2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) | 978-1-6654-9991-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAICA54878.2022.9844516

Support Vector Machine Algorithm in Machine Learning

Qiyu Wang
School of Mathematics, Faculty of Sciences, University of Nottingham, Nottingham, NG7 2RD, UK
ssyqw2@nottingham.ac.uk

***Abstract*—The Support Vector methods was proposed by V.Vapnik in 1965, when he was trying to solve problems in pattern recognition. In 1971, Kimeldorf proposed a method of constructing kernel space based on support vectors. In 1990s, V.Vapnik formally introduced the Support Vector Machine (SVM) methods in Statistical Learning. Since then, SVM has been widely applied in pattern recognition, natural language process and so on. Informally, SVM is a binary classifier. The model is based on the linear classifier with the optimal margin in the feature space and thus the learning strategy is to maximize the margin, which can be transformed into a convex quadratic programming problem. It uses the principle of structural risk minimization instead of empirical risk minimization to fit small data samples. Kernel trick is used to transform non-linear sample space into linear space, decreasing the complexity of algorithm. Even though, it still has broader prospects for development.**

Keywords—Machine Learning, Statistical Learning, Classification, Support Vector Machine

I. Introduction

Support Vector Machine has many applications such as text classification, image classification, biological sequence analysis and biological data mining, handwritten character recognition. It was firstly mentioned in Professor Vapnik's paper. Support Vector Machines (SVM): The input vectors x are mapped non-linearly into a high-dimensional feature space Z by choosing a priori, where a linear decision surface is built. The decision surface with some certain properties satisfies high generalization ability of the network (Vapnik and Cortes, 1995) [1-10]. The decision surface is also called as hyperplane, which is defined as a linear decision boundary between two classes with maximal margin (i.e., Everything on one side belongs to one class). Support Vectors are defined as a set of points or vectors o the margin of the separating hyperplane. Margin means finding out the closest point to the hyperplane and guaranteeing it to be as far away from the separating line.

The ideas of Support Vector Machine were firstly proposed in the framework for pattern recognition in 1962, which was called 'Generalized Portrait Method'. This algorithm was firstly published in 1964 [1-10].Given a certain Euclidian, unit vectors $\mathbf{x}\ (|\boldsymbol{x}| = \mathbf{1})$ represent the patterns and hence pictures are the subset of a unit sphere. The heuristic idea is that vectors or patterns may not be unit vectors in real world. The 'generalized portrait' of a class **S** is defined as the unit vector φ which satisfies:

$$\max_{\varphi} \min_{x \in S} (\varphi, x) = c. \qquad (1)$$

It can be interpreted that the vector φ should be closest to the vectors in the set **S** that most distant from φ and we hope $c > 0$.

II. Main Works

Given certain fixed restrictions, we try to minimize some collinear vector ψ instead of maximizing a fixed normal vector ϕ 's scalar product with marginal vectors. An equivalent statement is that given the restrictions $(\psi, x) \geq 1$. Find a vector ψ with minimum norm. Then we can find

$$\phi = \psi / ||\psi|| \qquad (2)$$

Considering the prime idea that we hope to find a unit vector ϕ which satisfies:

$$\max_{\varphi} \min_{x \in S} (\varphi, x) = c \qquad (3)$$

Based on the condition that $(\phi, x) \leq kc\ for\ all\ vectors\ x\ of\ the\ class\ S_1$, where $0 < k < 1$.

Moreover, we need to find the smallest volume segment of the sphere containing all class S_0's vectors and does not include S_1's vectors. We call the ϕ 'generalized portrait' of S_0 against S_1.

