Support Vector Machine for Classification (SVM)

In Support Vector Machine (SVM) for classification, given a training data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)\}$ of M samples, where $\mathbf{x}^i \in \mathbb{R}^N$ are multi-dimensional inputs and $y_i \in \{-1, 1\}$ are categorical (uni-dimensional) outputs, we seek to find a continuous mapping function $f : \mathbb{R}^N \to \mathbb{R}^F$ such that

$$\begin{aligned} h :\ & \mathbb{R}^N \to \{-1, 1\} \\ & \mathbf{x} \mapsto \operatorname{sgn}(f(\mathbf{x})) \end{aligned}$$

best predicts the labels (or class) of the training points with the function $y = h(\mathbf{x})$. Equivalently, this amounts to finding a continuous mapping function $f(\mathbf{x})$ such that the hyperplane $f(\mathbf{x}) = 0$ best separate the training datapoints.

In order to select among the infinite number of possible separating hyperplanes, the goal of SVM is to obtain a score function $f(\mathbf{x})$ such that[1]

$$\begin{cases} f(\mathbf{x}^i) \leq -1, & \text{if} \quad y_i = -1 \\ f(\mathbf{x}^i) \geq 1, & \text{if} \quad y_i = 1 \end{cases} \qquad \Longleftrightarrow \quad y_i f(x_i) \geq 1$$

and the margin, corresponding the width of the region between the hyperplanes defined by $f(\mathbf{x}) = 1$ and $f(\mathbf{x}) = -1$ (parallel to the separating hyperplane), is maximized.

Following the problem formulation of SVM described above, we seek to find a function $f(x)$ of the form:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^N$$

such that points that lie outside of the margin and on the correct side of the separating hyperplane defined by $f(\mathbf{x}) = 0$, i.e. for which $yf(x) \geq 1$, are not penalized. This function can be learned by solving the following optimization problem, also called primal problem:

$$\begin{aligned} & \min_{\mathbf{w}, b, \xi, \xi^*} \left(\frac{1}{2} ||\mathbf{w}||^2 + \frac{C}{M} \sum_{i=1}^{M} \xi_i \right) \\ & \text{s.t.} \quad y^i \left(\langle \mathbf{w}, \mathbf{x}^i \rangle + b \right) \geq 1 - \xi_i \\ & \qquad \xi_i \geq 0, \forall i = 1 \ldots M \end{aligned} \tag{1}$$

where $\mathbf{w}$ is weight vector, ξ_i are the slack variables, b is the bias and C is the penalty factor associated to points that lie in the margin or on the wrong side of the separating hyperplane.

In order to solve this convex optimization problem, we use the method of Lagrange multipliers. By defining the positive multipliers $\alpha = \{\alpha_1, ..., \alpha_M\}, \beta = \{\beta_1, ..., \beta_M\}$, the Lagrangian of the above problem can be written as:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{M} \sum_{i=1}^{M} \xi_i - \sum_{i=1}^{M} \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}^i \rangle + b) - 1 + \xi_i) - \sum_{i=1}^{M} \beta_i \xi_i$$

[1] In practice, the algorithm is given some slack and points lying inside the margin (or even on the wrong side of the separating hyperplane) are tolerated but penalized in the optimization.

To minimize with respect to the primal variables, the partial derivatives of the Lagrangian with respect to $\mathbf{w}$, b and ξ must vanish, i.e.:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \textstyle\sum_{i=1}^{M} y_i \alpha_i \mathbf{x}^i = 0 \iff \boxed{\mathbf{w} = \displaystyle\sum_{i=1}^{M} y_i \alpha_i \mathbf{x}^i} \\ \dfrac{\partial \mathcal{L}}{\partial b} = - \textstyle\sum_{i=1}^{M} y_i \alpha_i = 0 \iff \boxed{\displaystyle\sum_{i=1}^{M} y_i \alpha_i = 0} \\ \dfrac{\partial \mathcal{L}}{\partial \xi_i} = \dfrac{C}{M} - \alpha_i - \beta_i = 0 \iff \boxed{\beta_i = \dfrac{C}{M} - \alpha_i} \end{cases}$$

where the last equation and the positiveness of β_i imply that $\alpha_i \in \left[0; \dfrac{C}{M}\right]$.

