Fundamentals of Linear Algebra and Optimization
Soft Margin Support Vector Machines

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu

May 6, 2020

Soft Margin Support Vector Machine ($\mathrm{SVM}_{s2'}$)

In this section we consider the following version of the soft margin support vector machine:

Soft margin SVM ($\mathrm{SVM}_{s2'}$):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2} w^\top w - K_m \eta + K_s \begin{pmatrix} \epsilon^\top & \xi^\top \end{pmatrix} \mathbf{1}_{p+q} \\
&\text{subject to} \\
&\qquad w^\top u_i - b \geq \eta - \epsilon_i, \quad \epsilon_i \geq 0 \qquad i = 1, \ldots, p \\
&\qquad -w^\top v_j + b \geq \eta - \xi_j, \quad \xi_j \geq 0 \qquad j = 1, \ldots, q \\
&\qquad\qquad \eta \geq 0.
\end{aligned}
$$

Soft Margin Support Vector Machine ($\mathrm{SVM}_{s2'}$)

This version of the SVM problem was first discussed in Schölkopf, Smola, Williamson, and Bartlett under the name of ν-*SVC* (or ν-*SVM*), and also used in Schölkopf, Platt, Shawe–Taylor, and Smola.

For this problem it is no longer clear that if $(w, \eta, b, \epsilon, \xi)$ is an optimal solution, then $w \neq 0$ and $\eta > 0$.

In fact, if the sets of points are not linearly separable and if K_s is chosen too big, Problem $(\mathrm{SVM}_{s2'})$ *may fail to have an optimal solution*.

Conditions for Existence of an Optimal Solution

We show that in order for the problem to have a solution we must pick K_m and K_s so that

$$K_m \leq \min\{2pK_s, 2qK_s\}.$$

If we define ν by

$$\nu = \frac{K_m}{(p+q)K_s},$$

then $K_m \leq \min\{2pK_s, 2qK_s\}$ is equivalent to

$$\nu \leq \min\left\{\frac{2p}{p+q}, \frac{2q}{p+q}\right\} \leq 1.$$

Soft Margin Support Vector Machine ($\mathrm{SVM}_{s2'}$)

The reason for introducing ν is that $\nu(p+q)/2$ can be interpreted as the maximum number of points failing to achieve the margin $\delta = \eta/\|w\|$.

We will show later that if the points u_i and v_j are not separable, then we must pick ν so that $\nu \geq 2/(p+q)$ for the method to have a solution for which $w \neq 0$ and $\eta > 0$.

The objective function of our problem is convex and the constraints are affine.

Soft Margin Support Vector Machine ($\mathrm{SVM}_{s2'}$)

Consequently, by the duality theorem *if* the primal problem $(\mathrm{SVM}_{s2'})$ has an optimal solution, *then* the dual problem has a solution too, and the *duality gap is zero*.

This *does not* immediately imply that an optimal solution of the dual yields an optimal solution of the primal because the hypotheses of the duality theorem fail to hold.

$(\mathrm{SVM}_{s2'})$ *Notational Conventions*

Let X be the $n \times (p+q)$ matrix

$$X = \begin{pmatrix} -u_1 & \cdots & -u_p & v_1 & \cdots & v_q \end{pmatrix}.$$

- Let $\lambda \in \mathbb{R}^p_+$ be the Lagrange multipliers for $w^\top u_i - b \geq \eta - \epsilon_i$,
- Let $\mu \in \mathbb{R}^q_+$ be the Lagrange multipliers for $-w^\top v_j + b \geq \eta - \xi_j$.
- Let $\alpha \in \mathbb{R}^p_+$ be the Lagrange multipliers associated with $\epsilon_i \geq 0$.
- Let $\beta \in \mathbb{R}^q_+$ be the Lagrange multipliers associated with $\xi_j \geq 0$.
- Let $\gamma \in \mathbb{R}^+$ be the Lagrange multiplier associated with $\eta \geq 0$.