The problem can be transformed as a quadratic programming problem:

$$min(\psi, \psi).$$

given the conditions $(\psi, x) \geq 1, for\ \forall x \in S_0, (\psi, Y) \leq k, for\ \forall y \in S_1$ (4)

We try to find out the 'generalized portrait' using the decomposition of vectors of the training sequence instead of coordinators. We are looking for $\psi = \Sigma a_i x_i$. The function (ψ, ψ) can be written as

$$(\psi, \psi) = (\psi, \textstyle\sum_i a_i\, x_i) = \sum_{(i,j)} a_i a_j (x_i, x_j) \qquad (5)$$

The constraints are given

$$(\psi, x_j) = (\textstyle\sum_i a_i x_i\,, x_j) = \sum_i a_i (x_i, x_j) \qquad (6)$$

Thus, we can simply use the scalar product of the initial vectors and the coefficients of the decomposition of ψ. This method was firstly used for computation in analogue computers. Since digital computers emerged, people got used to representing data in the form of coordinates. However, V. Vapnik used kernel tricks to reuse this method. This is the most distinct difference between generalized portrait method and support vector machine algorithm. We firstly introduce the 'Jackknife' method. An object is deleted from the training sequence, and it constructs a decision rule, and the ignored object is tested. It repeats such procedure to other objects and computes the rate of correct and errors of the objects. Such method makes the generalization properties become obvious in the generalized portrait method. It will be an unbiased estimation theoretically.

The ignored vector can be accurately recognized, and the generalized portrait does not change if a non-marginal vector is removed from the training sequence. Hence the error rate is smaller than the share or support vectors belong to the training sequence. Moreover, if the class of support vectors is linearly independent, then there is a unique decomposition of the generalized portrait. If not, there exists some decompositions where the coefficients be non-negative. The support vectors which are called 'informative' cannot have a zero coefficient. The number of errors made by informative vectors is always greater than the 'Jackknife' method made, and the number of informative vectors is always less than the dimension of the parameter space. Since the 'Jackknife' estimation is unbiased, we can conclude that the error expectation for a test for linear separable classes and the generalized portrait method cannot exceed n/(l+1), say $E_{error} \leq n/(l+1)$, where n is the dimension of the space and l is the length of the training sequence. The number of support vectors for the optimal hyperplane often is larger than the generalized portrait for other constants k. Finding a value for the constant to provide the least number of support vectors as well as using the generalized portrait method seems to be reasonable even though V. Vapnik suggested only using the optimal hyperplane. Support vector machine is quite different from the generalized portrait method since there is no coordinate representation being used. Decomposition is the technique used in the support vectors and scalar products presented in kernel form. Its sill shares the generalization properties since the number of support vectors is quite small. Even though SVM method has been widely applied in face detection, text categorization and pedestrian detection in recent years, scientists did not explore the use of the SVM in the cure rate models deeply. In the paper A support Vector Machine Based Cure Rate Model for Interval Censored Data. They constructed a new cure rate model which used SVM algorithm to model the incidence part and a proportional hazards (PH) to model the latency part of survival data. This new cure rate model can still have traditional SVM's properties, and it can process more complicated classification boundaries. The standard errors were estimated by the non-parametric bootstrapping since the model's complexity. It illustrates that given a non-linear boundary, the SVM model performed better than the traditional logistic regression model. These results made sense for both incidence and latency sections.

Cconsidering a binary classification problem with labels y and features x. We will use $y \in -1,1$ to denote the class labels (i.e., the label for every data on one side is 1 and the other side is -1). Parameters w, b are used and the classifier can be written as:

$$h_{w,b}(x) = g(w^T x + b) \tag{7}$$

If $z \geq 0, g(z) = 1, g(z) = -1$

Given a training example $\left(x^{(i)}, y^{(i)}\right)$, functional margin of (w, b) related to the training example as

$$\hat{\gamma}^{(i)} = y^{(i)}\left(w^T x^{(i)} + b\right) \tag{8}$$

If $y^{(i)} = 1$, then we need $w^T x^{(i)} + b$ to be a large positive number to ensure the functional margin to be large. If $y^{(i)}\left(w^T x^{(i)} + b\right) > 0$, then the prediction on this example is correct. Therefore, a large functional margin can indicate a confident and correct prediction. Consider a point $A\left(x^{(i)}, y^{(i)}\right)$, where $y^{(i)} = 1$. $\gamma^{(i)}$ denotes the distance between A to the hyperplane (Line AB). Since $w/\|w\|$ is a unit vector and A represents $x^{(i)}$, we can derive that B represents $x^{(i)} - \gamma^{(\mathrm{i})} \cdot w/\|w\|$. B is on the hyperplane; therefore, B satisfies

$$w^T\left(x^{(i)} - \gamma^{(\mathrm{i})} \cdot w/\|w\|\right) + b = 0 \tag{9}$$

