The Mathematics of Machine Learning 2025

Lecture 9

Tim van Erven

University of Amsterdam

Outline

Support Vector Machines Refresher

1. Intuition and Terminology
2. Four Equivalent Formalizations

Convex Optimization (Optional!)

3. Convex Problems
4. Example: Dual Formulation of SVMs
5. KKT Optimality Conditions
6. Example: Primal SVM Solution from Dual Solution

Intuition and Terminology

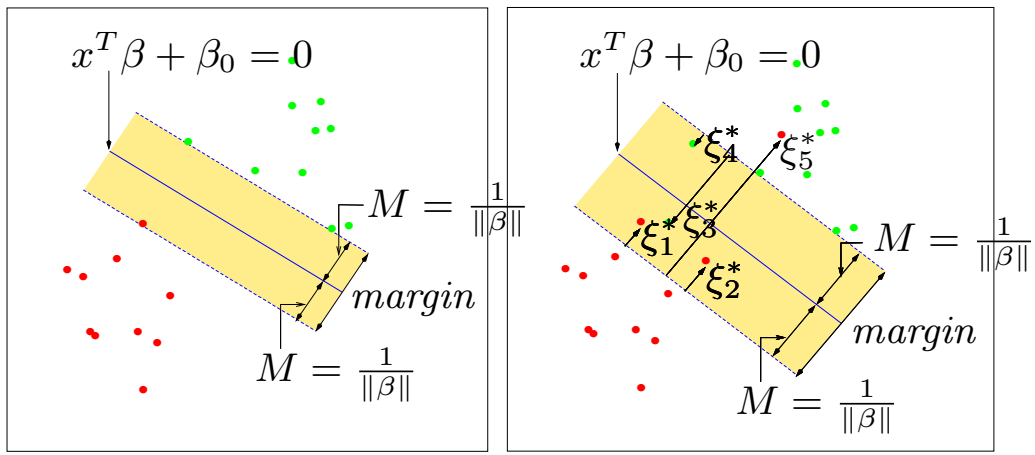
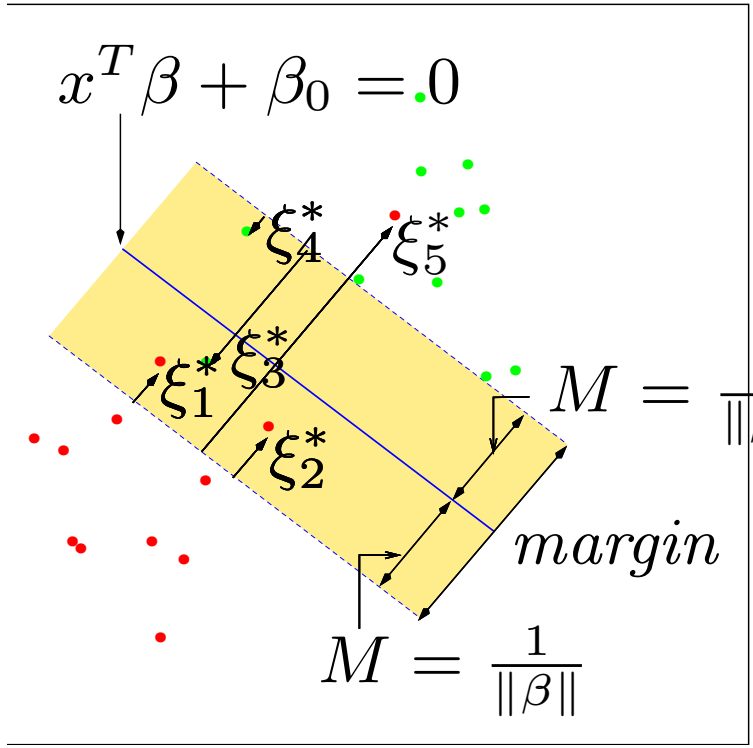


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width* $2M = 2/\|\beta\|$. *The right panel shows the nonseparable (overlap) case. The points labeled* ξ_j^* *are on the wrong side of their margin by an amount* $\xi_j^* = M\xi_j$; *points on the correct side have* $\xi_j^* = 0$. *The margin is maximized subject to a total budget* $\sum \xi_i \leq$ *constant. Hence* $\sum \xi_j^*$ *is the total distance of points on the wrong side of their margin.*

