Support vector machines

Victor Kitov

Yandex School of Data Analysis

Table of contents

Kuhn-Takker conditions

Consider the optimization task:

$$\begin{cases} f(x) \to \min_x \\ g_i(x) \le 0 \qquad i = 1, 2, \dots m \end{cases} \tag{1}$$

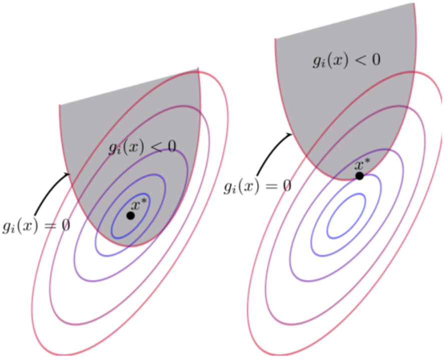
Theorem (necessary conditions for optimality):
Let

- x^* - be the solution to (1),
- $f(x^*)$ and $g_i(x^*)$, $i = 1, 2, \dots m$ - continuously differentiable at x^*.
- one of the conditions of regularity is satisfied

Then coefficients $\lambda_1, \lambda_2, \dots \lambda_m$ exist, such that x^* satisfies the conditions:

$$\begin{cases} \nabla f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) = 0 & \text{stationarity} \\ g_i(x) \le 0 & \text{feasibility} \\ \lambda_i \ge 0 & \text{non-negativity} \\ \lambda_i g_i(x^*) = 0 & \text{complementary slackness} \end{cases} \tag{2}$$

Illustration of constrained optimization



Kuhn-Takker conditions

Possible regularity conditions:

- $\{\nabla g_j, j \in J\}$ - linearly independent, where J - are indexes of active constraints $J = \{j : g_j(x^*) = 0\}$.
- Slater condition: $\exists x : \ g_i(x) < 0 \, \forall i$ (applicable only when $f(x)$ and $g_i(x)$, $i = 1, 2, ...m$ are convex)

Sufficient conditions of optimality:

If $f(x)$ and $g_i(x)$, $i = 1, 2, ...m$ are convex, Kuhn-Takker conditions (2) and Slater conditions become sufficient for x^* to be the solution of (1).

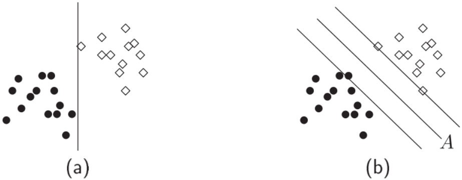
Convex optimization

Why convexity of $f(x)$ and $g_i(x)$, $i = 1, 2, ...m$ is convenient:

- All local minimums become global minimums
- The set of minimums is convex
- If $f(x)$ is strictly convex and minimum exists, then it is unique.

Table of contents

Support vector machines



Support vector machines

(a) (b)

Main idea

Select hyperplane maximizing the margin - the sum of distances from nearest ω_1 object to hyperplane and from nearest ω_2 object to hyperplane.

Support vector machines

Objects x_i for $i = 1, 2, ...n$ lie at distance $b/|w|$ from discriminant hyperplane if

$$\begin{cases} x_i^T w + w_0 \geq b, & y_i = +1 \\ x_i^T w + w_0 \leq -b & y_i = -1 \end{cases} \quad i = 1, 2, ...n.$$

This can be rewritten as

$$y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, ...n.$$

The margin is equal to $2b/|w|$. Since w, w_0 and b are defined up to multiplication constant, we can set $b = 1$.

Problem statement

Problem statement:

$$\begin{cases} w^T w \to \min\limits_{w,w_0} \\ y_i(x_i^T w + w_0) \ge 1, \quad i = 1, 2, ...n. \end{cases}$$

According to Kuhn-Takker theorem, solution satisfies the following problem:

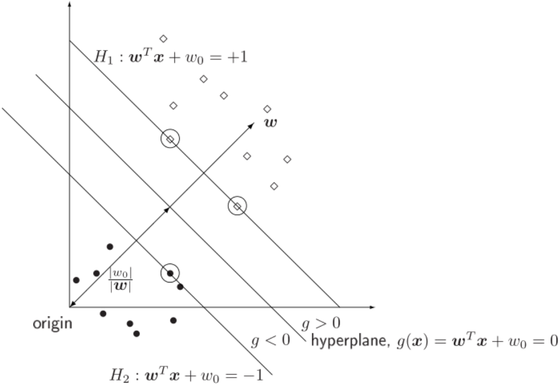
$$L_P = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x + w_0) - 1) \to \min_{w,w_0} \max_{\alpha}, \quad \alpha_i \ge 0, i = 1, 2, ...n$$

with the constraints:

$$\begin{cases} \alpha_i \ge 0, \\ y_i(x_i^T w + w_0) - 1 \ge 0, \\ \alpha_i (y_i (x_i^T w + w_0) - 1) = 0. \end{cases}$$

Support vectors

Condition $\alpha_i(y_i(x_i^T w + w_0) - 1) = 0$ is satisfied when either $\alpha_i = 0$ or $y_i(x_i^T w + w_0) - 1 = 0$. Second case describes support vectors, which lie at distance $1/|w|$ to separating hyperplane and which affect the weights. Other vectors don't affect the solution.