Thus

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{\|w\|} = \left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|} \tag{10}$$

Generally, geometric margin of (w, b) in terms of a training example $\left(x^{(i)}, y^{(i)}\right)$ is defined as

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}\right) \tag{11}$$

If $\|w\| = 1$, the function margin equals the geometric margin. Meanwhile, the geometric margin is invariant to rescaling the parameters (i.e., $w \to 2w, b \to 2b$, the geometric margin does not change). Given a training set

$$S = \left(x^{(i)}, y^{(i)}\right): i = 1, \ldots, n \tag{12}$$

the geometric margin can also be defined as

$$\gamma = \min_{i = 1,\ldots,n} \gamma^{(i)} \tag{13}$$

Assuming that there is a linearly separable training set, we now seek to find out the maximum geometric margin. The function margin is the same as the geometric margin, all the geometric margins are at least γ. Therefore, we can get the largest geometric margin if we solve this optimization problem. We are now trying to find out the maximal value.

Since that the change of scaling of the constraints w and b does not change anything, we set the scaling to be 1: $\hat{\gamma} = 1$. We now transform the original problem into another optimization problem.

$$\min_{w,b}\frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, i = 1, \ldots, n \tag{14}$$

This is a problem which only have linear constraints and a convex quadratic objective. The optimal margin classifier is given by solving this problem.

A. Lagrange duality

Define Lagrange as

$$\mathcal{L}(w,b) = f(w) + \sum_{i=1}^{l}\beta_i h_i(w) \tag{15}$$

here, β_i is called as the Lagrange multiplier. Then set $\frac{\partial\mathcal{L}}{\partial w_i} = 0\ and\ \frac{\partial\mathcal{L}}{\partial\beta_i} = 0$ and solve w and β.

Consider the problem: $\min_{w} fws.t.giw \leq 0, i = 1, \ldots, k\ and\ hiw = 0, i = 1, \ldots,$

To solve it, we define the $generalized\ Lagrangian$.

$$\mathcal{L}(w,a,b) = f(w) + \sum_{i=1}^{k}\alpha_i g_i(w) + \sum_{i=1}^{l}\beta_i h_i(w) \tag{16}$$

Here, the Lagrange multipliers are the $\alpha_i\ and\ \beta_i$. Consider the quantity

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i\geq 0}\mathcal{L}(w,\alpha,\beta) \tag{17}$$

Where, $\mathcal{P}$ denotes the 'primal'. Then,

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w)\ , if\ w\ satisfies\ primal\ constraints \\ \quad infinity, otherwise \end{cases} \tag{18}$$

$\theta_{\mathcal{P}}$ takes certain value if there exists w satisfying primal constraints and it would be positive infinity otherwise.

Now consider a different question:

$$\theta_{\mathcal{D}}(\alpha,\beta) = min_w\mathcal{L}(w,\alpha,\beta) \tag{19}$$

$\mathcal{D}$ denotes 'dual'.

