})$ now will become:

$$dn^{(j)} = y^{(j)} \left(\frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x}^{(j)} + \frac{b}{||\mathbf{w}||} \right),$$

And now, $D = \underset{j=1,2,...,M}{minimize} \ dn^{(j)}$ is called **geometric margin**.

2.3 SVM optimization

Now we have to construct a form for optimization to find the best that describe a optimum hyperplane.
Assume that we have a dataset with M samples $x^{(m)} = \{x^{(1)}, x^{(2)}, ..., x^{(M)}\}$ according to its label $y^{(m)} = \{y^{(1)}, y^{(2)}, ..., y^{(M)}\}$, and a hyperplane with $(\mathbf{w}, b)$. The **geometric margin** of the hyperplane is defined as:

$$D = \underset{j=1,2,...,M}{minimize} \ dn^{(j)}$$

Then the optimal hyperplane is obtained through $(\mathbf{w}, b)$ for which the geometric margin D is maximized.
Then, the SVM problem:

$$\underset{k=1,2,...,K}{maximize} \ (\underset{j=1,2,...,M}{minimize} \ dn_k^{(j)}),$$

or

$$\underset{k=1,2,...,K}{maximize} \ (minimize \ D)$$

can be recast as:

$$\begin{cases} \underset{\mathbf{w},b}{maximize} & D \\ subject\ to & dn^{(j)} \geq D, \ j = 1, 2, ..., M \end{cases}$$

As we can see, $D = \frac{F}{||\mathbf{w}||}$, and $dn^{(j)} = y^{(j)} \left(\frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x}^{(j)} + \frac{b}{||\mathbf{w}||} \right)$ then,

$$\begin{cases} \underset{\mathbf{w},b}{maximize} & \frac{F}{||\mathbf{w}||} \\ subject\ to & y^{(j)} \left(\frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x}^{(j)} + \frac{b}{||\mathbf{w}||} \right) \geq \frac{F}{||\mathbf{w}||}, \ j = 1, 2, ..., M \end{cases}$$

From previous section we have, $d^{(j)} = \mathbf{w} \cdot \mathbf{x} + b$, then,

$$\begin{cases} \underset{\mathbf{w},b}{maximize} & \frac{F}{||\mathbf{w}||} \\ subject\ to & \frac{d^{(j)}}{||\mathbf{w}||} \geq \frac{F}{||\mathbf{w}||},\ j = 1, 2, ..., M \end{cases}$$

Eliminate $||\mathbf{w}||$ in both denominator of condition term, we have:

$$\begin{cases} \underset{\mathbf{w},b}{maximize} & \frac{F}{||\mathbf{w}||} \\ subject\ to & d^{(j)} \geq F,\ j = 1, 2, ..., M \end{cases}$$

If we normalize this optimization by a value F in both maximize term and in conditional term, then finally we have:

$$\begin{cases} \underset{\mathbf{w},b}{maximize} & \frac{1}{||\mathbf{w}||} \\ subject\ to & d^{(j)} \geq 1,\ j = 1, 2, ..., M \end{cases}$$

As we can notice, maximize $\frac{1}{||\mathbf{w}||}$ is equivalent to minimize $||\mathbf{w}||$, then we finally have:

$$\begin{cases} \underset{\mathbf{w},b}{minimize} & ||\mathbf{w}|| \\ subject\ to & y^{(j)}(\mathbf{w} \cdot \mathbf{x}^{(j)} + b) \geq 1,\ j = 1, 2, ..., M \end{cases}$$

Hoever, in practical scenarioes, the minimize $||\mathbf{w}||^2$ is more prefered than minimize $||\mathbf{w}||$, and the term $\frac{1}{2}$ is added to help the optimization converge faster. Then, we can rewrite the optimization problem as:

$$\begin{cases} \underset{\mathbf{w},b}{minimize} & \frac{1}{2}||\mathbf{w}||^2 \\ subject\ to & y^{(j)}(\mathbf{w} \cdot \mathbf{x}^{(j)} + b) \geq 1,\ j = 1, 2, ..., M \end{cases}$$

