Support Vectors and the Margin in a Nutshell

Andreas Maunz
Freiburg Center for Data Analysis and Modelling (FDM), Hermann-Herder-Str. 3a, 79104 Freiburg, Germany

Leaving mathematical sophistication aside, this document briefly outlines the theory of support vectors and the concept of margin. It is a very condensed version of parts of chapters 1, 2, 7 and 9 of "Learning with Kernels" by Schölkopf and Smola, 2002. Kernel functions are not explained and merely assumed to be useful for non-linear solutions in input space. Support Vector Machines make heavy use of Lagrangians to solve constrained optimization problems. However, this technique is also not explained in detail here.

If not explicitly stated otherwise, i, j always run over $1, \ldots, m$.

1 Optimal margin hyperplanes

Consider the class of hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ corresponding to binary decision functions

$$sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b). \tag{1}$$

Based on empirical training data $(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^n, y \in \{-1, 1\}$, one can find a unique optimal hyperplane which is the solution of the following optimization problem:

$$\underset{\mathbf{w},b}{max}\ \underset{i}{min}\ \{||\mathbf{x} - \mathbf{x}_i|| \ |\langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}. \tag{2}$$

In words: find the hyperplane that has maximum distance to the nearest training vectors (*support vectors*). This can be achieved by minimizing $\mathbf{w}$ using the *objective function*

$$\underset{\mathbf{w}}{min}\ \frac{1}{2}||\mathbf{w}||^2, \tag{3}$$

subject to

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1. \tag{4}$$

The left-hand side of (4) divided by $||\mathbf{w}||$ gives the distance between $\mathbf{x}_i$ and the hyperplane and minimizing $||\mathbf{w}||$ thus maximizes this distance, called *margin*[a]. The following Lagrangian can be used to solve this optimization problem:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_i \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \tag{5}$$

We want the solution that maximizes (5) over the α_i and minimizes (5) over $\mathbf{w}, b$. The α_i (*primal variables*) are weights for the $\mathbf{x}_i$. If $\mathbf{x}_i$ violates (4) then (5) can be increased by increasing α_i. Therefore, $\mathbf{w}$ and b will have to change to satisfy more constraints and decrease $||\mathbf{w}||$. Note that the α_i for which the $\mathbf{x}_i$ fulfill (4) but are not precisely met as equalities have to be 0 to maximize (5), i.e. only the support vectors have non-zero α-weights.

Lagrangians are solved by finding zero points of their partial derivatives (*saddle point condition*). It is instructive to calculate $\frac{d}{db}L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$ and $\frac{d}{d\mathbf{w}}L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$ manually[b]. This leads to

$$\sum_i \alpha_i y_i = 0, \tag{6}$$

and

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i, \tag{7}$$

[a] Note that all support vectors then have distance of $\frac{1}{||\mathbf{w}||}$ to the hyperplane.

[b] The reason why (3) not directly minimizes $||\mathbf{w}||$ will become clear here.

repectively. The derivative (7) implies that the solution vector has an expansion in terms of some training vectors, namely those with non-zero α-weight: the *support vectors*. Substituting (6) and (7) into (5) yields

$$\max_{\alpha} W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle, \tag{8}$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$, cf. (6), the *dual optimization problem*. In practice, however, the kernel trick is used, modifying (8) to[c]

$$\max_{\alpha} W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j). \tag{9}$$

Using a kernel function lifts the algorithm to a higher dimensional *feature space*, thus enabling a non-linear solution in input space. The decision function (1) can be rewritten due to (7) and the kernel trick into

$$f(x) = sgn(\sum_i \alpha_i y_i k(x_i, x) + b). \tag{10}$$

To estimate b, one can make use of the fact that only support vectors have non-zero α-weight (*KKT conditions*):

$$y_j = \sum_i \alpha_i y_i k(x_i, x) + b. \tag{11}$$

Thus, b can be obtained by e.g. averaging over all y_j, which completes the decision function.

2 Soft margin hyperplanes

2.1 The important role of the margin

If a seperating hyperplane does not exist, the constraint (4) has to be relaxed with *slack variables* ξ_i:

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \tag{12}$$

subject to $\xi_i \geq 0$. The idea is to allow points to lie within the margin or being misclassified to improve robustness towards outliers.