The $dual$ optimization problem:

$$\max_{\alpha,\beta:\alpha_i\geq 0}\theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} min_w\mathcal{L}(w,\alpha,\beta) \tag{20}$$

This is the same problem as the primal problem except that now we are trying to find a minimum value. Define the value which is the optimal solution of the dual problem to be $d^* = \max_{\alpha,\beta:\alpha_i\geq 0}\theta_{\mathcal{D}}(w)$. The relationship between the primal problem and the dual problem is that

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} min_w\mathcal{L}(w,\alpha,\beta) \leq \min_{w}\max_{\alpha,\beta:\alpha_i\geq 0}\mathcal{L}(w,\alpha,\beta) = p^* \tag{21}$$

Suppose f and the g_i's are convex and feasible and the h_i's are affine. There exist w^*, α^*, β^* such that w^* is the solution to the primal problem and α^*, β^* are the solution to the dual problem. Besides, w^*, α^*, β^* satisfy the $Karush - Kuhn - Tucker\ (KKT)\ conditions$

KKT conditions:

$$\boldsymbol{f_i(\tilde{x}) \leq 0, i = 1, \ldots, n} \tag{22}$$

$$\boldsymbol{h_i(\tilde{x}) = 0, i = 1, \ldots, n}$$

$$\boldsymbol{\tilde{\lambda}_i \geq 0, i = 1, \ldots, n}$$

$$\boldsymbol{\tilde{\lambda}_i f_i(\tilde{x}) = 0, i = 1, \ldots, n}$$

$$\boldsymbol{\nabla f_0(\tilde{x}) + \sum_{i=1}^{m}\lambda_i\nabla f_i(\tilde{x}) + \sum_{i=1}^{p}\tilde{v}_i\nabla h_i(\tilde{x}) = 0}$$

Then $\tilde{x}$ and $(\bar{\lambda},\bar{v})$ are primal and dual optimal and there is a zero-duality gap.

According to the last condition, the gradient equals zero as $x = \tilde{x}$. Thus $\tilde{x}$ is the minimum of $\mathcal{L}(x,\bar{\lambda},\bar{v})$. Based on the first and second condition, we conclude that $\tilde{x}$ is primal feasible. Therefore,

$$g(\bar{\lambda},\bar{v}) = \mathcal{L}(x,\bar{\lambda},\bar{v}) = f_0(\tilde{x}) + \sum_{i=1}^{m}\bar{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^{p}\bar{v}_i h_i(\tilde{x}) = f_0(\tilde{x}) \tag{23}$$

This indicates that $\tilde{x}$ and $(\bar{\lambda},\bar{v})$ have a zero-duality gap.

Now we back to the SVM algorithm. Consider the problem:

$$\min_{w,b}\frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}\left(w^Tx^{(i)}+b\right)\geq 1, i=1,\ldots,n \qquad (24)$$

We need to solve this problem to find the maximum margin with respect to $f(x)=w^Tx+b$

The Lagrange Multiplier method is applied, that is we add a multiplier $\alpha_i\geq 0$, then the problem can be written as

$$\mathcal{L}(w,a,b)=\tfrac{1}{2}\|w\|^2+\textstyle\sum_{i=1}^{m}\alpha_i(1-y^{(i)}(w^Tx^{(i)}+b)) \qquad (25)$$

Take the partial derivative with w, b respectively, and let them be 0 i.e. $\frac{\partial\mathcal{L}(w,a,b)}{\partial w}=0$, $\frac{\partial\mathcal{L}(w,a,b)}{\partial b}=0$. We can get

$$w=\textstyle\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x_i} \qquad (26)$$

$$0=\textstyle\sum_{i=1}^{m}\alpha_i\,y_i \qquad (27)$$

Take the two equations back to (25) and consider the constraints provided by (27), we can derive the dual problem

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{I=1}^{M}\sum_{J=1}^{M}\alpha_i\alpha_j y_i y_j x_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{j}$$

$$\text{s.t. }\sum\nolimits_{i=1}^{m}\alpha_i\,y_i\ =0,\quad \alpha_{\mathrm{i}}\geq 0, \mathrm{i}=1,\ldots,\mathrm{n} \qquad (28)$$

Then, we can derive the model $\mathrm{f}(\mathbf{x})=\mathbf{w}^{\mathbf{T}}\mathrm{x}+\mathrm{b}=\sum_{\mathrm{i}=1}^{\mathrm{m}}\alpha_i y_i x_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{x}$ The computation above strictly follows the KTT condition.

