SVMs

20 April 2018 23:35

SVM:

Motivation:
Assume a **linearly separable training set**
Find a plane $f(x; w, b) = 0 \Leftrightarrow w^t x + b = 0$ that discriminates perfectly the training points and is "as far as possible" from all the training samples.
For the math that follows, assume the labels $y_i \in \{-1, +1\}$.

A feasible plane should put the positive training samples in one side of the plane and the negative samples in the other side (semi-plane):
$f(x_i; w, b) > 0, if\ y_i = +1$
$f(x_i; w, b) < 0, if\ y_i = -1$

In a more compact way:
$y_i \times f(x_i; w, b) > 0, \forall x_i$

Note that since the algebraic distance of a training point x_i to the planes is $\frac{f(x_i;w,b)}{||w||}$, for feasible planes, $\frac{y_i \times f(x_i;w,b)}{||w||}$ gives the absolute value of the distance.

The SVM model is the feasible plane that gives the "best worst case".
The SVM model is the feasible plane that maximizes the minimum of the distances of the training samples to the plane:

$$\text{argmax}_w \left(\min_i \frac{y_i \times f(x_i; w, b)}{||w||} \right)$$

S.t.:
$y_i \times f(x_i; w, b) > 0, \forall x_i$

Rewriting:

$$\text{argmax}_w \left(\min_i \frac{y_i(w^t x_i + b)}{||w||} \right) \quad \text{Prob(A)}$$

$y_i(w^t x_i + b) > 0, \forall x_i$

Since the plane $w^t x + b = 0$ is the same plane as $w^t x/k + b/k = 0$, for $k \neq 0$, one can work with the subset of equations (of planes) for which
$(\min_i y_i(w^t x_i + b) = 1)$
(see **Details A**)
The optimization problem simplifies to

$$\text{argmax}_w \left(\frac{1}{||w||} \right)$$

$y_i(w^t x_i + b) > 1, \forall x_i$

Finally, it can be rewritten as:
$\text{argmin}_w\ (w^t w)$
$y_i(w^t x_i + b) > 1, \forall x_i$

Nonlinearly separable datasets
Now, there is no plane to perfectly discriminate the points. Accepting that it's still better to stay with the linear model (maybe it's just due to the noise that we are not able to linearly discriminate the points).

However, now, prob(A) has no solution, the constraints cannot be satisfied for all points. This means that for some points $y_i(w^t x_i + b)$ will be negative, which we can compensate by adding a nonnegative quantity ϵ_i.
So, the constraints can be reformulated as
$y_i(w^t x_i + b) + \epsilon_i > 0, \forall x_i$
$\epsilon_i \geq 0, \forall x_i$

Without adapting the goal function these constraints lead to a trivial solution, very far away from the training points and where the points of one of the classes are all misclassified (with this trivial solution, the distance of the training samples to the plane are all very high).
The workaround is to penalize misclassifications in the goal function. One need to accept that some points have to be misclassified but we prefer solutions with just a few misclassifications.

$$\text{argmin}_w \left(C \sum_{n=1}^{N} \epsilon_n + \frac{||w||}{\min_i y_i(w^t x_i + b)} \right)$$

C is a nonnegative constant weighing the cost of the misclassifications with the importance of achieving a large margin (the margin is the minimum of the distances).

Using the same simplification as before, one finally gets

$$\text{argmax}_w \left(C \sum_{n=1}^{N} \epsilon_i + ||w|| \right)$$

$y_i(w^t x_i + b) + \epsilon_i > 1, \forall x_i$
$\epsilon_i \geq 0, \forall x_i$

Although valid, it's much more common in practice to relate the sum of the training errors with the square of the inverse of the margin, leading to the formulation widely adopted in SVMs:

$$\text{argmax}_w \left(C \sum_{n=1}^{N} \epsilon_n + w^t w \right)$$

$y_i(w^t x_i + b) + \epsilon_i > 1, \forall x_i$
$\epsilon_i \geq 0, \forall x_i$

Note that the two constraints can be combined in a single one:
$\epsilon_i \geq \max(0, 1 - y_i(w^t x_i + b)), \forall x_i$
Leading to

$$\text{argmax}_w \left(C \sum_{n=1}^{N} \epsilon_n + w^t w \right)$$

s.t.
$\epsilon_i \geq \max(0, 1 - y_i(w^t x_i + b)), \forall x_i$
Finally, there is always an optimal for which
$\epsilon_i = \max(0, 1 - y_i(w^t x_i + b)), \forall x_i$
(see **detailsEq**)

And therefore the constraints can be integrated in the goal function leading to

$$\text{Model} = \text{argmax}_w \left(C \sum_{i=1}^{N} \max(0, 1 - y_i(w^t x_i + b)) + w^t w \right) \quad \text{(EqSVMmodel)}$$