The dual problem is obtained by substituting the above equations into the following optimization program, equivalent to the primal:

$$\begin{aligned} &\max_{\alpha,,\beta} \min_{\mathbf{w},b,\xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) \\ &\text{s.t.} \quad \alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0 \\ &\qquad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \end{aligned}$$

After a few simplifications, the dual problem takes the following form:

$$\begin{aligned} &\max_{\alpha} \sum_{i=M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}^i, \mathbf{x}^j \rangle \\ &\text{s.t.} \qquad 0 \leq \alpha_i \leq \frac{C}{M} \quad \text{and} \quad \sum_{i=1}^{M} y_i \alpha_i = 0 \end{aligned}$$

Additionally, the complementary optimality conditions, also called the KKT (Karush-Kuhn-Tucker), can be written for all (training) points as:

$$\begin{cases} \alpha_i (y_i (\langle w, \mathbf{x}^i \rangle + b) - 1 + \xi_i) = 0 \\ \xi_i \beta_i = 0 \end{cases} \implies \boxed{\begin{cases} \alpha_i (y_i (\langle w, \mathbf{x}^i \rangle + b) - 1 + \xi_i) = 0 \\ \xi_i \left(\alpha_i - \dfrac{C}{M} \right) = 0 \end{cases}}$$

These latter conditions allow us to define the support vectors for which either $\alpha_i > 0$. The support vectors thus lie either on the boundaries of the margin, inside the margin or on the wrong side of the separating hyperplane. In addition, the last two conditions imply that all the points inside the margin or on the wrong side of the separating hyperplane, for which $\xi_i > 0$, must satisfy $\alpha_i = \dfrac{C}{M}$. Finally, all the other points satisfy $\alpha_i = 0$.

By solving the dual of this objective function, we obtain a score function of the following form:

$$\boxed{y = f(\mathbf{x}) = \sum_{i=1}^{M} y_i \alpha_i \langle \mathbf{x}, \mathbf{x}^i \rangle + b} \tag{2}$$

Nonlinear classification

In order to perform nonlinear classification, we proceed to a transformation of the original dataset through the mapping function

$$\begin{aligned}\phi :\ & \mathbb{R}^N \to \mathbb{R}^F \\ & \mathbf{x} \mapsto \phi(\mathbf{x})\end{aligned}$$

where F may be infinity. Instead of working directly in this potentially infinite-dimensional space, we replace the dot products in feature space, $\langle \phi(x^i), \phi(x^j) \rangle$, by the corresponding kernel function, $k(x^i, x^j)$ in our optimization problem, this is the so-called *kernel trick*, the regressive function then becomes:

$$\boxed{y = f(\mathbf{x}) = \sum_{i=1}^{M} y_i \alpha_i k(\mathbf{x}, \mathbf{x}^i) + b} \tag{3}$$

where we used:

$$\mathbf{w} = \sum_{i=1}^{M} y_i \alpha_i \phi(\mathbf{x}) \implies \langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{M} y_i \alpha_i \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}) \rangle = \sum_{i=1}^{M} y_i \alpha_i k(\mathbf{x}, \mathbf{x}^{\mathbf{i}})$$

Note: As shown in the above equations, in the nonlinear case, $\mathbf{w}$ lives in $\mathbb{R}^F$.

Kernels

The kernel function can be any type of function that corresponds to a dot-product of the features transformed to a high-dimensional space. In this course, we focus on only two types of kernels:

- *Linear*: $k(\mathbf{x}, \mathbf{x}^i) = \langle \mathbf{x}, \mathbf{x}^i \rangle$
  which is equivalent to the linear formulation of SVM.
- *Radial Basis Function (RBF)*: $k(\mathbf{x}, \mathbf{x}^i) = \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{x} - \mathbf{x}^i||^2\right\}$
  where σ is an hyper-parameter and corresponds to the width or scale of the Gaussian kernel centered at $\mathbf{x}^i$

Optimization parameters

Apart from the kernel hyper-parameters, C-SVM has one open-parameter:

- C: Cost $[0 \to \infty]$ represents the penalty associated to points inside the margin or on the wrong side of the separating hyperplane. Increasing cost value causes closer fitting to the training data.

The total number of hyper-parameters is thus 1 for linear SVM and 2 for nonlinear SVM with RBF kernel.

