International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

Statistical Approach to Support Vector Machine

Yashima Ahuja, Sumit Kumar Yadav

Abstract— This paper provides you a brief description about support vector machine. In this, we have explained other universal approximators rather than layered feed-forward network and radial-basis function networks. Here, we have also explained mathematical concepts of support vector machine.

Index Terms— Dual Problem, Linearly Separable, Non-linear Separable, Primal Problem, Support Vector Machine

I. INTRODUCTION

Support Vector Machine is a machine learning tool which can be mainly used for classification and regression. It represents complex patterns and excludes bogus patterns. Classification involves grouping entities on the basis of some common properties or attributes. Support Vector machine is a linear machine with several nice properties. The core purpose of support vector machine is to separate the data with decision boundary in linear approach and extend it to non-linear approach by using kernel trick. It is used for many applications such as text categorization, face recognition, pattern recognition and many more.

Traditional approaches and techniques perform weakly in case of high dimension data. Support vector machine overcomes all that pitfalls of previous techniques. In this, we have mainly focused on mathematical concepts rather than on theoretical.

II. SUPPORT VECTOR MACHINE

Let's discuss the working of support vector machine in terms of mathematical expressions and by using cases of separable patterns. Main Idea of support vector machine is to construct a hyper plane as the decision surface that provides maximum margin of separation between positive and negative examples. A line is not optimal if it passes too close to the points because it will be noise sensitive and it will not simplify appropriately. Hence, our purpose should be able to find the line passing as far as possible from all points.

Support Vector Machine can provide a fine generalization performance on pattern classification problems although the reality is that it doesn't include problem- domain knowledge. This feature is unique to support vector machine.

Support Vector Machine utilize 2 Mathematical Cases for classification purposes-

A. Linear Separable Case

First case is the **linearly separable** case in which mapping of higher dimension into two categories is always separable by hyper plane which has a equation

$$w^T x + b = 0 \qquad (1)$$

where w is adjustable weight vector, b is bias and x is Input Vector. Hyper plane is a straight line that we are using for defining the maximum separation between two categories. In mathematical terms, we can also compare it with line equation $y = mx + b$ where m stands for Slope (means that how steep the line is) and b stands for y- intercept.

Support Vectors plays major role in the operation of this class of learning machines. These are those points that lie closest to the decision surface and that's why these are the most difficult ones to classify.

Suppose you have a training sample $\{x_i, d_i\}_{i=1}^{N}$, where x_i = Input Pattern for the i^{th} element and d_i = corresponding desired response. On the basis of desired response we can separate this two cases like d_i =+1 belongs to first case and d_i =-1 belongs to second case.

In mathematical Terms,

$$w^T x_i + b \geq 0 \qquad For\ d_i = +1$$
$$w^T x_i + b < 0 \qquad For\ d_i = -1$$
(2)

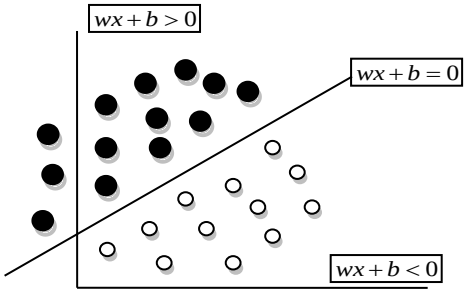

Fig 1: Illustration of Idea of Hyper plane

In General aspect Separation between hyper plane and closest data point is called margin of separation.

The Goal of support vector machine is to find the particular hyper plane for which margin of separation is maximized. Under this condition, decision surface is referred as optimal hyper plane which can be defined as

$$w_0^T x + b_0 = 0 \qquad (3)$$

The discriminant function gives a measure of the distance from x to the optimal hyper plane as written below:

Manuscript published on 28 February 2013.
* Correspondence Author (s)
Yashima Ahuja*, Computer Science and Engineering, Lovely Professional University, Phagwara, India.
Sumit Kumar Yadav, Computer Science and Engineering, Lovely Professional University, Phagwara, India.

$$g(x) = w_0^T x + b_0 \quad (4)$$

where $g(x)$ is a discriminant function

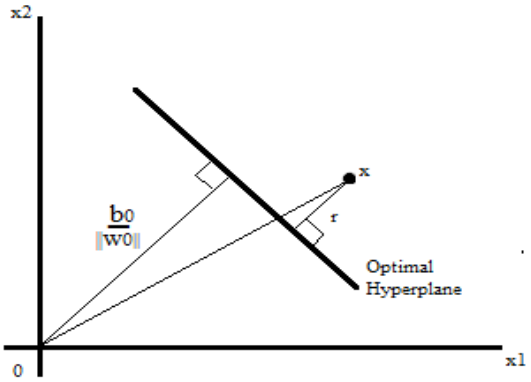

Fig 2: Geometric Interpretation of Algebric distance of points to the optimal hyper plane

We can also express this as

$$x = x_p + r\frac{w_o}{\|w_o\|}$$

Where x_p is the normal projection of x onto the optimal hyper plane and r is the desired algebraic distance.
On the basis of this, we can say that r is positive if x is on the positive side of the optimal hyper plane and negative if x is on the negative side.