Dual problem

$$\frac{\partial L}{\partial w_0} = 0 : \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} = 0 : \ w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Substituting into Lagrangian L_P, we get:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_{\alpha}$$

α_i can be found from the dual optimization problem:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_{\alpha} \\ \alpha_i \geq 0, \ i = 1, 2, \ldots n; \ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Solution

Denote $\mathcal{SV}$ - the set of indexes of support vectors.
Optimal α_i determine weights directly:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 can be found from any edge equality for support vectors:

$$y_i(x_i^T w + w_0) = 1,\ i \in \mathcal{SV}$$

Solution from summation over n_{SV} equation provides a more robust estimate of w_0:

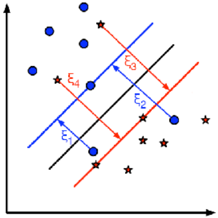
$$n_{\mathcal{SV}} w_0 + \sum_{i \in \mathcal{SV}} x_i^T w = \sum_{i \in \mathcal{SV}} y_i$$

Linearly non-separable case

No separating hyperplane exists. Errors are permitted by including slack variables ξ_i:

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i \to \min_{w,\xi} \\ y_i(w^T x_i + w_0) \ge 1 - \xi_i,\ i = 1, 2, ...n \\ \xi_i \ge 0,\ i = 1, 2, ...n \end{cases}$$

- Parameter C is the cost for misclassification and controls the bias-variance trade-off.
- It is chosen on validation set.
- Other penalties are possible, e.g. $C\sum_i \xi_i^2$.



Linearly non-separable case

According to Karush-Kuhn-Takker theorem, the solution satisfies:

$$L_P = \frac{1}{2}w^T w + C\sum_i \xi_i - \sum_{i=1}^n \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) - \sum_{i=1}^n r_i\xi_i$$

$$L_P \to \min_{w,w_0,\xi} \max_{\alpha,r}$$

under constraints:

$$\begin{cases} \xi_i \geq 0,\ \alpha_i \geq 0,\ r_i \geq 0 \\ y_i(x_i^T w + w_0) \geq 1 - \xi_i, \\ \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) = 0 \\ r_i\xi_i = 0 \end{cases}$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 : \ C - \alpha_i - r_i = 0 \quad \Rightarrow \quad \alpha_i \in [0, C].$$

Classification of training objects

- **Non-informative objects:**
  - have $\alpha_i = 0$ ($\Leftrightarrow r_i = C \Leftrightarrow \xi_i = 0 \Leftrightarrow y_i(w^T x_i + w_0) \geq 1$)
- **Support vectors:**
  - have $\alpha_i > 0$ ($\Leftrightarrow y_i(w^T x_i + w_0) = 1 - \xi_i$)
  - **boundary support vectors:**
    - have $\xi_i = 0$ ($\Leftrightarrow r_i > 0 \Leftrightarrow \alpha_i \in (0, C) \Leftrightarrow y(w^T x_i + w_0) = 1$) then support vector lies at $1/|w|$ distance to separating hyperplane and is called boundary support vector.
  - **violating support vectors:**
    - have $\xi_i > 0$ ($\Leftrightarrow r_i = 0 \Leftrightarrow \alpha_i = C$), so lies closer than $1/|w|$ to separating hyperplane.
    - If $\xi_i \in (0, 1)$ then violating support vector is correctly classified.
    - If $\xi_i > 1$ then violating support vector is misclassified.

Linearly non-separable case - dual problem

$$\frac{\partial L_P}{\partial w_0} = 0 : \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L_P}{\partial w} = 0 : w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 : C - \alpha_i - r_i = 0$$

Substituting these constraints into L_P, we obtain the dual problem:

$$\begin{cases} L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max_\alpha \\ \sum_{i=1}^n \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \end{cases}$$

Solution

Denote $\mathcal{SV}$ - the set of indexes of support vectors with $\alpha_i > 0$ ($\Leftrightarrow y(w^T x_i + w_0) = 1 - \xi_i$) and $\widetilde{\mathcal{SV}}$ - the set of indexes of support vectors with $\alpha_i \in (0, C)$ ($\Leftrightarrow \xi_i = 0$, $y(w^T x_i + w_0) = 1$)

Optimal α_i determine weights directly:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 can be found from any edge equality for support vectors, having $\xi_i = 0$:

$$y_i(x_i^T w + w_0) = 1,\ i \in \widetilde{\mathcal{SV}}$$

Solution from summation of equations for each $i \in \widetilde{SV}$ provides a more robust estimate of w_0:

$$n_{\widetilde{\mathcal{SV}}} w_0 + \sum_{i \in \widetilde{\mathcal{SV}}} x_i^T w = \sum_{i \in \widetilde{\mathcal{SV}}} y_i$$

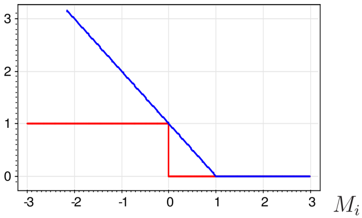
Another view on SVM

Optimization problem:

$$\begin{cases}\frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i \to \min_{w,\xi}\\ y_i(w^T x_i + w_0) = M_i(w, w_0) \geq 1 - \xi_i,\\ \xi_i \geq 0,\ i = 1, 2, ...n\end{cases}$$

can be rewritten as

$$\frac{1}{2C}|w|^2 + \sum_{i=1}^{n}[1 - M_i(w, w_0)]_+ \to \min_{w,\xi}$$



Thus SVM is linear discriminant function with cost approximated with $\mathcal{L}(M) = [1 - M]_+$ and L_2 regularization.