Formalizing Intuition



$$\max_{\beta_0,\beta,M,\xi_1,\ldots,\xi_N} \quad M$$

subject to: $\xi_i \geq 0, \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1-\xi_i) \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

Reformulation as Convex Optimization Problem

Intuitive Formulation:

$$\max_{\beta_0,\beta,M,\xi_1,\ldots,\xi_N} \quad M$$

$$\text{subject to:} \quad \xi_i \geq 0, \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1-\xi_i) \qquad \text{for } i = 1, \ldots, N$$

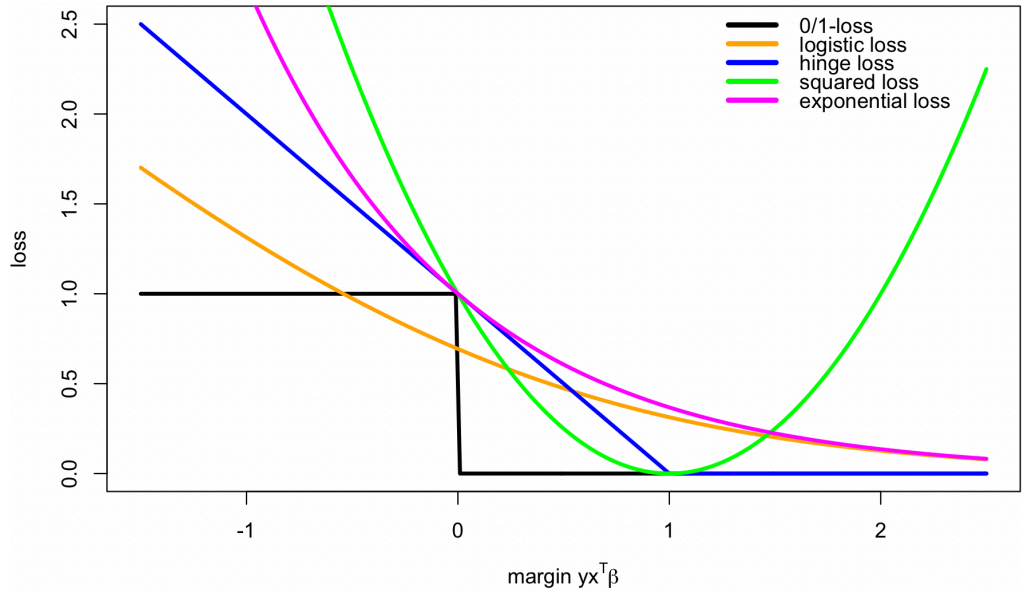
$$\sum_{i=1}^{N} \xi_i \leq t$$

Convex Optimization Problem:

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{subject to:} \quad \xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad \text{for } i = 1, \ldots, N$$

Reformulation as Penalized ERM with Hinge Loss



Hinge loss: $L(Y_i, f(X_i)) = \max\{0, 1 - Y_i f(X_i)\}$

Solution $(\hat{\beta}_0, \hat{\beta})$ to ★ equals solution to:

$$\min_{\beta_0, \beta} \quad \sum_{i=1}^{N} L(Y_i, X_i^\top \beta + \beta_0) + \frac{1}{2C}\|\beta\|^2$$

ERM for hinge loss with ℓ_2-penalty, $\lambda = \frac{1}{2C}$.

Dual Formulation

$X_i^\top X_k$

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k Y_i Y_k \langle X_i, X_k \rangle$$

subject to: $\quad 0 \leq \alpha_i \leq C \qquad \text{for } i = 1, \ldots, N$

$$\sum_{i=1}^{N} \alpha_i Y_i = 0$$

Dual Formulation

$X_i^\top X_k$

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N \alpha_i\alpha_k Y_i Y_k \langle X_i, X_k\rangle$$

subject to: $\quad 0 \le \alpha_i \le C \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^N \alpha_i Y_i = 0$$

Then solution to $\bigstar$ is:

$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i Y_i X_i$$

$$\hat{\beta}_0 = Y_i - X_i^\top \hat{\beta} \qquad \text{for any } i \text{ s.t. } 0 < \hat{\alpha}_i < C.$$