(25) is a quadratic programming problem, there is a effective algorithm to solve such problem and that is SMO (Sequential Minimal Optimization) algorithm. The basic idea for SVO is that all the parameters are fixed except α_i, and find the extreme value of α_i. Since the constraint $\sum_{i=1}^{m}\alpha_i\,y_i\ =0$, α_i can be derived by other parameters. Hence, we can only set variables α_i and α_{j} and run the SVO algorithm until it converges. Considering a binary classification problem with labels y and features x. We will use $y\in\{-1,1\}$ to denote the class labels (i.e., the label for every data on one side is 1 and the other side is -1). Parameters w, b are used and the classifier can be written as:

$h_{w,b}(x)\ =\ g(w^Tx\ +b)$. If $z\geq 0, g(z)=1,$ and $g(z)=-1$ otherwise (29)

Given a training example $\left(x^{(i)},y^{(i)}\right)$, functional margin of (w,b) related to the training example as $\hat{\gamma}^{(i)}=y^{(i)}\left(w^Tx^{(i)}+b\right)$. If $y^{(i)}=1$, then we need $w^Tx^{(i)}+b$ to be a large positive number to ensure the functional margin to be large. If $y^{(i)}\left(w^Tx^{(i)}+b\right)>0$, then the prediction on this example is correct. Therefore, a large functional margin can indicate a confident and correct prediction. Consider a point $A\left(x^{(i)},y^{(i)}\right)$, where $y^{(i)}=1$. $\gamma^{(i)}$ denotes the distance between A to the hyperplane (Line AB). Since $w/\|w\|$ is a unit vector and A represents $x^{(i)}$, we can derive that B represents $x^{(i)}-\gamma^{(\mathrm{i})}\cdot w/\|w\|$. B is on the hyperplane; therefore, B satisfies $w^T\left(x^{(i)}-\gamma^{(\mathrm{i})}\cdot w/\|w\|\right)+b=0$.

Generally, geometric margin of (w,b) in terms of a training example $\left(x^{(i)},y^{(i)}\right)$ is defined as

$$\gamma^{(i)}=y^{(i)}\left(\left(\tfrac{w}{\|w\|}\right)^Tx^{(i)}+\tfrac{b}{\|w\|}\right) \qquad (30)$$

If $\|w\|\ =\ 1$, the function margin equals the geometric margin. Meanwhile, the geometric margin is invariant to rescaling the parameters (i.e., $w\rightarrow 2w, b\rightarrow 2b$, the geometric margin does not change).

Given a training set $S=\left(x^{(i)},y^{(i)}\right): i=1,\ldots,n$, the geometric margin can also be defined as

$$\gamma=\ \min_{i\ =\ 1,\ldots,n}\gamma^{(i)} \qquad (31)$$

Assuming that there is a linearly separable training set, we now seek to find out the maximum geometric margin, i.e., $\max_{\gamma,w,b}\gamma$ s.t. $y^{(i)}\left(w^Tx^{(i)}+b\right)\geq\gamma, i=1,\ldots,n\ and\ \|w\|=1$.

Since $\|w\|=1$, the function margin is the same as the geometric margin, all the geometric margins are at least γ. Therefore, we can get the largest geometric margin if we solve this optimization problem.

Consider: $\max_{\gamma,w,b}\hat{\gamma}/\|w\|$ s.t. $y^{(i)}\left(w^Tx^{(i)}+b\right)\geq\hat{\gamma}, i=1,\ldots,n$. We are now trying to find out the maximal value of $\frac{\hat{\gamma}}{\|w\|}$. Since that the change of scaling of the constraints w and b does not change anything, we set the scaling to be 1: $\hat{\gamma} = 1$. Thus, we are calculating $\max_{\gamma,w,b} \frac{1}{\|w\|}$, which is equivalent to finding out $\min_{w,b} \|w\|^2$. We now transform the original problem into another optimization problem.

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, i = 1, \ldots, n \tag{32}$$

This is a problem which only have linear constraints and a convex quadratic objective. The optimal margin classifier is given by solving this problem.