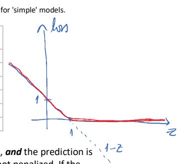
Hinge loss

It is instructive to analyse the previous goal function as a sum of two terms. The first part, the data fitness,

$$\left(\sum_{i=1}^{N} \max(0, 1 - y_i(w^t x_i + b)) \right)$$

is penalizing errors in the training set, while the second is setting preference for 'simple' models.
Let $z = y(w^t x + b)$. If the prediction of the model

y	$w^t x + b$	$z = y(w^t x + b)$	Loss = $\max(0, 1 - y_i(w^t x_i + b)) = \max(0, 1 - z_i)$
+1	-5	-5	6
+1	5	5	0
+1	0.5	0.5	0.5
-1	5	-5	6
-1	-5	5	0
-1	-0.5	0.5	0.5



If the prediction agrees in sign with the sign of the true label, ***and*** the prediction is 'strong' (greater than 1 in absolute value), the prediction is not penalized. If the prediction agrees in sign with the sign of the true label ***but*** the prediction is weak (less than 1 in absolute value), in the sense that is very close to the boundary, the prediction is weakly penalized. If the prediction is with the wrong sign, the prediction is strongly penalized, proportional to absolute value of the prediction.
This loss function - named the hinge loss - by penalizing predictions close to the boundary tries to push the values away from the boundary (and on the right side).

Kernel formulation: towards nonlinear models
So far, we have only been able to work with linear models (planes).
Towards the goal of achieving nonlinear models, we need to rewrite, again, our goal function.
Consider the goal function as given in (EqSVMmodel).
In optimal solution (w^*, b^*), the w^* vector can always be written as a linear combination of the training observations.
As such, on can restrict the optimization process to the set of feasible solutions that can be written as a linear combination of the training samples.

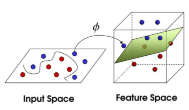
$$\text{Model} = \text{argmax}_{\alpha, b} \left(C \sum_{i=1}^{N} \max(0, 1 - y_i(\textstyle\sum_{j=1}^{N} \alpha_j x_j^t x_i + b)) + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j x_i^t x_j \right)$$

The important part to note now is that input observations are only used as the inner product between pairs of observations: $K(x_i, x_j) = x_i^t x_j$.
This is true both when training (as seen in the previous equation) and for inference.
The prediction for a test observation x_i is

$$w^t x_i + b = \sum_{j=1}^{N} \alpha_j x_j^t x_i + b$$

(Note that b and α_i were learnt using only the inner products between pairs of training samples. When making prediction, one uses those values and the inner products between training samples and the test sample.)

Assume now that a linear model in the input space is not a good idea (see figure).
One may then choose to apply the above described methodology, not directly in the input x-space but in the new (transformed) z-space.
The effort is not much. One just needs to replace the inner product between to observations x_1 and x_2 by the corresponding inner product between z_1 and z_2.



Assume as an example that $z^{(1)} = x^{(1)}$, $z^{(2)} = x^{(2)}$, $z^{(3)} = (x^{(1)})^2 + (x^{(2)})^2$.
In here $x^{(i)}$ is the i-component of vector x.
In this case the inner product between vectors z_1 and z_2 is $z_1^t z_2 = x_1^t x_2 + ||x||^2 ||y||^2$. On can say that the inner product in the new, feature space, is given by a function computed over the input space:
$z_1^t z_2 = K(x_1, x_2) = x_1^t x_2 + ||x||^2 ||y||^2$.

This function is called the kernel function. The advantage is that often we can avoid the explicit computation of the feature z-space. By working with suitable kernel functions we are indeed exploring implicit feature transformations.

For sure, not every function between two vectors to a scalar $K(x_1, x_2)$ will correspond to the inner product of some implicit feature space (see theory related with Mercer's theorem - https://en.wikipedia.org/wiki/Mercer%27s_theorem).
But some well-known kernel functions are often used in practice with good results. RBF, polynomial, linear are among the most typical kernels.
It is a good idea to think of kernels as our measure of similarity between observations. This kernel can be application specific. For instance in computer vision, if an object is represented by the histogram of the grayscale values, one can use as kernel the correlation between histograms.

Support vectors
ZZZZ
ZZZZZZZ

Optimization
ZZZZZZZZZZ

Distance between a point and a plane

Let the function $f(x; w) = w^t x + b, x \in R^d$

The set of points that make the equation $f(x; w) = 0$ true is a plane.

The algebraic distance between a point x and the plane $f(x; w) = 0$ is given by $\frac{f(x;w)}{||w||} = \frac{w^t x + b}{||w||}$.
Note that this is a signed distance: points in one of the semi-planes have positive distance; points in the other semi-plane have negative distance.