Soft margin hyperplanes are a generalization of optimal margin hyperplanes. The support vectors for the latter lie exactly on the margin (the rest contributes nothing to the solution), while for the former, support vectors are also allowed to lie *inside* the margin. The latter support vectors are called *margin errors*.

To allow for margin errors, we will see that the α-values have to be constrained, i.e. $0 \leq \alpha_i \leq b$ for some upper bound b. If $\alpha_i = b$ (or $\alpha_i = 0$), we call α_i *at bound*.

2.2 C-SVC

To allow for margin errors modify the objective function (3) to

$$\min_{\mathbf{w}, \xi_i} \frac{1}{2} ||\mathbf{w}||^2 + \frac{C}{m} \sum_i \xi_i, \tag{13}$$

subject to (12), where $C > 0$ determines the tradeoff between margin maximization and error minimization. Transforming this into a dual yields again (9), but subject to the new constraints $0 \leq \alpha_i \leq \frac{C}{m}$ and $\sum_i \alpha_i y_i = 0$. The solution can also be shown to have expansion (7) and decision function (10). The threshold b can be evaluated as in (11), too, but only for support vectors x_i (defined by meeting the equality of (12)) which additionally have $\xi_i = 0$, i.e. that sit directly on the edge of the margin.

[c] Note that the x_i need not be vectors from an inner-product space anymore, extending the approach to all inputs that a positive definite kernel function k is defined for.

2.3 ν-SVC

The parameter C is rather unintuitive and hard to choose a priori. A modification is the following objective function, governed by a parameter ν:

$$\underset{\mathbf{w},\xi_i,\rho,b}{min} \quad \frac{1}{2}||\mathbf{w}||^2 - \nu\rho + \frac{1}{m}\sum_i \xi_i, \tag{14}$$

subject to the constraints $y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq \rho - \xi_i$ and $\xi_i \geq 0$, as well as $\rho \geq 0$. To understand the role of ρ, observe that for $\boldsymbol{\xi} = 0$ the first constraint states that the two classes be separated by a margin of width $\frac{2\rho}{||w||}$.

The problem can be formulated as the corresponding Lagrangian

$$\frac{1}{2}||w||^2 - \nu\rho + \frac{1}{m}\sum_i \xi_i - \sum_i(\alpha_i(y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - \rho + \xi_i) + \beta_i\xi_i) - \delta\rho. \tag{15}$$

Setting the partial derivatives for the four primal variables $\mathbf{w}, \xi, b, \rho$ to 0 yields the four Lagrangian constraints $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$, $\alpha_i + \beta_i = \frac{1}{m}$, $\sum_i \alpha_i y_i = 0$ and $\sum_i \alpha_i - \delta = \nu$. Injection of those constraints into (15) yields

$$\underset{\alpha}{max}\ W(\alpha) = -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i, x_j), \tag{16}$$

subject to $0 \leq \alpha_i \leq \frac{1}{m}$, $\sum_i \alpha_i y_i = 0$ and $\sum_i \alpha_i \geq \nu$. The solution can also be shown to have decision function (10).

If a ν-SVC run yields a positive margin ($\rho > 0$, and therefore $\sum_i \alpha_i = \nu$ according to the fourth Lagrangian constraint), then ν is

- an *upper bound on the fraction of margin errors*: each α_i can be at most $\frac{1}{m}$ and only a fraction ν of the examples can have this α-value, which all margin errors do.
- also a *lower bound on the fraction of support vectors*: since every sv can have an alpha-value of at most $\frac{1}{m}$, there must be at least νm of them (including margin errors which are also support vectors).

However, for large datasets, the fraction of support vectors sitting directly on the margin can be neglected and the two numbers converge.

3 SV-Regression

To model quantitative targets, proceed analogous to the qualitative case. At each point, allow an error ϵ. Everything above ϵ is covered in slack variables $\xi_i^{(*)}$, which are penalized in the objective function [d]. Specifically, for $\mathbf{y} \in \mathbb{R}$, use the *$\epsilon$-insensitive loss function* to preserve sparse representation of the solution:

$$|y - f(\mathbf{x})|_\epsilon = max\{0, |y - f(\mathbf{x})| - \epsilon\} \tag{17}$$

This takes the form of the normal $|\cdot|$ function with y-intercept of $-\epsilon$, but with a *zero-error-tube* of width ϵ around the center. Therefore, errors inside the tube are not penalized. Note, that this "swaps" the penalization area: now the margin errors lie in the area *outside* the tube. Consequently, points *inside* the tube do not appear in the extension of the solution.