Lagrange duality

Define $Lagrangian$ as

$$\mathcal{L}(w,b) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w) \tag{33}$$

here, β_i is called as the $Lagrange\ multipliers$. Then set $\frac{\partial \mathcal{L}}{\partial w_i} = 0\ and\ \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$ and solve w and β

Consider the problem:

$$\min_{w} fws.t.\,giw \leq 0, i = 1, \ldots, k\ and\ hiw = 0, i = 1, \ldots, \tag{34}$$

To solve it, we define the $generalized\ Lagrangian$

$$\mathcal{L}(w,a,b) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) \tag{35}$$

Here, the Lagrange multipliers are the $\alpha_i\ and\ \beta_i$.

Consider the quantity $\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta)$. $\mathcal{P}$ denotes the 'primal'.

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w)\ , if\ w\ satisfies\ primal\ constraints \\ infinity, otherwise \end{cases} \tag{36}$$

$\theta_{\mathcal{P}}$ takes certain value if there exists w satisfying primal constraints and it would be positive infinity otherwise.

Now consider a different question:

$$\theta_{\mathcal{D}}(\alpha,\beta) = min_w \mathcal{L}(w,\alpha,\beta) \tag{37}$$

$\mathcal{D}$ denotes 'dual'.

The $dual$ optimization problem:
$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} min_w \mathcal{L}(w,\alpha,\beta)$

This is the same problem as the primal problem except that now we are trying to find a minimum value. Define the value which is the optimal solution of the dual problem to be $d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \theta_{\mathcal{D}}(w)$. The relationship between the primal problem and the dual problem is that

$$d^* = \max_{\alpha,\beta:\alpha_i \geq 0} min_w \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^* \tag{38}$$

Suppose f and the g_i's are convex and feasible and the h_i's are affine. There exist w^*, α^*, β^* such that w^* is the solution to the primal problem and α^*, β^* are the solution to the dual problem. Besides, w^*, α^*, β^* satisfy the $Karush - Kuhn - Tucker\ (KKT)\ conditions$.

KKT conditions:

$$\boldsymbol{f_i(\tilde{x}) \leq 0, i = 1, \ldots, n} \tag{39}$$

$$\boldsymbol{h_i(\tilde{x}) = 0, i = 1, \ldots, n}$$

$$\boldsymbol{\tilde{\lambda}_i \geq 0, i = 1, \ldots, n}$$

$$\boldsymbol{\tilde{\lambda}_i f_i(\tilde{x}) = 0, i = 1, \ldots, n}$$

$$\boldsymbol{\nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{v}_i \nabla h_i(\tilde{x}) = 0}.$$

Then $\tilde{x}$ and $(\bar{\lambda}, \bar{v})$ are primal and dual optimal and there is a zero-duality gap. According to the last condition, the gradient equals zero as $x = \tilde{x}$. Thus $\tilde{x}$ is the minimum of $\mathcal{L}(x, \bar{\lambda}, \bar{v})$. Based on the first and second condition, we conclude that $\tilde{x}$ is primal feasible. Therefore,

$$g(\bar{\lambda}, \bar{v}) = \mathcal{L}(x, \bar{\lambda}, \bar{v}) = f_0(\tilde{x}) + \sum_{i=1}^{m} \bar{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^{p} \bar{v}_i h_i(\tilde{x}) = f_0(\tilde{x}) \tag{40}$$

This indicates that $\tilde{x}$ and $(\bar{\lambda}, \bar{v})$ have a zero-duality gap.