Dual Formulation

$X_i^\top X_k$

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N \alpha_i\alpha_k Y_i Y_k \langle X_i, X_k\rangle$$

subject to: $\quad 0 \le \alpha_i \le C \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^N \alpha_i Y_i = 0$$

Then solution to ★ is:

$$\hat\beta = \sum_{i=1}^N \hat\alpha_i Y_i X_i$$

$$\hat\beta_0 = Y_i - X_i^\top \hat\beta \qquad \text{for any } i \text{ s.t. } 0 < \hat\alpha_i < C.$$

$\hat\alpha_i = 0$ for X_i outside margin and correctly classified

$0 < \hat\alpha_i < C$ for X_i exactly on margin and correctly classified

$\hat\alpha_i = C$ for X_i strictly inside margin or misclassified

Outline

Support Vector Machines Refresher

1. Intuition and Terminology
2. Four Equivalent Formalizations

Convex Optimization (Optional!)

3. Convex Problems
4. Example: Dual Formulation of SVMs
5. KKT Optimality Conditions
6. Example: Primal SVM Solution from Dual Solution

Optional

Today we do a speed run of convex optimization, which is a very interesting research field in itself.

This is optional background material, and will not be tested on the exam.

2. Convex Optimization (Optional!)

a) Convex Problems:

$$\min_{x \in \mathbb{R}^d} \quad f(x) \qquad \text{(} f \text{ convex)}$$

subject to constraints

$$g_i(x) \leq 0 \qquad i = 1, \ldots, m \qquad \text{(} g_i \text{ convex)}$$

$$h_j(x) = 0 \qquad j = 1, \ldots, k \qquad \text{(} h_j \text{ affine)}$$

"Primal Problem"

b) Dual Problem

$\lambda = (\lambda_1, \ldots, \lambda_m)$

$\mu = (\mu_1, \ldots, \mu_k)$

Lagrangian:

$$\mathcal{L}(z, \lambda, \mu) = f(z) + \sum_{i=1}^{m} \lambda_i g_i(z) + \sum_{j=1}^{k} \mu_j h_j(z)$$

$$\sup_{\substack{\lambda : \lambda_i \geq 0, \\ \mu}} \mathcal{L}(z, \lambda, \mu) = \begin{cases} f(z) & \text{if } z \text{ satisfies constraints} \\ \infty & \text{otherwise} \end{cases}$$

Primal Problem:

$$\min_{z} \sup_{\substack{\lambda : \lambda_i \geq 0 \\ \mu}} \mathcal{L}(z, \lambda, \mu)$$

Weak Duality: Always holds!

$$\min_{z} \sup_{\substack{\lambda,\mu \\ \lambda_i \geq 0}} \mathcal{L}(z,\lambda,\mu) \geq \sup_{\substack{\lambda,\mu \\ \lambda_i \geq 0}} \min_{z} \mathcal{L}(z,\lambda,\mu)$$

Strong Duality:

$$\min_{z} \sup_{\substack{\lambda,\mu \\ \lambda_i \geq 0}} \mathcal{L}(z,\lambda,\mu) = \sup_{\substack{\lambda,\mu \\ \lambda_i \geq 0}} \min_{z} \mathcal{L}(z,\lambda,\mu)$$

Holds e.g. under <u>Slater's condition</u>:

Exists "strictly feasible" point $\tilde{z}$:

$$g_i(\tilde{z}) < 0 \quad \text{for all } i$$
$$h_j(\tilde{z}) = 0 \quad \text{for all } j$$

(N.B. strict inequality!)

Dual Problem:

$$\max_{\lambda, \mu} \quad d(\lambda, \mu) \quad \text{where } d(\lambda, \mu) = \inf_{z} L(z, \lambda, \mu)$$

$$\text{s.t.} \quad \lambda_i \geq 0 \quad \text{for } i = 1, \ldots, m$$

* $d(\lambda, \mu)$ concave, so $-d(\lambda, \mu)$ convex
* value of dual is lower bound on value of primal,
  with equality if strong duality holds.

