Support Vector Machines

Mannem Charan AI21BTECH11019

Abstract

This report consists of my basic understanding of one of the popular Ml methods "Support Vector Machines".

1 Support Vector Machines

SVM, Support Vector Machines is a supervised learning model which is used to solve classification and regression problems.Though we say regression, it is predominantly used for classification problems.SVM in general is used to find a decision boundary known as "**hyperplane**" to classify the data.And the dimensions of hyper plane depends on features/predictors of the data point.It is used for binary classification and in case of multi class it breaks the problem into multiple binary classification problems.

2 Types of SVM

Based on the dataset given, SVM are of two types,

1) **Linear SVM**:We use linear SVM when the data is **"linearly Seperable"**, when we say linearly seperable we mean that we that one can find a line(if one line exists then **infinitely** many lines are possible) which can classify the data into two classes.
   Ex: Hard-margin SVM
2) **Non-linear SVM**:We use Non-linear SVM when the data is not linearly seperable i.e., we can't find a straight line which can classify the data into two classes.
   Ex: Soft-margin classifier, kernals

3 Linear SVM

If the data is linearly seperable,then we can find infinitely many lines/hyperplanes to classify the data but some may lead to the chance of overfitting.So we choose a hyperplane which can classify the given data and have less generalisation error.To do that, we take the help of some data points known as "Support Vectors",the data points(of either class) which are close to the hyperplane.And the gap between these data points and the hyperplane is known as "Margin".So we use these support vectors to adjust the hyperplane by maximising the margin(in case of linear SVM).Since these vectors supporting the hyperplane, they are known as "Support Vectors".That's why the algorithm is named after them as "Support Vector Machines".First it is introduced as Support Vector Networks but then machine word became popular and people started using it as SVM.

4 Understanding Linear SVM

As we said earlier, we use Linear SVM when we have linearly seperable data.Let us define the labelled data as $\{(\mathbf{x_1}, y_1), \ldots (\mathbf{x_k}, y_k)\}$ where ,

1) k = no. of features
2) x_i = i^{th} data point
3) y_i = label of x_i and $y_i \in \{-1, 1\}$

Mathematically, to say that the dataset is linearly seperable, we should be able to find a $(\mathbf{w}, \theta)$ such that,

$$\mathbf{w}^T\mathbf{x} - \theta = \begin{cases} \geq 0 & , y_i = 1 \\ < 0 & , y_i = -1 \end{cases} \tag{4.1}$$

The above expression follows from the fact that the hyperplane classifies the data correctly.So one can find infinetly many $\mathbf{w}$ and θ that can satisfy above expression.So one of the classical methods to choose the optimal hyper plane is to **maximise** the margin.First, to find the margin we will draw parralel lines to the hyperplane that are passing through the support vectors.As you can see in the figure 4.1,
let say the two parallel hyperplanes are

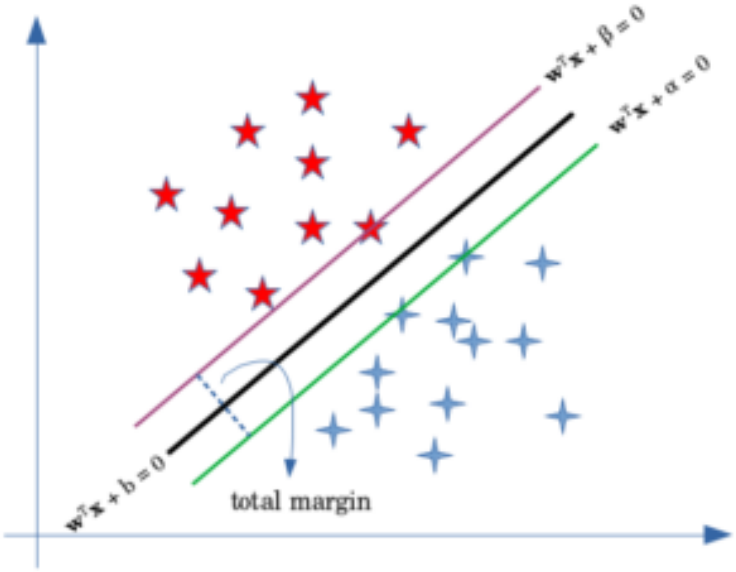

Fig. 4.1: Linearly seperable data

$$\mathbf{w}^T\mathbf{x} + \alpha = 0 \tag{4.2}$$

$$\mathbf{w}^T\mathbf{x} + \beta = 0 \tag{4.3}$$

And in this case let say $\mathbf{w}^T\mathbf{x} + b = 0$ be the hyperplane.So since we want to maximise the margin,we should maximise the distance between the two hyperplanes i.e, we want to maximise

