Support Vector Machine Classification: Application of Quadratic Programming and Lagrange duality

Vojtěch Franc

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

xfrancv@cmp.felk.cvut.cz, http://cmp.felk.cvut.cz

Linear classifier

- **Linear classification rule** is $h\colon \mathbb{R}^n \to \{+1,-1\}$ defined by

$$h(\boldsymbol{x};\boldsymbol{w},b) = \begin{cases} +1 & \text{if} \quad \boldsymbol{x}^T\boldsymbol{w} + b > 0 \\ -1 & \text{if} \quad \boldsymbol{x}^T\boldsymbol{w} + b < 0 \end{cases}$$

  where a vector $\boldsymbol{w} \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$ are parameters.

- Linear classifier splits the input space $\mathbb{R}^n$ into three sub-spaces:

$$\begin{aligned} H^+(\boldsymbol{w},b) &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T\boldsymbol{w} + b > 0\} && \text{positive decisions} \\ H^0(\boldsymbol{w},b) &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T\boldsymbol{w} + b = 0\} && \text{hyperplane of undecided inputs} \\ H^-(\boldsymbol{w},b) &= \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T\boldsymbol{w} + b < 0\} && \text{negative decisions} \end{aligned}$$

Linearly separable examples

- **Training examples**

$$\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\} \in (\mathbb{R}^n \times \{+1, -1\})^m$$

- **Linearly separable training examples:** There exist $(\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ such that the linear rule $h(\cdot; \boldsymbol{w}, b)$ classifies all examples in $\mathcal{T}$ correctly, i.e., $(\boldsymbol{w}, b)$ is a solution of

$$\left.\begin{array}{rcll} \boldsymbol{x}_i^T \boldsymbol{w} + b & > & 0\,, & \forall i \in I^+ \\ \boldsymbol{x}_i^T \boldsymbol{w} + b & < & 0\,, & \forall i \in I^- \end{array}\right\} \quad \text{which is the same as} \quad y_i\big(\boldsymbol{x}_i^T \boldsymbol{w} + b\big) > 0\,, \forall i \in I$$

where $I = \{1, \dots, m\}$, $I^+ = \{i \in I \mid y_i = +1\}$ and $I^- = \{i \in I \mid y_i = -1\}$.

- **Separting hyperplane** is any $H^0(\boldsymbol{w}, b) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}^T \boldsymbol{w} + b = 0\}$ such that $(\boldsymbol{w}, b)$ is a solution of $y_i\big(\boldsymbol{x}_i^T \boldsymbol{w} + b\big) > 0\,, \forall i \in I$.

- *Remark:* Note that a given separating hyperplane has infinite number of parametrizations: $H^0(\boldsymbol{w}, b) = H^0(\lambda \boldsymbol{w}, \lambda b)$, $\forall \lambda > 0$.

Finding a separating hyperplane

Task 1: Assume that the training examples $\mathcal{T}$ are linearly separable. The task is to find any separating hyperplane.

- Task 1 requires to find $(\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) > 0\,, \quad \forall i \in I \tag{1}$$

- Provided $(\boldsymbol{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ solves (1) then $\exists \varepsilon > 0$ such that

$$y_i\big(\boldsymbol{x}_i^T \boldsymbol{w} + b\big) \geq \varepsilon\,, \forall i \in I \qquad \Rightarrow \qquad y_i\left(\boldsymbol{x}_i^T \frac{\boldsymbol{w}}{\varepsilon} + \frac{b}{\varepsilon}\right) \geq 1\,, \forall i \in I$$

- Any separating hyperplane $H^0(\boldsymbol{w}', b')$ can be parametrized by $(\boldsymbol{w}, b)$ which satisfies the following set of non-strict linear inequalities

$$y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) \geq 1\,, \quad \forall i \in I \tag{2}$$

- As a result, a separating hyperplane can be found by solving (2) which is an instance of linear programming (with zero objective).