Example: dual formulation SVMs

Primal:

$$\min_{\beta, \beta_0, \xi} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } \xi_i \geq 0 \;,\; y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \quad i = 1, \ldots, N$$

$$z = (\beta, \beta_0, \xi)$$

$$f(z) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$g_i(z) = -y_i(x_i^T\beta + \beta_0) + (1 - \xi_i) \quad \text{for } i = 1, \ldots, N$$

$$g_{N+i}(z) = -\xi_i \quad \text{for } i = 1, \ldots, N$$

(No equality constraints.)

Slater's condition satisfied for $\beta = 0$, $\beta_0 = 0$, $\xi_i = 2$.

Dual :

$$J(\alpha, \nu) = \inf_{\beta, \beta_0, \xi} \mathcal{L}(\beta, \beta_0, \xi, \alpha, \nu)$$

$$\mathcal{L}(\beta, \beta_0, \xi, \alpha, \nu) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$- \sum_{i=1}^{N} \alpha_i \left(y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)\right)$$

$$- \sum_{i=1}^{N} \nu_i \xi_i$$

$$0 = \nabla_\beta \mathcal{L} = \beta - \sum_{i=1}^{N} \alpha_i y_i x_i \quad \Rightarrow \quad \beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$0 = \frac{\partial}{\partial \beta_0} \mathcal{L} = -\sum_{i=1}^{N} \alpha_i y_i \quad \Rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (2)$$

$$0 = \frac{\partial}{\partial \xi_i} \mathcal{L} = C - \alpha_i - v_i \Rightarrow v_i = C - \alpha_i \quad (2)$$

If (1) + (2), then:

$$d(\alpha, v) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \langle x_i, x_k \rangle$$

$$- \sum_{i=1}^{N} \alpha_i y_i x_i^T \sum_{k=1}^{N} \alpha_k y_k x_k + \sum_{i=1}^{N} \alpha_i$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i \alpha_k y_i y_k \langle x_i, x_k \rangle$$

If not (1) or not (2), then

$$d(\alpha, \mu) = -\infty$$

Dual Problem:

$$\max_{\alpha, \nu} \quad d(\alpha, \nu)$$

$$\text{s.t} \quad \alpha_i \geq 0, \quad \nu_i \geq 0 \qquad \text{for } i = 1, \ldots, N$$

$$\max_{\alpha, \nu} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i \alpha_k y_i y_k \langle x_i, x_k \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad \nu_i \geq 0, \quad \nu_i = C - \alpha_i \qquad \text{for } i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i \alpha_k y_i y_k \langle x_i, x_k \rangle$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \qquad \text{for } i = 1, \ldots, N,$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

c) Karush-Kuhn-Tucker (KKT) optimality conditions

If strong duality:

z^* primal optimal $\Leftrightarrow$ $\exists \lambda^*, \nu^*$ such that

(z^*, λ^*, ν^*) satisfy

KKT conditions

(and then λ^*, ν^* are dual optimal)

KKT conditions:

$$g_i(z^*) \le 0 \quad \text{for } i = 1, \dots, m$$
$$h_j(z^*) = 0 \quad \text{for } j = 1, \dots, k$$

} (primal feasibility)

$$\lambda_i^* \ge 0 \quad \text{for } i = 1, \dots, m$$ (dual feasibility)

$$\lambda_i^* g_i(z^*) = 0 \quad \text{for } i = 1, \dots, m$$

(complementary slackness)

$$\nabla_z \mathcal{L}(z^*, \lambda^*, \nu^*) = 0$$ (z^* minimizes $\mathcal{L}(z, \lambda^*, \nu^*)$)

Example: Primal SVM solution from Dual solution

$$0 = \nabla_\beta \mathcal{L}(\beta, \beta_0, \xi, \alpha^*, \nu^*) \quad \Rightarrow \quad \beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\alpha_i \left(-y_i (x_i^T \beta + \beta_0) + (1 - \xi_i)\right) = 0$$

$$(C - \alpha_i)\, \xi_i = \nu_i \xi_i = 0$$

For any i such that

$$0 < \alpha_i < C \Rightarrow \begin{cases} \xi_i = 0 \\ \alpha_i \left(y_i (x_i^T \beta + \beta_0) - 1\right) = 0 \end{cases}$$

use to solve for β_0.