International Journal of Pure and Applied Mathematics
Volume 87 No. 6 2013, 741-750
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v87i6.2

ijpam.eu

SVM TRADE-OFF BETWEEN MAXIMIZE THE MARGIN AND MINIMIZE THE VARIABLES USED FOR REGRESSION

M. Premalatha[1], C. Vijaya Lakshmi[2]

[1]Department of Mathematics
Sathyabama University, Chennai, INDIA
[2]Department of Mathematics
VIT University
Chennai, INDIA

Abstract: Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In classification problems generalization control is obtained by maximizing the margin, which corresponds to minimization of the weight vector. The minimization of the weight vector can be used in regression problems, with a loss function. The problem of classification for linearly separable data and introduces the concept of margin and the essence of SVM - margin maximization. In this paper gives the soft margin SVM introduces the idea of slack variables and the trade-off between maximizing the margin and minimizing the number of misclassified variables. A presentation of linear SVM followed by its extension to nonlinear SVM and SVM regression is then provided to give the basic mathematical details. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It also develops the concept of SVM technique can be used for regression. SVR attempts to minimize the generalization error bound so as to achieve generalized performance instead of minimizing the observed training error.

AMS Subject Classification: 62H30, 26A24, 90C51, 90C20

Received: September 6, 2013

Key Words: linear and non-linear classification, machine learning, SVM mathematical, SVM trade-off, SVM regression

1. Introduction

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. SVM is a useful technique for data classification. Classification in SVM is an example of Supervised Learning. A step in SVM classification involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction [1] [3]. Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. In the same manner as the non-linear SVC approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior, to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks.

2. Basic Machine Learning

Given a collection of data, a machine learner explains the underlying process that generated the data in a general and simple fashion.
Different learning paradigms:
Supervised *learning*
Unsupervised *learning*
Semi-supervised *learning*
Reinforcement *learning*

2.1. Supervised Learning

Supervised Learning: Classification. Each element in the sample is labeled as belonging to some class (e.g., apple or orange). The learner builds a model to predict classes for all input data. There is no order among classes [6], [5].

Supervised Learning: Regression. Each element in the sample is associated with one or more continuous variables. The learner builds a model to predict the value(s) for all input data. Unlike classes, values have an order among them [2], [4].

3. SVM Mathematically

Consider the problem of separating the set of training vectors belonging to the separate classes,

$$D = \{(x^1, y^1), (x^2, y^2), \dots, (x^n, y^n), x \in R^n, y \in \{-1, +1\}$$

with the hyperplane
$(w, x) + b = 0$
The set of vectors is said to be optimally separated by the hyper plane
$\min |(w, x^i) + b| = 1$
Where the parameters w, b

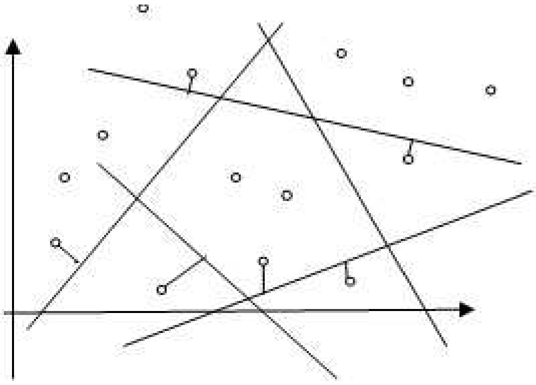

Figure 1: Hyper planes

The norm of the weight vector should be equal to the inverse of the distance, of the nearest point in the data set to the hyper plane [10], [7].

$$y^i[(w, x^i) + b] \geq 1 \quad i = 1, 2, \dots, n \tag{1}$$

The distance at a point x from the hyperplane

$$d = \frac{|(w, x^i) + b|}{||w||} \tag{2}$$

The optimal hyper plane is given by maximizing the margin, ρ, subject to the constraints of Equation (1)
The margin is given by,

$$\begin{aligned}\rho(w,b) &= \min_{x^i,y^i=-1} d(w,b;x^i) + \min_{x^i,y^i=1} d(w,b;x^i)\\ &= \min_{x^i,y^i=-1} \frac{|(w,x^i)+b|}{||w||} + \min_{x^i,y^i=1} \frac{|(w,x^i)+b|}{||w||}\\ &= \frac{1}{||w||}\left(\min_{x^i,y^i=-1} |(w,x^i)+b| + \min_{x^i,y^i=1} |(w,x^i)+b|\right)\\ &= \frac{2}{||w||}\end{aligned} \tag{3}$$

3.1. Support Vector Machine

Fix the empirical risk and minimize the VC confidence. SVM learns the best separating hyper plane. We should maximize the margin,

$$\rho(w,b) = \frac{2}{||w||}$$

Three main ideas: Define what an optimal hyper plane: maximize margin. Non-linearly separable problems: have a penalty term for misclassifications [9]. Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space