Finding maximal margin hyperplane

Task 2: Assume that the training examples $\mathcal{T}$ are linearly separable. The task is to find the maximal margin separating hyperplane, i.e. a separating hyperplane with the maximal margin

$$m(\boldsymbol{w}, b) = \min_{i \in I} y_i \frac{(\boldsymbol{x}_i^T \boldsymbol{w} + b)}{\|\boldsymbol{w}\|}$$

- Note that the margin $m(\boldsymbol{w}, b)$ is given by a minimal signed distance over the training examples $\mathcal{T}$.
- The signed distance is

$$y_i \frac{(\boldsymbol{x}_i^T \boldsymbol{w} + b)}{\|\boldsymbol{w}\|} = \begin{cases} d(\boldsymbol{x}_i, \boldsymbol{w}, b) & \text{if} \quad h(\boldsymbol{x}_i; \boldsymbol{w}, b) = y_i \\ -d(\boldsymbol{x}_i, \boldsymbol{w}, b) & \text{if} \quad h(\boldsymbol{x}_i; \boldsymbol{w}, b) \neq y_i \end{cases}$$

where

$$d(\boldsymbol{x}, \boldsymbol{w}, b) = \min\{\|\boldsymbol{x} - \boldsymbol{x}'\| \mid \boldsymbol{x}' \in H^0(\boldsymbol{w}, b)\} = \frac{|\boldsymbol{x}^T \boldsymbol{w} + b|}{\|\boldsymbol{w}\|}$$

is the Euclidean distance between $\boldsymbol{x}$ and its closest points on $H^0(\boldsymbol{w}, b)$.

Finding maximal margin hyperplane in canonical form

- The separating hyperplane $H^0(\boldsymbol{w}, b)$ is in a canonical form if

$$\min_{i \in I} y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) = 1$$

which implies that its margin is

$$m(\boldsymbol{w}, b) = \min_{i \in I} y_i \frac{(\boldsymbol{x}_i^T \boldsymbol{w} + b)}{\|\boldsymbol{w}\|} = \frac{1}{\|\boldsymbol{w}\|}$$

- Finding the maximal margin separating hyperplane in a canonical form leads to solving

$$\begin{aligned}
(\boldsymbol{w}^*, b^*) &= \operatorname*{argmax}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{\|\boldsymbol{w}\|} \quad \text{s.t.} \quad \min_{i \in I} y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) = 1 \\
&= \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad \min_{i \in I} y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) = 1 \\
&= \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) \geq 1\,, \forall i \in I
\end{aligned}$$

Finding maximal margin hyperplane by quadratic programming

- Finding the maximal margin hyperplane leads to solving a convex quadratic programming task (**PRIMAL-SVM-QP**)

$$(\boldsymbol{w}^*, b^*) = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}} \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \text{s.t.} \qquad y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) \geq 1\,, \forall i \in I$$

- The resulting linear rule $h(\boldsymbol{x}; \boldsymbol{w}^*, b^*)$ is called the maximal margin classifier.
- The PRIMAL-SVM-QP has $n + 1$ variables and m constraints.
- The SVM classifiers are often used in applications when the dimension n is very large and solving the primal PRIMAL-SVM-QP is not tractable.
- If $n >> m$, solving the PRIMAL-SVM-QP can be replaced by solving its Lagrange dual problem which has m variables and $m + 1$ constraints.

Primal and dual form of the SVM learning problem

- **Lagrange function** of the PRIMAL-SVM-QP reads

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i \Big[y_i \big(\boldsymbol{x}_i^T \boldsymbol{w} + b \big) - 1 \big) \Big]$$

  where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T \in \mathbb{R}^m$ are the Lagrange multipliers.

- **Primal problem**, which is equivalent to PRIMAL-SVM-QP, is defined as

$$(\boldsymbol{w}^*, b^*) = \underset{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}}{\operatorname{argmin}} \; P(\boldsymbol{w}, b)$$

  where

$$P(\boldsymbol{w}, b) = \max \big\{ L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \mid \boldsymbol{\alpha} \succeq \mathbf{0} \big\} = \begin{cases} \infty & \text{if} \quad \exists i \in I,\; y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) < 1 \\ \frac{1}{2}\|\boldsymbol{w}\|^2 & \text{if} \quad y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) \geq 1 \,, \forall i \in I \end{cases}$$

- **Dual problem** is defined as

$$\alpha^* = \underset{\boldsymbol{\alpha} \succeq \mathbf{0}}{\operatorname{argmax}} \, D(\boldsymbol{\alpha}) \quad \text{where} \quad D(\boldsymbol{\alpha}) = \min \big\{ L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \mid \boldsymbol{w} \in \mathbb{R}^n \,, b \in \mathbb{R} \big\}$$

Useful results of the Lagrange duality

- **Weak duality** holds in general

$$P(\boldsymbol{w}, b) \geq P(\boldsymbol{w}^*, b^*) \geq D(\boldsymbol{\alpha}^*) \geq D(\boldsymbol{\alpha})$$

  holds for all feasible $(\boldsymbol{w}, b)$ and $\boldsymbol{\alpha} \succeq \mathbf{0}$.