Now we back to the SVM algorithm. Consider the problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, i = 1, \ldots, n \tag{41}$$

We need to solve this problem to find the maximum margin with respect to $f(x) = w^T x + b$

The Lagrange Multiplier method is applied, that is we add a multiplier $\alpha_i \geq 0$, then the problem can be written as

$$\mathcal{L}(w, a, b) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y^{(i)}(w^T x^{(i)} + b)) \tag{42}$$

Take the partial derivative with w, b respectively, and let them be 0

We can get

$$w = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} \tag{43}$$

$$0 = \sum_{i=1}^{m} \alpha_i \, y_i \tag{44}$$

Take the two equations back to (15) and consider the constraints provided by (17), we can derive the dual problem

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{I=1}^{M} \sum_{J=1}^{M} \alpha_i \alpha_j y_i y_j x_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{j}$$

$$\text{s.t.} \sum_{i=1}^{m} \alpha_i \, y_i \; = 0, \tag{45}$$

Then we can derive the model

$$\mathrm{f}(\mathbf{x}) = \mathbf{w^T}\mathrm{x} + \mathrm{b} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_i y_i x_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{x} \tag{46}$$

The computation above strictly follows the KTT condition.

(43) is a quadratic programming problem, there is an effective algorithm to solve such problem and that is SMO (Sequential Minimal Optimization) algorithm. The basic idea for SVO is that all the parameters are fixed except α_i, and find the extreme value of α_i . Since the constraint $\sum_{i=1}^{m} \alpha_i \, y_i \; = 0$, α_i can be derived by other parameters. Hence, we can only set variables α_i and α_j and run the SVO algorithm until it converges.

We now introduce kernels briefly.

Definition: Given an input space X, let H be the Hilbert space. If there exists a mapping from

X to H

ϕ(x): X→H for any x,z∈X,the function K(x,z) satisfies K(x,z)=ϕ(x)ϕ(z). Then, K(x,z) is the kernel function and ϕ(x) is the mapping funtion.

A non-linearly separable input space can be mapped into a linear separable space by kernel methods. Typically, this space is higher-dimensional and likely to be infinite. The distance between two elements is returned rather than mapping features to this space and the implicit mapping is called as the Kernel Trick.

III. Conclusion

PROS: 1)A few support vectors determine the final decision function, and the complexity of computation depends on the number of support vectors instead of the dimension of the sample space, avoiding the 'dimension disaster' efficiently.

2) The kernel tricks are introduced, which can help solve nonlinear regression and nonlinear classification problems.

3) It can be proofed and interpreted by rigorous mathematics such as Lagrange dual optimization, and does not rely on statistical methods, simplifying classification and regression problems.

4) It can identify the key samples (i.e., the support vectors) which are critical to problem.

CONS:

1) Since the kernel tricks are used, the kernel matrix needs to be stored and the spatial complexity is $O(n^2)$.

2) The training time is longer and since a couple of parameters should be selected during each time, the time complexity is $O(N^2)$ where N is the number of training samples.

3) SVM is only suitable for tasks with small batch samples and cannot applied to tasks with millions of samples because the prediction time is proportional to the number of vectors.

Future:

There still exists a broader prospect for future research and development.

References

[1] Vapnik, V.N. (1995) The nature of statistical learning theory. Springer.

[2] B. Schölkopf A. Smola and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," Neural Computation, Vol. 10, 1998, pp. 1299-1319.

[3] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.

[4] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P Numerical Recipes: The Art of Scientific Computing 3rd. New York: Cambridge University Press. 2007 [2016-11-06]. ISBN 978-0-521-88068-8.

[5] Barghout, Lauren. "Spatial-Taxon Information Granules as Used in Iterative Fuzzy-Decision-Making for Image Segmentation." Granular Computing and Decision-Making. Springer International Publishing, 2015. 285-318.

[6] R. Cuingnet, C. Rosso, M. Chupin, S. Lehéricy, D. Dormont, H. Benali, Y. Samson and O. Colliot, Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome, Medical Image Analysis, 2011, 15 (5): 729–737

[7] Statnikov, A., Hardin, D., & Aliferis, C. (2006). Using SVM weight-based methods to identify causally relevant and non-causally relevant variables. sign, 1, 4.

[8] Ferris, M. C.; Munson, T. S. Interior-Point Methods for Massive Support Vector Machines. SIAM Journal on Optimization. 2002, 13 (3): 783.

[9] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. 1992: 144.

[10] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.