Determining b

In order to determine the value of b from the optimization program, one can use the KKT conditions. Indeed, for the support vectors that lie on the boundaries of the margin, we have:

$$\xi_i = \xi_i^* = 0$$

and

$$\begin{cases} \alpha_i \in \left]0; \dfrac{C}{M}\right[\\ y_i(\langle w, \mathbf{x}^i \rangle + b) - 1 = 0 \end{cases} \iff \begin{cases} \alpha_i \in \left]0; \dfrac{C}{M}\right[\\ \langle w, \mathbf{x}^i \rangle + b - y_i = 0 \end{cases}$$

Hence, we can compute b by averaging the error over the subset of support vectors that lie on the boundary of the margin:

$$\boxed{b = \frac{1}{|I|} \sum_{i \in I} y_i - \langle w, \mathbf{x}^i \rangle}$$

or, equivalently

$$\boxed{b = \frac{1}{|I|} \sum_{i \in I} y_i - \sum_{j=1}^{M} y_j \alpha_j k(\mathbf{x}^j, \mathbf{x}^i)}$$

where I is the set of indices of the support vectors that lie on the boundary of the margin, i.e. for which $\alpha_i \in \left]0; \frac{C}{M}\right[$,

$$|I| = \sum_{i=1}^{M} \mathbb{1}_{]0;\frac{C}{M}[}(\alpha_i) \quad \text{where} \quad \mathbb{1}_{]0;\frac{C}{M}[}(\alpha) = \begin{cases} 1, & \text{if} \quad \alpha \in \left]0; \frac{C}{M}\right[\\ 0, & \text{otherwise} \end{cases}$$

Support Vector Machine for Regression (SVR)

In Support Vector Regression (SVR), given a training data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)\}$ of M samples, where $\mathbf{x}^i \in \mathbb{R}^N$ are multi-dimensional inputs and $y_i \in \mathbb{R}$ are continuous uni-dimensional outputs, we seek to find a continuous mapping function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ that best predicts the set of training points with the function $y = f(\mathbf{x})$.

The goal of SVR is to obtain a function $f(\mathbf{x})$ that has at most an ϵ-deviation from the training outputs $\{y_1, \ldots, y_M\}$ and is as *flat* as possible. Intuitively, this means that we do not mind having some errors in our regression, as long as they remain within an acceptable range defined by the ϵ-deviation of $f(\mathbf{x})$[2]. This type of function is often called ϵ-insensitive loss function and the allowed deviation is called ϵ-insensitive tube.

Following the problem formulation of SVR described above, we seek to find an estimate $f(\mathbf{x})$ of the true function:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad \text{with} \quad \mathbf{w} \in \mathbb{R}^N \tag{4}$$

such that points that are contained withing the ϵ-tube, i.e. for which $|f(\mathbf{x}) - y| \leq \epsilon$, are not penalized. This function can be learned by solving the following optimization problem, also called primal problem:

$$\begin{aligned} &\min_{\mathbf{w}, b, \xi, \xi^*} \left(\frac{1}{2} ||\mathbf{w}||^2 + \frac{C}{M} \sum_{i=1}^{M} (\xi_i + \xi_i^*) \right) \\ &\text{s.t.} \quad y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle - b \leq \epsilon + \xi_i \\ &\qquad \langle \mathbf{w}, \mathbf{x}^i \rangle + b - y^i \leq \epsilon + \xi_i^* \\ &\qquad \xi_i, \xi_i^* \geq 0, \forall i = 1 \ldots M \end{aligned} \tag{5}$$

where $\mathbf{w}$ is weight vector, ξ_i, ξ_i^* are the slack variables, b is the bias, ϵ the allowable error and C is the penalty factor associated to errors larger than ϵ.

In order to solve this convex optimization problem, we use the method of Lagrange multipliers. By defining the positive multipliers $\alpha = \{\alpha_1, ..., \alpha_M\}, \alpha^* = \{\alpha_1^*, ..., \alpha_M^*\}, \beta = \{\beta_1, ..., \beta_M\}, \beta^* = \{\beta_1^*, ..., \beta_M^*\}$, the Lagrangian of the above problem can be written as:

$$\begin{aligned} \mathcal{L}(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{M} \sum_{i=1}^{M} (\xi_i + \xi_i^*) + \sum_{i=1}^{M} \alpha_i^* (\langle \mathbf{w}, \mathbf{x}^i \rangle + b - y_i - \epsilon - \xi_i^*) \\ + \sum_{i=1}^{M} \alpha_i (y_i - \langle \mathbf{w}, \mathbf{x}^i \rangle - b - \epsilon - \xi_i) - \sum_{i=1}^{M} \beta_i^* \xi_i^* - \sum_{i=1}^{M} \beta_i \xi_i \end{aligned}$$

[2] In practice, the algorithm is given some slack and larger errors are tolerated but penalized in the optimization.