- **Strong duality** applies for some problems including the PRIMAL-SVM-QP

$$P(\boldsymbol{w}^*, b^*) = D(\boldsymbol{\alpha}^*)$$

- If the strong duality holds and $\boldsymbol{\alpha}^*$ is an optimal solution of the dual, then **the primal solution** $(\boldsymbol{w}^*, b^*)$ is a minimizer of the unconstrained problem

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}^*)$$

- Assume that the strong duality holds and $(\boldsymbol{w}^*, b^*)$ is a primal and $\boldsymbol{\alpha}^*$ dual optimal solution, then the **complementary slackness** holds

$$\alpha_i^* \Big[y_i \big(\boldsymbol{x}_i^T \boldsymbol{w}^* + b^* \big) - 1 \big) \Big] = 0\,, \qquad \forall i \in I$$

Derivation of the SVM dual problem

- By definition the dual objective is

$$D(\boldsymbol{\alpha}) = \min \left\{ L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \mid \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R} \right\}$$

where

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i \Big[y_i \big(\boldsymbol{x}_i^T \boldsymbol{w} + b \big) - 1 \big) \Big]$$

- For a fixed $\boldsymbol{\alpha}$, the $\boldsymbol{w}(\boldsymbol{\alpha})$ minimizing L is obtained by

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1} \alpha_i y_i \boldsymbol{x}_i = \mathbf{0} \qquad \Rightarrow \qquad \boldsymbol{w}(\boldsymbol{\alpha}) = \sum_{i=1} \alpha_i y_i \boldsymbol{x}_i$$

thus

$$L(\boldsymbol{w}(\boldsymbol{\alpha}), b, \boldsymbol{\alpha}) = \sum_{i \in I} \alpha_i - \frac{1}{2} \sum_{i \in I} \sum_{j \in I} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j - b \sum_{i \in I} \alpha_i y_i$$

- Minimizing $L(\boldsymbol{w}(\boldsymbol{\alpha}), b, \boldsymbol{\alpha})$ w.r.t. b yields

$$D(\boldsymbol{\alpha}) = \begin{cases} \boldsymbol{\alpha}^T \boldsymbol{e} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} & \text{if} \quad \boldsymbol{\alpha}^T \boldsymbol{y} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where $\boldsymbol{e}$ is vector of all ones, $\boldsymbol{y} = (y_1, \ldots, y_m)^T$ is a vector containing labels and $\mathbf{H}$ is a symmetric positive semi-definite matrix with $H_{ij} = y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}$.

The dual SVM problem

- The dual of the primal SVM problem is a convex Quadratic Program (**DUAL-SVM-QP**)

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\operatorname{argmax}} \left[\boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}\right] \qquad \text{s.t.} \qquad \boldsymbol{y}^T\boldsymbol{\alpha} = 0\,, \quad \boldsymbol{\alpha} \succeq \mathbf{0}$$

- The DUAL-SVM-QP has m variables and $m+1$ constraints of a simple form.
- Given solution the dual solution $\boldsymbol{\alpha}^*$, the primal solution vector $\boldsymbol{w}^*$ can be obtained by

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{R}^n}{\operatorname{argmin}}\, L(\boldsymbol{w}, b, \boldsymbol{\alpha}^*) = \sum_{i\in I} \alpha_i^* y_i \boldsymbol{x}_i$$

- The optimal b^* can be determined from the complementary slackness (shown on the next slide) or by selecting b^* to satisfy the constraints

$$\boldsymbol{x}_i^T\boldsymbol{w}^* + b^* \geq 1\,, \forall i \in I^+ \qquad \text{and} \qquad \boldsymbol{x}_i^T\boldsymbol{w}^* + b^* \leq -1\,, \forall i \in I^-$$

so that

$$b^* = -\frac{1}{2}\left(\min_{i\in I^+} \boldsymbol{x}_i^T\boldsymbol{w}^* + \max_{i\in I^-} \boldsymbol{x}_i^T\boldsymbol{w}^*\right)$$