$$d = \frac{|\alpha - \beta|}{||w||} \tag{4.4}$$

So to the maximise the margin $\frac{|\alpha-\beta|}{||w||}$, we should minimise the norm $||w||$. Now to make things simple let us take $\alpha = b + 1$ and $\beta = b - 1$, now the constraints will be

$$\mathbf{w}^T\mathbf{x} - \theta = \begin{cases} \geq 1 & , y_i = 1 \\ < -1 & , y_i = -1 \end{cases} \tag{4.5}$$

So the same thing can be written as ,

$$y_i\left(\mathbf{w}^T\mathbf{x} - \theta\right) \geq 1 \tag{4.6}$$

Because we want to classify the data point to class $+1$ whenever $\mathbf{w}^T\mathbf{x} + b \geq 1$ and to class -1 whenever $\mathbf{w}^T\mathbf{x} + b < -1$.So the margin in this case is $\frac{2}{||w||}$ and to maximise it we should minimise,

$$\frac{||w||}{2} \tag{4.7}$$

Now to make math simple we will try to minimise,

$$\frac{||w||^2}{2} = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{4.8}$$

So here the cost function will be

$$J = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{4.9}$$

with the constraints mentioned in (4.5).By minimising the cost function we will find the optimal hyperplane needed.
This method is also known as ”Hard - margin” SVM since we are strictly classifying the data points using the margin.But often we don’t have linearly seperable data in those cases we either relax the margin or use some kernel tricks.

5 Understanding Non-linear SVM

When we can’t divide the data using a straight line, we generally follow two approaches,

1) **Soft -margin**: Where we try to relax the margin to allow some data points to misclassify.

2) **Kernal besed approach**: In this case, we try to transform the training data(low dimensional data) into a high dimesional space in order to make the data linearly seperable.

6 Soft-Margin

When the data is not linearly seperable then maximising the margin is not a suitable method,in that case we try to relax the margin and allow some data points to misclassify.To do that we will add some new variables known as "Slack variables". These variables try to relax the margin and allow some points to misclassify.So the new constraints will be,

$$y_i\left(\mathbf{w}^T\mathbf{x} - \theta\right) \geq 1 - \xi_i \, (i = 0,\ 1,\ \ldots,\ k) \tag{6.1}$$

And if $0 \leq \xi_i \leq 1$ then the inner product stil have the same sign but with a reduced margin but if $\xi_i > 1$ then that data point will be misclassified.
Since we are allowing the data points to misclassify, we should minimise the error as low as possible.We call this error as "Hinge loss" and this is nothing but the lossoccured to a misclassified point i.e., ξ_i.That's why the cost function for soft-margin classifier is,

$$||\mathbf{w}||^2 + C\sum_{i=1}^{k}\xi_i \tag{6.2}$$

with the constraints mentioned in (6.1).And constant C is a tuning parameter which helps in adjusting the count of how many points to be misclassifed.If $C = 0$, then model reverts back to hard-margin classifier.And as C increases the violation of data points increases.Therefore soft-margin classifier can be considered as a more generalised form of Hard-margin classifier.You can visualise the same in Fig 6.2.
And moreover even though data is linearly seperable if the margin is too low,then the model have a risk of overfitting or it may become sensitive to the outliers.Hence in that case we can opt for a large margin by choosing soft-margin SVM in order to help the model generalise better.

7 Kernal Based Approach

In SVM, to deal with non-linear data although we have soft-margin SVM we will try use some techniques known as kernel tricks which actually classifies the given data in the higher dimensional space.First we try to map our feature set to an enlarged feature set,to understand that let us see an example Fig 7.3
The Fig 7.3 represents a dataset with two features x and y. And you can see, the data