3 Dua-form optimization using Lagrange and KKT

To solve the optimization in previous subsection, the Lagrange method is used. The method help to find the local maxima and minima of a objective function with some equity constraints.
The Lagrange method can be summaried 3-step process as:

- Build a Lagrange function by adding all constriants in which each constraint multiply with one new variable called nultiplier.
- Derivative the Lagrange function $L(x, multipliers)$
- Solve the equation derivation $L(x, multipliers) = 0$

By using Lagrange methods, we introduce M multipliers $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_M\}$ because we have M constraints.
By following 3-step process of Lagrange method, we can transform the optimization in previous section into:

$$\begin{cases} \underset{\alpha}{Maximize} & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \cdot x^{(j)}) \\ subject\ to & \alpha_i \geq 0 \ \ for \ \ i = 1, 2, ..., M \\ & \sum_{i=1}^{M} \alpha_i y_i = 0 \end{cases}$$

where, $(x^{(i)} \cdot x^{(j)})$ means dot product between two vector training samples.
Remember that Lagrange only solve with equity contraints, but M constraints in the optimization is inequity contraints. So, the KKT conditions must be satisfy.
From apply KKT conditions, we can say that **support vectors** are equivalent to training sample having positive lagrange multipliers.
Variable $\mathbf{w}$ and b can be computed as:

$$\mathbf{w} = \sum_{j=1}^{M} \alpha_j y^{(j)} \mathbf{x}^{(j)}$$

We can average b over the support vectors.

$$b = \frac{1}{P} \sum_{i=1}^{P} (y^{(i)} - \mathbf{w} \cdot \mathbf{x}^{(i)}),$$

where P is the number of support vectors.
The hypothesis function $h(x^{(i)}) = sign(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$ now can become

$$h(x^{(i)}) = \sum_{j=1}^{M} \alpha_j y^{(j)} (x^{(j)} \cdot x^{(i)})$$

4 Soft-margin

Until now, we only try to solve for problem with linearly seperable dataset. When some noise, or outliers happen to dataset, the classes will not linearly seperable anymore. And our SVM will fail to give us a solutin $\mathbf{w}$ and b. To avoid this annoy, we will introduce a simple way to let our SVM at least give us a solution instead of failure.
A slack variable ζ is introduce and relax the constraint in every training sample. Then the optimization procedure can be rewrite as:

$$\begin{cases} \underset{\mathbf{w},b}{minimize} & \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{M} \zeta_i \\ subject\ to & y^{(j)}(\mathbf{w} \cdot \mathbf{x}^{(j)} + b) \geq 1 - \zeta_j,\ j = 1, 2, ..., M \end{cases}$$

The same process as before, dual-form of the optimization by using lagrange and KKT.

$$\begin{cases} \underset{\alpha}{Maximize} & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)} \cdot x^{(j)}) \\ subject\ to & 0 \leq \alpha_i \leq C \ \ for\ \ i = 1, 2, ..., M \\ & \sum_{i=1}^{M} \alpha_i y_i = 0, \end{cases}$$

where constant C now is related to the relax value from hard margin to soft margin.
The dua-form of soft margin is mostly the same with dua-form of hard margin (linearly separating data), with only different is that now $\alpha_i,\ i = 1, 2, ..., M$ now is constrained in range from 0 to C.

- if C is +infinity then, this is the hard margin, dealing with the linearly separable data. if not linearly, SVM can not give a solution.
- if C is small then, some misclassified data points will happen, due to noisy data or ouliers. Then, the SVM still gives the solution although data is noisy and some outliers.

5 Kernel

The strong characteristic of SVM is the ability to solve for problem with linear non-separatable by using kernel.
Kernel is the ability to transform the non linearly separable data the current ambient space to another space with higher dimension and in this transformed space, the data now become linearly separable.