Since by definition $g(x_p) = 0$, we get

$$g(x) = w_o^T x + b_o = r\|w_o\|$$

$$\text{So, } r = \frac{g(x)}{\|w_o\|} \quad (5)$$

In particular, the distance from the origin (i.e., x=0) to the optimal hyper plane is given by $b_o / \|w_o\|$. The issue is to find the parameters w_o and b_o for optimal hyper plane.

Consider a support vector $x^{(s)}$ for which $d^{(s)} = +1$. Then by definition, we have

$$g(x^{(s)}) = w_o^T x^{(s)} \mp b_o = \mp 1 \text{ , For } d^{(s)} = \mp 1 \quad (6)$$

Algebraic Distance from the support vector $x^{(s)}$ to the optimal hyper plane is

$$r = \frac{g(x^{(s)})}{\|w_o\|}$$

$$= \begin{cases} \frac{1}{\|w_o\|} & if \quad d^{(s)} = +1 \\ -\frac{1}{\|w_o\|} & if \quad d^{(s)} = -1 \end{cases} \quad (7)$$

where the plus sign indicated that $x^{(s)}$ lies on the positive side of the optimal hyper plane and the minus sign indicates that $x^{(s)}$ lies on the negative side of the optimal hyper plane.

Let ρ denote the optimum value of margin of separation between the two classes that constitute the training set τ.

We can also state it as $\rho = 2r = \frac{2}{\|w_o\|}$ (8)

It states that maximizing the margin of separation between subsets or classes is equivalent to minimizing the Euclidean norm of weight vector w.

It is very difficult to find the value of w_o and b_o by solving linear equation. So, we will find optimal hyper plane by using quadratic optimization.

Quadratic Optimization to uncover the optimal hyper plane-

Given training sample $\tau = \{x_i, d_i\}_{i=1}^N$ Find the optimal values of weight vector w and bias b such that they satisfy the constraints

$$d_i(w^T x_i + b) \geq 1 \quad \text{For } i = 1,2.....N \quad (9)$$

And the weight vector w minimized the cost function:

$$\Phi(w) = \frac{1}{2} w^T w$$

This scaling vector 1/2 is included here for convenience of presentation. This constrained optimization problem is called the **primal problem**. It is characterized as below:

- The cost function is a convex function of w.
- The constraints are linear in w.

Accordingly we may solve this by using method of Lagrange multiplier. Firstly let's see Lagrangian function:

$$J(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \left[d_i (w^T x_i + b) - 1 \right] \quad (10)$$

where the auxiliary nonnegative variables α_i is called Lagrange multiplier.

The key to solve the optimization problem is determined by the saddle point of the Lagrangian function $J = (w,b,\alpha)$ which has to be minimized with respect to w and b and maximized with respect to α. Thus by differentiating $J = (w,b,\alpha)$ with respect to w and b and setting the results equal to zero.

In mathematical terms,

Condition 1: $\frac{\partial J(w,b,\alpha)}{\partial w} = 0$

Condition 2: $\frac{\partial J(w,b,\alpha)}{\partial b} = 0$

After differentiating condition 1 it comes,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$w = \sum_{i=1}^{N} \alpha_i d_i x_i \qquad (11)$$

And from condition 2,

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad (12)$$

Though, this way out is exclusive by virtue of convexity of Lagrangian but the same cannot be said about the Lagrange coefficients.

It is also important to note that at the saddle point for each Lagrangian multiplier α_i, the product of that multiplier with its corresponding constraint vanishes as shown:

$$\alpha_i \left[d_i \left(w^T x_i + b \right) - 1 \right] = 0 \text{ For } i = 1,2,.....N \qquad (13)$$

Therefore only those multipliers exactly meeting equation (13) can assume non-zero values. This property follows from Kuhn-Tucker conditions of optimization theory.

Given such optimization problem, it is also possible to construct another problem called **dual problem**.