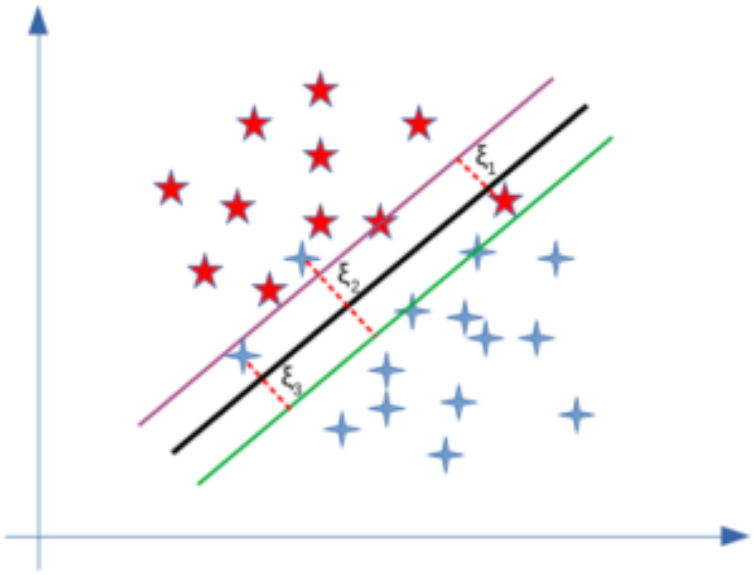

Fig. 6.2: A soft-margin classifier allows the data points(can be outliers) to misclassify in order to generalise better

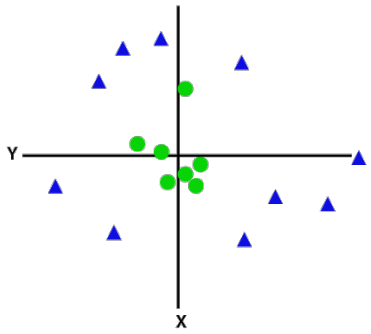

Fig. 7.3: Non -linear data

is not linearly seperable so it try to add a new dimension to the feature set let say z which depends on x and y as follows,

$$z = x^2 + y^2 \tag{7.1}$$

So the new data looks like Fig 7.4, and since it is linearly seperable we can find the suitable hyperplane to classify this new data.Since the data has 3 dimensions the hyperplane will be of 2 dimensions i.e., a plane. So when we see it from above the data

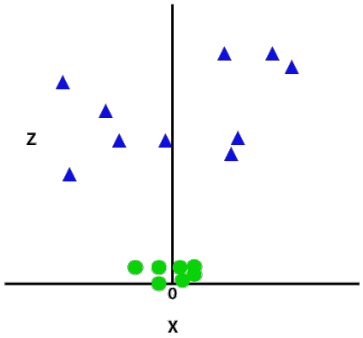

Fig. 7.4: Linear data after transforamation

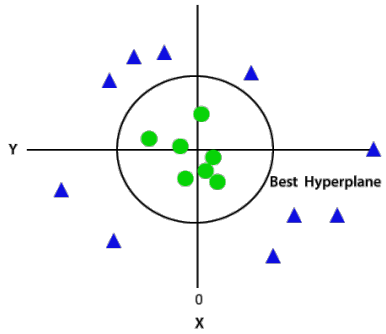

Fig. 7.5: Data is classified with a circle

will appeared to be classified with a curve as shown in fig 7.5. So as we saw in this example,in general we try to map the given feature dataset to a enlarged dataset and then classify it using our knowledge of linear SVM.

8 Questions

1) What is a hyperplane?
2) What are support vectors in SVM?
3) What is the difference between hard-margin and soft-margin SVM?
4) What is the role of C in soft-margin SVM?
5) What is a kernel trick?

9 Solutions

1) Hyperplane is a decision boundary used to classify the data into two classes. The dimensions of the hyper plane depends of the feature set of the dataset and it is 1 less than the dimensions of features.

2) Support vectors are the data points(of either class) which are close to the hyper plane.And these support vectors help in building the hyper plane.
3)
   - Hard-margin SVM is used for linear data whereas soft-margin SVM is used for non-linear data
   - In soft-margin we will allow some points to misclassify whereas in hard-margin, we will try classify the data correctly.
   - Soft-margin generalises better in case of outliers whereas the performance of hard-margin depends on the position of outlier(whether there are close to hyperplane or not)
4) The C is the tuning parameter for the soft-margin model, where
   - The smaller the value of C, the more sensitive the algorithm is to the training data.
   - The larger the value of C, the less sensitive the algorithm is to the training data.
5) The kernel technique is a method to transform a non-linear data into a higher dimensional space to make the data suitable for classification.It involves adding new features using the orginal feature set.