Kernel Trick

5.1 Kernel definitions

Kernel trick is the way we manipulate the formulation of SVM optimization to find to optimal hyperplane.

For example, if we define $K(x^{(i)}, x^{(j)}) = (x^{(i)} \cdot x^{(j)})$, then the SVM optimization in the original form and dua-form using lagrange and KKT can be expressed by just substitude all the terms $(x^{(i)} \cdot x^{(j)}$ by the term $K(x^{(i)}, x^{(j)})$.

$$\begin{cases} \underset{\alpha}{Maximize} & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)}) \\ subject\ to & 0 \leq \alpha_i \leq C \ \ for \ \ i = 1, 2, ..., M \\ & \sum_{i=1}^{M} \alpha_i y_i = 0, \end{cases}$$

and the hypothesis can be obtained in the same way:

$$h(x^{(i)}) = \sum_{j=1}^{M} \alpha_j y^{(j)} K(x^{(i)}, x^{(j)})$$

5.2 Kernel types

Linear kernel

The simplest one is linear kernel where the data is linearly separable. Out task is to find a optimum straight line that could separate the data into two classes.
$K(x^{(i)}, x^{(j)}) = (x^{(i)} \cdot x^{(j)})$

Polynomial kernel

In case the data is not really linearly separable, polynomial can be a candidate to try. Just immagine, for example, in the current space, the data is linear not separable, but in higher dimension space, the data is really linear separable.
$K(x^{(i)}, x^{(j)}) = (x^{(i)} \cdot x^{(j)} + constant)^d$

Radial basic kernel

The radial basis function or gaussian is considered as a infinity space domain. This is the strongest tool of SVM when dealing with data that are not linearly separable.
$K(x^{(i)}, x^{(j)}) = exp^{(-\gamma ||x^{(i)} - x^{(j)}||^2)}$

6 Results

6.1 Radial basic kernel

In this example, assume that we want to classified two classes with labels 'blue' and 'red'. Clearly, it is impossible to find a straight line to separate those classes. Then, the kernel using radial basis function, or gaussian function is the way to try.
As mention the formulation in last section, the kernel has this formulation

$$K(x^{(i)}, x^{(j)}) = exp^{(-\gamma||x^{(i)}-x^{(j)}||^2)},$$

the value gamma is set to 0.5. Run the software and we have the results as follows.
The green contours show us regions which red and blue data points belong to. It shows that the SVM is able to separate complicated linear non-separable data points.

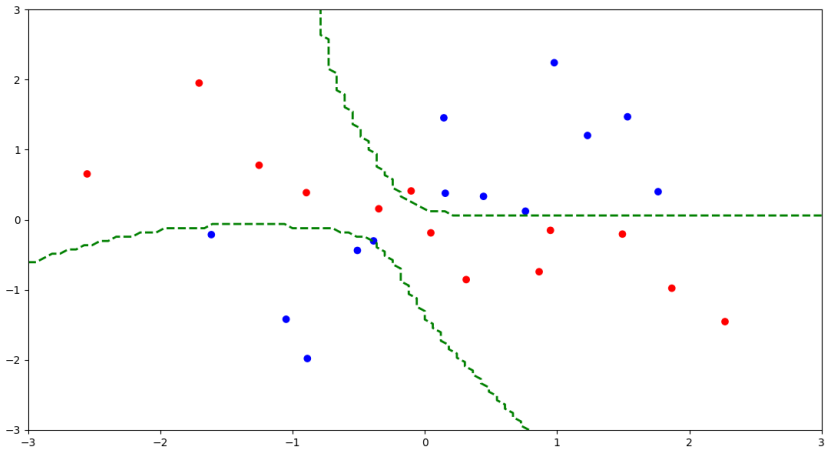

Figure 3: Contours seperation $\gamma = 0.5$

References