We can state the following Duality Theorem:

- If primal problem has an optimal solution and the dual problem has also an optimal solution then the corresponding optimal values are identical.
- In order for w_o to be an optimal primal solution and α_o to be an optimal dual solution, it is essential and adequate that w_o is feasible for primal problem and $\phi(w) = J(w_o, b_o, \alpha_o) = \min J(w, b_o, \alpha_o)$

To postulate dual problem for our primal problem, we will expand Lagrangian function as below:

$$J(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i w^T x_i - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i \qquad (14)$$

The third term on the right hand is zero by virtue of the optimality condition that we found in condition 2 of Lagrange function and from condition 1 we can have,

$$w^T w = \sum_{i=1}^{N} \alpha_i d_i w^T x_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j$$

By setting objective function $J(w,b,\alpha) = Q(\alpha)$, we may reformulate as:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j \qquad (15)$$

where α_i are non-negative.

Dual Problem can also be stated as:

Given training sample $\tau = \{x_i, d_i\}_{i=1}^{N}$ Find Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j$$

Subject to the constraints

1. $\sum_{i=1}^{N} \alpha_i d_i = 0$
2. $\alpha_i \geq 0$ For i=1,2,…,N

This was all about linear separable case. Let's discuss another case.

B. Non-Linearly Separable case

Second case is to handle **non-linear separable** data. In this section, it is not feasible to construct a separating hyper plane without encountering classification errors. For this, we need to find an optimal hyper plane that minimizes the probability of classification error.

The margin of separation between subsets or classes is assumed to be soft if a data point (x_i, d_i) breaks the following condition

$$d_i \left(w^T x_i + b \right) \geq +1 , \quad \text{For } i = 1,2,.....,N \qquad (16)$$

This can be done in two ways:

1. The data point (x_i, d_i) falls inside the region of separation but on the right side of the decision surface as shown in Fig 3.
2. The data point (x_i, d_i) falls on the wrong side of the decision surface as shown in Fig 4.

Here we can declare that we have accurate classification in case 1 but misclassification in case 2 shown in below figures:

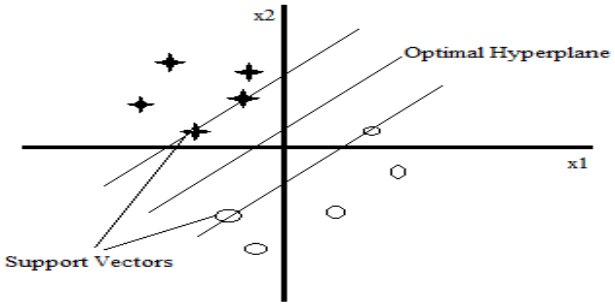

Fig 3: Data point (belongs to class C_1) falls inside the region of separation but on the right side of the decision surface.

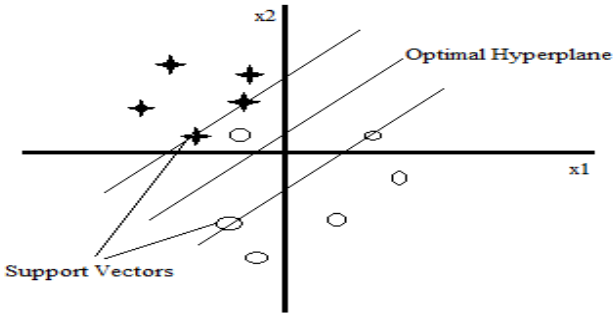

Figure 4: Data point (belongs to class C_2) falls on the wrong side of the decision surface)

We can also define for separating the hyper plane as

$$d_i \left(w^T x_i + b \right) \geq 1 - \varepsilon_i \quad \text{For i=`1, 2…, N} \qquad (17)$$

The ε_i called slack variables which help to measure the deviation of data point from the ideal condition of pattern separability.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

First case comes under $0 \le \varepsilon_i \le 1$ and Second case considered as $\varepsilon_i > 1$ .The support vectors are those particular data points that satisfy above equation even if $\varepsilon_i > 0$.