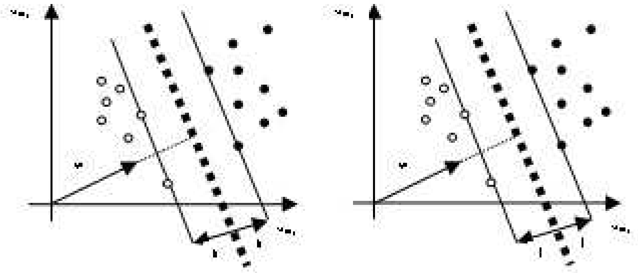

Figure 2: Support Vector Setting up to the Optimization Problem

The width of the margin is: $\rho(w,b) = \frac{2|k|}{||w||}$
So, the problem is max max $\frac{2|k|}{||w||}$

Subject to the Constraint
$(w.x+b) \geq k, \ \forall\ x$ of class 1
$(w.x+b) \leq -k, \ \forall\ x$ of class 2
There is a scale and unit for data so that $k=1$. Then problem becomes:
So, the problem is max $\frac{2}{||w||}$
Subject to the Constraint
$(w.x+b) \geq 1, \ \forall\ x$ of class 1
$(w.x+b) \leq -1, \ \forall\ x$ of class 2
If class 1 corresponding to 1 and class 2 corresponds to -1
$(w.x^i+b) \geq 1, \ \forall\ x$ with $y^i=1$
$(w.x^i+b) \leq -1, \ \forall\ x$ with $y^i=-1$
$y^i[(w,x^i)+b] \geq 1\ i=1,2,\ldots,n$
$y^i[(w,x^i)+b]-1 \geq 0\ i=1,2,\ldots,n$
So the problem becomes
Objective function of the maximization and minimization

max $\frac{2}{\lVert w \rVert}$	min $\frac{1}{2}\lVert w \rVert^2$
Subject to the constrain	Subject to the constraint
$y^i(w.x^i+b) \geq 1\ \forall\ x^i$	$y^i(w.x^i+b) \geq 1\ \forall\ x^i$
with $y^i=\pm 1$	with $y^i=\pm 1$

Now SVM formulation
min $\frac{1}{2}||w||^2$
Subject to the constraint
$y^i(w.x^i+b) \geq 1\ \forall\ x^i$
So, there is a unique global minimum value (when feasible). There is also a unique minimizer, i.e. weight and b value that provides the minimum. Non-solvable if the data is not linearly separable. Quadratic Programming Very efficient computationally with modern constraint optimization engines (handles thousands of constraints and training instances) [7]. There are theoretical
upper bounds on the error on unseen data for SVM
The larger the margin, the smaller the bound
The smaller the number of SV, the smaller the bound

3.2. Introduction of Slack Variables C Tradeoff

Objective function penalizes for misclassified instances and those within the margin

min $\frac{1}{2}||w||^2 + C\sum_i \xi_i$

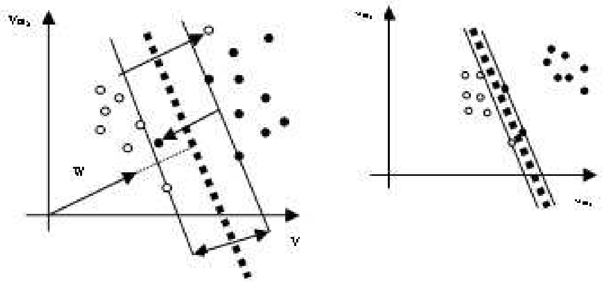

Figure 3: (a) C tradeoff margin width, (b) Hard margin SVM

C trades-off margin width and misclassifications.
Define $\xi_i = 0$ if there is no error for x_i
ξ_i Are just "slack variables"
Subject to the constraint
$y^i(w.x^i + b) \geq 1 - \xi_i \;\; \forall \; x^i y^i = 1$
$y^i(w.x^i + b) \leq -1 + \xi_i \; \forall \; x^i y^i = -1$
$\xi_i \geq 0 \;\; \forall \; x^i$
Subject to the constraint
$y^i(w.x^i + b) \geq 1 - \xi_i \; \forall \; x^i$
$\xi_i \geq 0$

Algorithm tries to maintain ξ_i to zero while maximizing margin, algorithm does not minimize the number of mis classifications, but the sum of distances from the margin hyper planes. Other formulations use ξ_i^2 instead. As $C \to \infty$, we get closer to the hard-margin solution. Soft-Margin always has a solution. Hard-Margin does not require guessing the cost parameter.