To minimize with respect to the primal variables, the partial derivatives of the Lagrangian with respect to $\mathbf{w}$, b, ξ, ξ^* must vanish, i.e.:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^{M} \alpha_i^* \mathbf{x}^i - \sum_{i=1}^{M} \alpha_i \mathbf{x}^i = 0 \iff \boxed{\mathbf{w} = \displaystyle\sum_{i=1}^{M} (\alpha_i - \alpha_i^*)\, \mathbf{x}^i} \\ \dfrac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{M} \alpha_i^* - \sum_{i=1}^{M} \alpha_i = 0 \iff \boxed{\displaystyle\sum_{i=1}^{M} (\alpha_i - \alpha_i^*) = 0} \\ \dfrac{\partial \mathcal{L}}{\partial \xi_i} = \dfrac{C}{M} - \alpha_i - \beta_i = 0 \iff \boxed{\beta_i = \dfrac{C}{M} - \alpha_i} \\ \dfrac{\partial \mathcal{L}}{\partial \xi_i^*} = \dfrac{C}{M} - \alpha_i^* - \beta_i^* = 0 \iff \boxed{\beta_i^* = \dfrac{C}{M} - \alpha_i^*} \end{cases}$$

where the last two equations and the positiveness of β_i and β_i^* imply that $\alpha_i, \alpha_i^* \in \left[0; \dfrac{C}{M}\right]$

The dual problem is obtained by substituting the above equations into the following optimization program, equivalent to the primal:

$$\begin{aligned} &\max_{\alpha,\alpha^*,\beta,\beta^*} \min_{\mathbf{w},b,\xi,\xi^*} \mathcal{L}(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) \\ &\text{s.t.} \quad \alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0 \\ &\qquad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \xi_i} = \frac{\partial \mathcal{L}}{\partial \xi_i^*} = 0 \end{aligned}$$

After a few simplifications, the dual problem takes the following form:

$$\begin{aligned} &\max_{\alpha,\alpha^*} -\epsilon \sum_{i=M} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} (\alpha_i - \alpha_i^*)\, y_i - \frac{1}{2} \sum_{i,j=1}^{M} (\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right) \langle \mathbf{x}^i, \mathbf{x}^j \rangle \\ &\text{s.t.} \qquad 0 \leq \alpha_i, \alpha_i^* \leq \frac{C}{M} \quad \text{and} \quad \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) = 0 \end{aligned}$$

Additionally, the complementary optimality conditions, also called the KKT (Karush-Kuhn-Tucker), can be written for all (training) points as:

$$\begin{cases} \alpha_i^*(\langle w, \mathbf{x}^i \rangle + b - y_i - \epsilon - \xi_i^*) = 0 \\ \alpha_i(y_i - \langle w, \mathbf{x}^i \rangle - b - \epsilon - \xi_i) = 0 \\ \xi_i^* \beta_i^* = 0 \\ \xi_i \beta_i = \ 0 \end{cases} \Longrightarrow \boxed{\begin{cases} \alpha_i^*(y_i - \langle w, \mathbf{x}^i \rangle - b + \epsilon + \xi_i^*) = 0 \\ \alpha_i(y_i - \langle w, \mathbf{x}^i \rangle - b - \epsilon - \xi_i) = 0 \\ \xi_i^* \left(\alpha_i^* - \dfrac{C}{M}\right) = 0 \\ \xi_i \left(\alpha_i - \dfrac{C}{M}\right) = 0 \end{cases}}$$

These latter conditions allow us to define the support vectors for which either $\alpha_i > 0$ or $\alpha_i^* > 0$[3]. The support vectors thus lie either outside the ϵ-tube or on its boundaries. In addition, the last two conditions imply that all the points outside the ϵ-tube, for which either $\xi_i > 0$ or $\xi_i^* > 0$, must satisfy $\alpha_i = \dfrac{C}{M}$ or $\alpha_i^* = \dfrac{C}{M}$. Finally, all the other points satisfy $\alpha_i = \alpha_i^* = 0$.