For finding optimal hyper plane, we can do this by –

$$\Phi(\varepsilon) = \sum_{i=1}^{N} I(\varepsilon_i - 1) \quad (18)$$

The function $\Phi(\varepsilon)$ is an indicator function defined by-

$$I(\varepsilon) = \begin{cases} 0 & if \varepsilon \le 0 \\ 1 & if \varepsilon > 0 \end{cases} \quad (19)$$

Minimization of $\Phi(\varepsilon)$ with respect to w is a nonconvex optimization problem that is NP-Complete. To trace it mathematically, we can also write it as

$$\Phi(\varepsilon) = \sum_{i=1}^{N} \varepsilon_i \quad (20)$$

Moreover we simplify the computation by formulating the functional to be minimized with respect to the weight vector w as follows:

$$\Phi(w, \varepsilon) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i \quad (21)$$

Primal Problem for non-separable case is:

Given training sample $\{(x_i, d_i)\}_{i=1}^{N}$ find the optimal values of the weight vector w and bias b such that they satisfy the constraint

$$d_i (w^T x_i + b) \ge 1 - \varepsilon_i \quad \text{For i=1, 2…, N}$$

$$\varepsilon_i \ge 0 \quad \text{For all i}$$

And such that weight vector w and the slack variables ε_i minimizes the cost functional

$$\Phi(w, \varepsilon) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i$$

where c is a user specified parameter. The parameter C controls the tradeoff between complexity of machine and the number of non-separable points. It can also be viewed as Regularization parameter.

Using method of Lagrange multipliers we can formulate **dual problem for non-separable patterns** as:

Given training sample $\tau = \{x_i, d_i\}_{i=1}^{N}$ Find Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j$$

Subject to the constraints

1. $\sum_{i=1}^{N} \alpha_i d_i = 0$
2. $0 \le \alpha \le C$ For i=1,2,….,N

Dual problem of non-separable case differs only in one constraint but it is the important difference.

The optimization problem for non-separable patterns includes the optimization problem for linearly separable patterns as a special case.

It can be done with the help of kernel technique. Main advantage of kernel trick is that it reduces computational overhead.

III. CONCLUSION

Support vector machine mainly focus on the study a classification of data points into two classes which can be further categorized as positives and negatives classes. The benefit of kernel functions extends support vector machine to nonlinear decision boundaries or surface, while keeping the computational ease of linear classification. In non-linear technique, use of slack variables allows the training data to include incorrectly labeled points. Support Vector Machines are used in many fields ranging from spam filters, face recognition, pattern classification, engineering, bioinformatics and medicine.

REFERENCES

[1] Simon Haykin, “Neural Networks- A Comprehensive Foundation”, Published by Pearson, Second Edition.

[2] Jiawei Han, Micheline Kamber and Jian Pei, “Data Mining- Concepts and Techniques” Published by Elsevier, Third Edition.

[3] Ingo Steinwart and Andreas Christmann, “Support Vector Machines”, Published by Springer 2008.

[4] Yashima Ahuja, Sumit Kumar Yadav, “Multiclass Classification and Support Vector Machine, *Published in Global Journals of Computer Science and Engineering (GJCST) December 2012, Volume 12, Issue 11, Version 1.0, pp 15-19.*

[5] Kristin P. Bennett, Colin Campbell, “Support Vector Machine: Hype or Hallelujah?” Published in SIGKDD Explor December 2000, Volume 2 and Issue 2, pp 1-13.

[6] Andre Guggenberger, “Another Introduction to Support Vector Machine”, Unpublished.

[7] Cunhe Li, Kangwei Liu, Hongxia Wang, “The incremental learning algorithm with support vector machine based on hyper plane-distance”, Published by Springer April 2009, pp 19-27.

[8] Leon Bottou, Chih-Jen Lin, “Support Vector Machine Solvers”, Unpublished.

[9] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin, “A Practical Guide to Support Vector Classification” pp 1-12.

[10] Tristan Fletcher, “Support Vector Machines Explained” Unpublished pp 1-19.

[11] Support Vector Machine, *Available on http://homepages.cae.wisc.edu/~ece539/spring00/notes/wordfile/svm.doc*

[12] Support Vector Machine, http://www.mathworks.in/help/bioinfo/ug/support-vector-machines-svm.html#bs3tbev-11

ABOUT AUTHORS:

Yashima Ahuja has completed B.Tech in Computer Science and engineering in 2011. Currently, she is pursuing her Master Degree (M.Tech) in Computer Science and Engineering from Lovely Professional University, Jalandhar. She has keen interest in Data Mining, Support vector machine, Classification Methods and some other mathematical concepts based on Data Mining. Her research is concentrated on Support Vector Machine and has also published some papers in International Journals related to same earlier.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Sumit Kumar Yadav has done his graduation (B.Tech) from UPTU in 2008 and achieved M.S degree in Computer Science and Engineering from IIIT, Hyderabad, India. Currently, He is working as Assistant Professor of Computer Science and Engineering in Lovely Professional University, Jalandhar. His research area includes Database, Data Mining and Data Warehouse. He has done a lot of research on this field and has published many papers in International Journals and Conferences.