[d] The notation $x^{(*)}$ references the variables x and x^* at the same time.

3.1 ϵ-SVR

For a fixed ϵ, the corresponding constrained optimization problem is analogous to ϵ-SVC given by

$$\underset{\mathbf{w},\xi^{(*)},b}{min} \quad \frac{1}{2}||\mathbf{w}||^2 + C\frac{1}{m}\sum_i(\xi_i + \xi_i^*), \tag{18}$$

subject to

$$\begin{aligned}(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i &\leq \epsilon + \xi_i,\\ (y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle + b) &\leq \epsilon + \xi_i^*,\\ \xi_i^{(*)} &\geq 0.\end{aligned} \tag{19}$$

Note, that for $|\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i| \leq \epsilon$, we have $\xi_i^{(*)} = 0$. Transforming (18) and constraints (19) into a Lagrangian yields

$$\begin{aligned}&\frac{1}{2}||\mathbf{w}||^2 + \frac{C}{m}\sum_i(\xi_i + \xi_i^*) - \sum_i(\eta_i\xi_i + \eta_i^*\xi_i^*)\\ &- \sum_i \alpha_i(\epsilon + \xi_i + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b)\\ &- \sum_i \alpha_i^*(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle + b),\end{aligned} \tag{20}$$

which must be minimized with respect to the primal variables $\mathbf{w}, b, \xi_i^{(*)}$ and maximized with respect to the dual variables $\alpha_i^{(*)}$. Hence, the saddle point condition yields the three constraints $\sum_i(\alpha_i - \alpha_i^*) = 0$, $\mathbf{w} - \sum_i(\alpha_i^* - \alpha_i)\mathbf{x}_i = 0$ and $\frac{C}{m} - \alpha_i^{(*)} - \eta_i^{(*)} = 0$ which can be inserted into (20), yielding the following dual problem:

$$\begin{aligned}\underset{\boldsymbol{\alpha}^{(*)}}{max} \quad &-\frac{1}{2}\sum_{i,j}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle \mathbf{x}_i, \mathbf{x_j}\rangle\\ &-\epsilon\sum_i(\alpha_i^* + \alpha_i) + \sum_i y_i(\alpha_i^* - \alpha_i),\end{aligned} \tag{21}$$

subject to $\sum_i(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i^{(*)} \in [0, \frac{C}{m}]$. The solution has again an expansion in terms of support vectors:

$$f(\mathbf{x}) = \sum_i(\alpha_i^* - \alpha_i)\langle \mathbf{x}, \mathbf{x}_i\rangle + b. \tag{22}$$

At the point of solution, due to the KKT conditions, the product between dual variables and constraints has to vanish:

$$\begin{aligned}&\alpha_i(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle + b) = 0 \qquad \text{and}\\ &\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b) = 0\ ,\end{aligned} \tag{23}$$

as well as

$$\begin{aligned}&(\frac{C}{m} - \alpha_i)\ \xi_i = 0 \qquad \text{and}\\ &(\frac{C}{m} - \alpha_i^*)\ \xi_i^{(*)} = 0.\end{aligned} \tag{24}$$

The parameter ϵ has taken the role of the margin parameter ρ here. Support vectors lie either directly on the edge of the tube or outside the tube.

Condition (23) allows to compute b by exploiting the former points, i.e. the cases where $0 < \alpha_i^{(*)} < \frac{C}{m}$ due to the second factor of (23) being 0, but additionally $\xi_i^{(*)} = 0$ holds for these cases, too. Furthermore, conclude that only points with $\alpha_i^{(*)} = \frac{C}{m}$ can have $\xi_i^{(*)} > 0$, i.e. can lie outside the tube and there is no i for which $\alpha_i > 0$ *and* $\alpha_i^* > 0$ (c.f. proof in section 4).