Determining **w** *of an Optimal Solution*

We show that *if* the primal problem has an optimal solution $(w, \eta, \epsilon, \xi, b)$ with $w \neq 0$, then *any* optimal solution of the dual problem determines λ and μ, which in turn determine w *via* the equation

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p} \lambda_i u_i - \sum_{j=1}^{q} \mu_j v_j. \qquad (*_w)$$

Illustration of a Soft Margin SVM

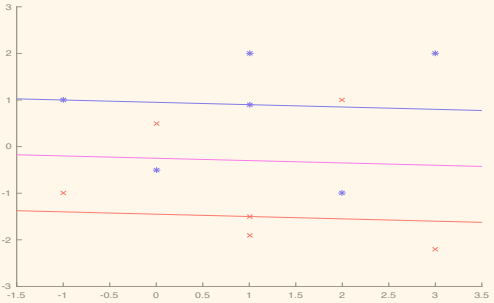

Figure 1: Soft margin ν-SVM for two sets of six points for $\nu = 0.6$. There are two sets of six points each. Two blue points are on the wrong side of the separating hyperplane and two red points are on the wrong side of the separating hyperplane. Two blue points are on the wrong side of the blue margin and three red points are on the wrong side of the red margin. Two blue points are on the blue margin and one red points is on the the red margin.

Lagrangian of the Optimization Problem

The derivation of the Lagrangian of the above optimization problem is somewhat laborious. After some algebra, the Lagrangian can be written as

$$\begin{aligned} L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma) = \frac{1}{2} w^\top w + w^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + (\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - K_m - \gamma)\eta \\ + \epsilon^\top (K_s \mathbf{1}_p - (\lambda + \alpha)) + \xi^\top (K_s \mathbf{1}_q - (\mu + \beta)) \\ + b(\mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu). \end{aligned}$$

To find the dual function $G(\lambda, \mu, \alpha, \beta, \gamma)$, we minimize $L(w, \epsilon, \xi, b, \eta, \lambda, \mu, \alpha, \beta, \gamma)$ with respect to w, ϵ, ξ, b, and η.

Determining the Dual From the Lagrangian

Since the Lagrangian is convex and $(w, \epsilon, \xi, b, \eta) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R} \times \mathbb{R}$, a convex open set, by a previous theorem, the Lagrangian has a minimum in $(w, \epsilon, \xi, b, \eta)$ iff $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$.

So we compute its gradient with respect to $w, \epsilon, \xi, b, \eta$, and we get

$$\nabla L_{w,\epsilon,\xi,b,\eta} = \begin{pmatrix} X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} + w \\ K_s \mathbf{1}_p - (\lambda + \alpha) \\ K_s \mathbf{1}_q - (\mu + \beta) \\ \mathbf{1}_p^\top \lambda - \mathbf{1}_q^\top \mu \\ \mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu - K_m - \gamma \end{pmatrix}.$$

Determining the Dual From the Lagrangian

By setting $\nabla L_{w,\epsilon,\xi,b,\eta} = 0$ we get the equations

$$w = -X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \qquad (*_w)$$

$$\begin{aligned} \lambda + \alpha &= K_s \mathbf{1}_p \\ \mu + \beta &= K_s \mathbf{1}_q \\ \mathbf{1}_p^\top \lambda &= \mathbf{1}_q^\top \mu, \end{aligned}$$

and

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu = K_m + \gamma. \qquad (*_\gamma)$$

The Box Constraints

The second and third equations are equivalent to the *box constraints*

$$0 \leq \lambda_i, \mu_j \leq K_s, \quad i = 1, \ldots, p,\ j = 1, \ldots, q,$$

and since $\gamma \geq 0$ equation $(*_\gamma)$ is equivalent to

$$\mathbf{1}_p^\top \lambda + \mathbf{1}_q^\top \mu \geq K_m.$$

Dual Function of $(\mathrm{SVM}_{s2'})$

Plugging back w from $(*_w)$ into the Lagrangian, after simplifications we get

$$\begin{aligned} G(\lambda, \mu, \alpha, \beta) &= \frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} - \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \\ &= -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}, \end{aligned}$$

so the dual function is independent of α, β and is given by

$$G(\lambda, \mu) = -\frac{1}{2} \begin{pmatrix} \lambda^\top & \mu^\top \end{pmatrix} X^\top X \begin{pmatrix} \lambda \\ \mu \end{pmatrix}.$$

Finally, the dual program is equivalent to the minimization program:

Dual Program of $(\mathrm{SVM}_{s2'})$

Dual of Soft margin SVM $(\mathrm{SVM}_{s2'})$:

$$\text{minimize} \quad \frac{1}{2}\begin{pmatrix}\lambda^\top & \mu^\top\end{pmatrix} X^\top X \begin{pmatrix}\lambda \\ \mu\end{pmatrix}$$

subject to

$$\sum_{i=1}^{p} \lambda_i - \sum_{j=1}^{q} \mu_j = 0$$

$$\sum_{i=1}^{p} \lambda_i + \sum_{j=1}^{q} \mu_j \geq K_m$$

$$0 \leq \lambda_i \leq K_s, \quad i = 1, \ldots, p$$

$$0 \leq \mu_j \leq K_s, \quad j = 1, \ldots, q.$$

Solving the Dual and Computing w

It is shown in a following section how the dual program is solved using ADMM.

If the primal problem is solvable, this yields solutions for λ and μ. Once a solution for λ and μ is obtained, we have

$$w = -X\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \sum_{i=1}^{p}\lambda_i u_i - \sum_{j=1}^{q}\mu_j v_j.$$

It remains to determine b, η, ϵ and ξ. The solution of the dual *does not* determine b, η, ϵ, ξ directly, and we are not aware of necessary and sufficient conditions that ensure that they can be determined.

Computing Remaining $(\mathrm{SVM}_{s2'})$ *Parameters*

The best we can do is to use the KKT conditions.

If $(w, \eta, \epsilon, \xi, b)$ is an optimal solution of Problem $(\mathrm{SVM}_{s2'})$ with $w \neq 0$ and $\eta \neq 0$, then the complementary slackness conditions yield a classification of the points u_i and v_j in terms of the values of λ and μ.

Indeed, we have $\epsilon_i \alpha_i = 0$ for $i = 1, \ldots, p$ and $\xi_j \beta_j = 0$ for $j = 1, \ldots, q$.

Also, if $\lambda_i > 0$, then the corresponding constraint is active, and similarly if $\mu_j > 0$. Since $\lambda_i + \alpha_i = K_s$, it follows that $\epsilon_i \alpha_i = 0$ iff $\epsilon_i(K_s - \lambda_i) = 0$, and since $\mu_j + \beta_j = K_s$, we have $\xi_j \beta_j = 0$ iff $\xi_j(K_s - \mu_j) = 0$.

Standard Margin Hypothesis

In order to determine b and η we assume the following condition:

Standard Margin Hypothesis for $(\mathrm{SVM}_{s2'})$: There is some i_0 such that $0 < \lambda_{i_0} < K_s$, and there is some μ_{j_0} such that $0 < \mu_{j_0} < K_s$.

Under the **Standard Margin Hypothesis** for $(\mathrm{SVM}_{s2'})$, there is some i_0 such that $0 < \lambda_{i_0} < K_s$ and some j_0 such that $0 < \mu_{j_0} < K_s$, and by the complementary slackness conditions $\epsilon_{i_0} = 0$ and $\xi_{j_0} = 0$, so we have the *two active constraints*

$$w^\top u_{i_0} - b = \eta, \quad -w^\top v_{j_0} + b = \eta,$$

Standard Margin Hypothesis

and we can solve for b and η and we get

$$
\begin{aligned}
b &= \frac{w^\top u_{i_0} + w^\top v_{j_0}}{2} \\
\eta &= \frac{w^\top u_{i_0} - w^\top v_{j_0}}{2} \\
\delta &= \frac{\eta}{\|w\|}.
\end{aligned}
$$

Computing Soft Margin SVM Parameters

Due to numerical instability, when writing a computer program it is preferable to compute the lists of indices I_λ and I_μ given by

$$I_\lambda = \{i \in \{1, \ldots, p\} \mid 0 < \lambda_i < K_s\}$$
$$I_\mu = \{j \in \{1, \ldots, q\} \mid 0 < \mu_j < K_s\}.$$

Then it is easy to compute b and η using the following averaging formulae:

$$b = w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| + \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu| \right) /2$$

$$\eta = w^\top \left(\left(\sum_{i \in I_\lambda} u_i \right) / |I_\lambda| - \left(\sum_{j \in I_\mu} v_j \right) / |I_\mu| \right) /2.$$