By solving the dual of this objective function, we obtain a regressive function of the following form:

$$\boxed{y = f(\mathbf{x}) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) \langle \mathbf{x}, \mathbf{x}^i \rangle + b} \tag{6}$$

Nonlinear regression

In order to perform nonlinear regression, we proceed to a transformation of the original dataset through the mapping function

$$\begin{aligned} \phi : \ & \mathbb{R}^N \to \mathbb{R}^F \\ & \mathbf{x} \mapsto \phi(\mathbf{x}) \end{aligned}$$

where F may be infinity. Instead of working directly in this potentially infinite-dimensional space, we replace the dot products in feature space, $\langle \phi(x^i), \phi(x^j) \rangle$, by the corresponding kernel function, $k(x^i, x^j)$ in our optimization problem, this is the so-called *kernel trick*, the regressive function then becomes:

$$\boxed{y = f(\mathbf{x}) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(\mathbf{x}, \mathbf{x}^i) + b} \tag{7}$$

where we used:

$$\mathbf{w} = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*)\, \phi(\mathbf{x}) \implies \langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*)\, \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}) \rangle = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*)\, k(\mathbf{x}, \mathbf{x}^{\mathbf{i}})$$

Note: As shown in the above equations, in the nonlinear case, $\mathbf{w}$ lives in $\mathbb{R}^F$.

Kernels

The kernel function can be any type of function that corresponds to a dot-product of the features transformed to a high-dimensional space. In this course, we focus on only two types of kernels:

- *Linear*: $k(\mathbf{x}, \mathbf{x}^i) = \langle \mathbf{x}, \mathbf{x}^i \rangle$
  which is equivalent to the linear formulation of SVR.
- *Radial Basis Function (RBF)*: $k(\mathbf{x}, \mathbf{x}^i) = \exp\left\{-\frac{1}{2\sigma^2} ||\mathbf{x} - \mathbf{x}^i||^2\right\}$
  where σ is an hyper-parameter and corresponds to the width or scale of the Gaussian kernel centered at $\mathbf{x}^i$

[3] By definition $\alpha_i \alpha_i^* = 0$. Hence, if $\alpha_i > 0$, then $\alpha_i^* = 0$ (and vice versa)

Optimization parameters

Apart from the kernel hyper-parameters, ϵ-SVR has two open-parameters:

- C: Cost $[0 \to \infty]$ represents the penalty associated with errors larger than epsilon. Increasing cost value causes closer fitting to the training data.
- ϵ: epsilon represents the minimal required precision.

The total number of hyper-parameters is thus 2 for linear SVR and 3 for nonlinear SVR with RBF kernel.

Determining b

In order to determine the value of b from the optimization program, one can use the KKT conditions. Indeed, for the support vectors that lie on the boundaries of the ϵ-insensitive tube, we have:

$$\xi_i = \xi_i^* = 0$$

and

$$\begin{cases} \alpha_i \in \left]0; \dfrac{C}{M}\right[\\ y_i - \langle w, \mathbf{x}^i \rangle - b - \epsilon = 0 \end{cases} \quad \text{or} \quad \begin{cases} \alpha_i^* \in \left]0; \dfrac{C}{M}\right[\\ y_i - \langle w, \mathbf{x}^i \rangle - b + \epsilon = 0 \end{cases}$$

Hence, we can compute b by averaging the error over the subset of support vectors that lie on the boundary of the ϵ-insensitive tube:

$$\boxed{b = \frac{1}{|I|} \sum_{i \in I} y_i - \langle w, \mathbf{x}^i \rangle - \epsilon_i}$$

or, equivalently

$$\boxed{b = \frac{1}{|I|} \sum_{i \in I} y_i - \sum_{j=1}^{M} \left(\alpha_j - \alpha_j^*\right) k(\mathbf{x}^j, \mathbf{x}^i) - \epsilon_i}$$

where I is the set of indices of the support vectors that lie on the boundary of the ϵ-insensitive tube, i.e. for which $(\alpha_i + \alpha_i^*) \in \left]0; \frac{C}{M}\right[$[4],

$$|I| = \sum_{i=1}^{M} \mathbb{1}_{\left]0;\frac{C}{M}\right[}(\alpha_i + \alpha_i^*) \quad \text{where} \quad \mathbb{1}_{\left]0;\frac{C}{M}\right[}(\alpha) = \begin{cases} 1, & \text{if} \quad \alpha \in \left]0; \frac{C}{M}\right[\\ 0, & \text{otherwise} \end{cases}$$

and

$$\epsilon_i = \begin{cases} +\epsilon, & \text{if} \quad \alpha_i > 0 \\ -\epsilon, & \text{if} \quad \alpha_i^* > 0 \end{cases}$$

Author: Brice Platerrier

[4] This condition is equivalent to $\alpha_i \in \left]0; \frac{C}{M}\right[$ or $\alpha_i^* \in \left]0; \frac{C}{M}\right[$, since either $\alpha_i \alpha_i^*$ and $\alpha_i, \alpha_i^* \geq 0$ (by definition)