3.2 ν-SVR

Instead of using a fixed ϵ, optimize ϵ. Use as objective function:

$$\min_{\mathbf{w},\xi^{(*)},\epsilon,b} \quad \frac{1}{2}||\mathbf{w}||^2 + C\,(\nu\epsilon + \frac{1}{m}\sum_i(\xi_i + \xi_i^*)), \tag{25}$$

subject to

$$\begin{aligned}(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i &\leq \epsilon + \xi_i,\\ (y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle + b) &\leq \epsilon + \xi_i^*,\\ \xi_i^{(*)} &\geq 0,\\ \epsilon &\geq 0.\end{aligned} \tag{26}$$

Note the similarity to ϵ-SVR. Transforming (25) and constraints (26) into a Lagrangian yields

$$\begin{aligned}&\frac{1}{2}||\mathbf{w}||^2 + C\nu\epsilon + \frac{C}{m}\sum_i(\xi_i + \xi_i^*) - \sum_i(\eta_i\xi_i + \eta_i^*\xi_i^*)\\ &- \sum_i \alpha_i(\epsilon + \xi_i + y_i - \langle \mathbf{w}, \mathbf{x}_i\rangle - b)\\ &- \sum_i \alpha_i^*(\epsilon + \xi_i - y_i + \langle \mathbf{w}, \mathbf{x}_i\rangle + b)\end{aligned} \tag{27}$$

The saddle point condition yields this time the shorter dual:

$$\begin{aligned}\max_{\boldsymbol{\alpha}^{(*)}} \quad &-\frac{1}{2}\sum_{i,j}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle \mathbf{x_i}, \mathbf{x_j}\rangle\\ &+ \sum_i y_i(\alpha_i^* - \alpha_i),\end{aligned} \tag{28}$$

subject to $\sum_i(\alpha_i - \alpha_i^*) = 0$, $\alpha_i^{(*)} \in [0, \frac{C}{m}]$ and $\sum_i(\alpha_i + \alpha_i^*) \leq Cv$.

Again, b and this time also ϵ can be computed using the KKT conditions (23), i.e. computing thickness and vertical position of the tube by using points that sit exactly on the border of the tube.

Again, ν is

- an *upper bound on the fraction of margin errors*
- a *lower bound on the fraction of support vectors*

The formal proof is analogous to the corresponding proof of ν-SVC, however, the following "sloppy" argumentation is more instructive:
The first statement can be seen by observing that, for increasing ϵ, the first term in $\nu\epsilon + \frac{1}{m}\sum_i(\xi_i + \xi_i^*)$ increases proportionally to ν (i.e. with constant gradient $g_1 = \nu$), while the second term decreases (monotonically and) proportionally to the fraction of margin errors h_1 (the points on the edge of the tube cost nothing), inducing a varying gradient $g_2 = -h_1$. Since the terms are added up, a minimum of the sum can not be found until $g_1 + g_2 = 0$ (the combined gradient is the sum of the single gradients), i.e. $h_1 = \nu$.
Analogously, for the second statement, observe that, for decreasing ϵ, the first term decreases with gradient $-\nu$ while the second term increases (monotonically and) with a gradient that equals the fraction of support vectors (the points on the edge gain nothing) and which must reach ν for the minimum.

4 Proofs

4.1 Problem 9.1

For $\epsilon > 0$ the solution of the ϵ-SVR dual satisfies $\alpha_i \alpha_i^* = 0$: Assume the contrary, i.e. $\alpha_i > 0$ and $\alpha_i^* > 0$ for $\epsilon > 0$. It follows from (23) that

$$\begin{aligned} (i) \qquad & \epsilon + \xi_i = y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \qquad \text{and} \\ (ii) \qquad & \epsilon + \xi_i^* = \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i. \end{aligned}$$

Inserting (ii) in (i) by substituting y_i yields

$$\begin{aligned} \epsilon + \xi_i &= -\epsilon - \xi_i^* \qquad \Leftrightarrow \\ \epsilon &= -\frac{1}{2}(\xi_i + \xi_i^*). \end{aligned}$$

Since $\xi_i^{(*)} \geq 0$, it follows that $\epsilon \leq 0$, which is a contradiction. □