In the limit, $C \to \infty$ the solution converges toward the solution obtained by the optimal separating hyper plane In the limit, $C \to 0$, the solution converges to one where the margin maximisation term dominates, there is now less emphasis on minimizing the misclassification error, but purely on maximising the margin, producing a large width margin. Consequently as C decreases the width of the margin increases. The useful range of C lies between the point where all the Lagrange Multipliers are equal to C and when only one of them is just bounded by C [8].

4. Support Vector Regression – Linear Regression

Consider the problem of approximating the set of data,

$$D = \{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n), \quad x \in R^n, y \in R$$

$$\text{with a linear function} \tag{4}$$

$$(w, x) + b = 0$$

$$f(x) = w.x + b \tag{5}$$

Regression function is given by the minimum of the function.

$$\phi(w, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i^- + \xi_i^+) \tag{6}$$

Where C is pre-specified value and $\xi_i^- + \xi_i^+$ is slack representing upper and lower constraints on the outputs of the system.

Table 1: Classification Data's

Linear separable Classification Data			Non-Linear Separable Classification Data		
X1	X2	Y	X1	X2	y
1	1	-1	1	1	-1
3	3	1	3	3	1
1	3	1	1	3	1
3	1	-1	3	1	-1
2	3	1	2	2.5	1
3	2.5	-1	3	2.5	-1
4	3	-1	4	3	-1
4	3.5	1	1.5	1.5	1
2	1	-1	1	2	-1
			1.5	2.5	-1

4.1. Regression Loss Function

$$L_\epsilon(y) = \{0 \ \text{ for } |f(x) - y| < \epsilon|f(x) - y| - \epsilon; \ \text{otherwise} \tag{7}$$

The solution is given by

$$\overline{\alpha}, \overline{\alpha^*} = \arg\min_{\alpha,\alpha^*} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i + \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)\epsilon \tag{8}$$

With constraints

$$(\alpha_i + \alpha_i^*) \leq C; \quad (\alpha_i + \alpha_i^*) \geq 0 \tag{9}$$

$$\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) = 0$$

Solving equation (8), (9) determines the Lagrange multipliers $(\alpha_i + \alpha_i^*)$ and the regression function is given by the equation (4) where

$$\overline{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i \tag{10}$$

$$b^* = -\frac{1}{2}(\overline{w}, (x_l + x_m))$$

Using a quadratic loss function

$$L_{quad}(f(x) - y) = (f(x) - y)^2 \tag{11}$$

$$\max_{\alpha,\alpha^*} W = \max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \frac{1}{2c}\sum_{i=1}^{n}(\alpha_i^2 + (\alpha_i^*)^2) \tag{12}$$

Regression function is given by Equation (4) and (10).

5. Conclusion

Support Vector Machines are an attractive approach to data modelling. They combine generalization control with a technique to address the curse of dimensionality. The formulation results in a global quadratic optimization problem with box constraints. SVM are trained by solving a constrained quadratic optimization problem. SVM, implements mapping of inputs onto a high dimensional space using a set of nonlinear basis functions. In removing the training patterns that are not support vectors, the solution is unchanged and hence a fast method for validation may be available when the support vectors are sparse.

References

[1] A. Ben-Hur, D. Horn, H.T. Siegelmann and V. Vapnik, Support vector clustering, *Journal of Machine Learning Research*, **2** (2001), 125–137.

[2] C.M. Bishop, Pattern Recognition and Machine Learning, *Information Science and Statistics*, Springer (2006).

[3] N. Cristianini, J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*, Cambridge University Press, New York, NY, USA (2000).

[4] J.B. Gao, S.R. Gunn, and C.J. Harris, SVM Regression through Variational Methods and its Sequential Implementation, *Neurocomputing*, **55** (2003), 151–167.

[5] M.O. Stitson and J.A.E. Weston, Implementation issues of support vector machines, *Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London* (1996).

[6] D.M.J. Tax and R.P.W. Duin, Support vector domain description, *Pattern Recognition Letters*, **20**(1113) (1999), 1191–1199.

[7] V. Vapnik, S. Golowich and A. Smola, Support vector method for function approximation, regression estimation, and signal processing, In *M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems*, Cambridge, MA, 1997. MIT Press, Vol. 9 (1997), 281–287,

[8] C.W. Yap and Y.Z. Chen, Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines, *J. ChemInf. Model.*, **45** (2005), 982–992.

[9] Y.Q. Zhan and D.G. Shen, Design Efficient Support Vector Machine for Fast Classification, *Pattern Recognition*, **38** (2005), 157–161.

[10] Zweiri, Yahya H, Optimization of a three Back propagation Algorithm used for Neural Network learning, *International Journal of Computational Intelligence* **3**(4) (2007), 322–327.