Complementary slackness

- The complementary slackness guarantee that

$$\alpha_i^* \Big[y_i \big(\boldsymbol{x}_i^T \boldsymbol{w}^* + b^* \big) - 1 \Big] = 0\,, \qquad \forall i \in I$$

which implies

$$\begin{array}{rcll} y_i \big(\boldsymbol{x}_i^T \boldsymbol{w}^* + b^* \big) & = & 1\,, & \text{for} \quad i \in I^{SV} = \{ i \in I \mid \alpha_i^* > 0 \} \\ y_i \big(\boldsymbol{x}_i^T \boldsymbol{w}^* + b^* \big) & \geq & 1\,, & \text{for} \quad i \in I \setminus I^{SV} \end{array}$$

- The training examples $\{ \boldsymbol{x}_i \mid i \in I^{SV} \}$, called support vectors, have the shortest distance (equal to $\frac{1}{\|\boldsymbol{w}^*\|}$) to the hyperplane $H^0(\boldsymbol{w}^*, b^*)$.
- Removing the support vectors from the training set does not change the solution of the PRIMAL-SVM-QP.
- The optimal b^* can be computed by

$$b^* = y_i - \boldsymbol{x}_i^T \boldsymbol{w}^*\,, \qquad \forall i \in I^{SV}$$

or, for better numerical stability, using the average $b^* = \frac{1}{I^{SV}} \sum_{i \in I^{SV}} \big(y_i - \boldsymbol{x}_i^T \boldsymbol{w}^* \big)$.

Learning SVM classifier from non-separable examples

- **Task 3:** Given examples $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\} \in (\mathbb{R}^n \times \{+1, -1\})^m$, the goal is to find parameters $(\boldsymbol{w}^*, b^*)$ of the linear SVM classifier by solving a convex QP task (PRIMAL-C-SVM-QP)

$$(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^m} \left[\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i \in I} \xi_i \right]$$

subject to

$$\begin{array}{rcll} y_i(\boldsymbol{x}_i^T \boldsymbol{w} + b) & \geq & 1 - \xi_i \,, & \forall i \in I \\ \xi_i & \geq & 0 \,, & \forall i \in I \end{array}$$

- $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)^T \in \mathbb{R}^m$ are the slack variables relaxing the linear inequalities and $C > 0$ is a prescribed constant.
- The sum of the slack variables upper bounds the number of training errors, i.e.

$$\sum_{i \in I} \xi_i \geq \sum_{i \in I} [\![h(\boldsymbol{x}_i; \boldsymbol{w}, b) \neq y_i]\!]$$

- The PRIMAL-C-SVM-QP has $m + n + 1$ variables and $2m$ constraints. The corresponding dual problem has m variables and $2m + 1$ constraints.

Dual SVM problem for non-separable case

- **Lagrange function** of the PRIMAL-C-SVM-QP reads

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\Big[y_i(\boldsymbol{x}_i^T\boldsymbol{w} + b) - 1 + \xi_i)\Big] - \sum_{i=1}^{m}\mu_i\xi_i$$

  where $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}^m$ are the Lagrange multipliers.

- $\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m} y_i\alpha_i\boldsymbol{x}_i = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{m} y_i\alpha_i\boldsymbol{x}_i$
- $\mu_i \geq 0$ and $\mu_i = C - \alpha_i \quad \Rightarrow \quad \sum_{i=1}^{m} \xi_i(C - \mu_i - \alpha_i) = 0$
- $\sum_{i=1}^{m} \alpha_i y_i = 0$
- The dual objective $D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^m} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ simplifies to

$$D(\boldsymbol{\alpha}) = \begin{cases} \boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} & \text{if} \quad \boldsymbol{\alpha}^T\boldsymbol{y} = 0 \quad \text{and} \quad C\boldsymbol{e} \succeq \boldsymbol{\alpha} \succeq \mathbf{0} \\ \infty & \text{otherwise} \end{cases}$$

- The dual problem of the PRIMAL-C-SVM-QP is a convex QP

$$\alpha^* = \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^m}\left[\boldsymbol{\alpha}^T\boldsymbol{e} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}\right] \quad \text{s.t.} \quad \boldsymbol{\alpha}^T\boldsymbol{y} = 0\,, \quad C\boldsymbol{e} \succeq \boldsymbol{\alpha} \succeq \mathbf{0}$$