-If $x \in plane$, the distance is zero.
-If x is the origin, the proof A or the prof B
-for any $x \in R^d$, see proof C

Details A

Let w^*, b^* be the optimal solution.

Let j be index for which $\left(\min_i \frac{y_i(w^{*t} x_i + b^*)}{||w^*||} \right)$ is minimized:

$$\left(j = \text{argmin}_i \frac{y_i(w^t x_i + b)}{||w||} \right)$$

$k = f(x_j; w^*, b^*)$

Then the plane $w^{*\prime} = \frac{w^*}{k}$, $b^{*\prime} = \frac{b^*}{k}$ is also an optimal plane/solution and $\left(\min_i \frac{y_i(w^{*\prime t} x_i + b^{*\prime})}{||w^{*\prime}||} \right) = 1$

detailsEq

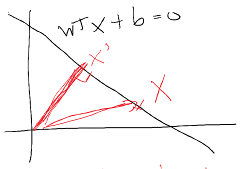
Assume in the optimal solution some ϵ_i is stricly greater than $\max(0, 1 - y_i(w^t x_i + b))$.
$\epsilon_i^* > \max(0, 1 - y_i(w^t x_i + b))$

Rewriting
$\epsilon_i^* = \delta_i + \max(0, 1 - y_i(w^t x_i + b)), \delta_i > 0$.

Consider now the solution
$\epsilon_i^{*\prime} = \epsilon_i^* - \delta_i$

This solution is still feasible (obeys the constraints) and attains a lower value in the goal function.

Assume that the optimal w, w^*, cannot be written as a linear combination of the training samples, including an additional term $w_\perp$: $w^* = \sum_{i=1}^{N} \alpha_i^* x_i + w_\perp = w_{//} + w_\perp$.
The value attained by the goal function will be
$\left(C \sum_{i=1}^{N} \max(0, 1 - y_i(w^{*t} x_i + b)) + w^{*t} w^* \right) =$
$\left(C \sum_{i=1}^{N} \max(0, 1 - y_i(w_{//}^t x_i + b)) + w_{//}^t w_{//} + w_\perp^t w_\perp \right)$
This value can always be decreased (improved the goal function) by setting $w_\perp$ to zero.



the gradients should be proportional

Ⓐ $\begin{cases} \min ||\cancel{x}|| \\ \text{st. } w^T x + b = 0 \end{cases} \Leftrightarrow \begin{cases} \min x^T x \\ \text{st } w^T x + b = 0 \end{cases} \Leftrightarrow \begin{cases} 2x = \lambda w \\ w^T x + b = 0 \end{cases} \begin{cases} 2w^T x = \lambda w^T w \\ w^T \cancel{x} + b = 0 \end{cases} \begin{cases} w^T x = -b \\ \lambda = -\frac{2b}{w^T w} \end{cases} \begin{cases} x = -\frac{b}{||w||^2} \cdot w \\ ||x|| = \frac{|b|}{||w||} \end{cases}$

$w^T x = -b$
$\lambda = \frac{-2b}{w^T w}$

Ⓑ distance = projection of any $x \in$ plane over the w direction (w is orthogonal to the plane)

$\frac{|\langle x, w \rangle|}{||w||} = \frac{|w^T \cancel{x}|}{||w||} = \frac{|b|}{||w||} = \frac{|f(0; w)|}{||w||}$

$\Updownarrow$

x_A and $x_B \in$ plane
$w^T(x_A - x_B) = 0$

Ⓒ modify A accordingly

$x^T x =$

$= \left(x + \frac{b}{||w||} \hat{w} - \frac{b\hat{w}}{||w||} \right)^T \left(\cancel{x} + \frac{b\hat{w}}{||w||} - \frac{b\hat{w}}{||w||} \right)$

$= \left(x + \frac{b\hat{w}}{||w||} \right)^T \left(x + \frac{b\hat{w}}{||w||} \right) - \left(x + \frac{b\hat{w}}{||w||} \right)^T \cdot \frac{b\hat{w}}{||w||}$ $= 0$ if $w^T x + b = 0$

$- \frac{b\hat{w}^T}{||w||} \cdot \left(x + \frac{b\hat{w}}{||w||} \right)$ $= 0$ if $w^T x + b = 0$

$+ \frac{b\hat{w}^T b\hat{w}}{||w||^2}$

$= \left(x + \frac{b\hat{w}}{||w||} \right)^T \left(x + \frac{b\hat{w}}{||w||} \right) + \frac{b^2 \hat{w}^T \hat{w}}{||w||^2}$

minimized for $x = -\frac{b\hat{w}}{||w||}$