Lecture 12

Soft Margin SVM

Recall the hard-margin SVM

Labels $y_i \in \{-1, +1\}$ $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$

$$f(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0$$

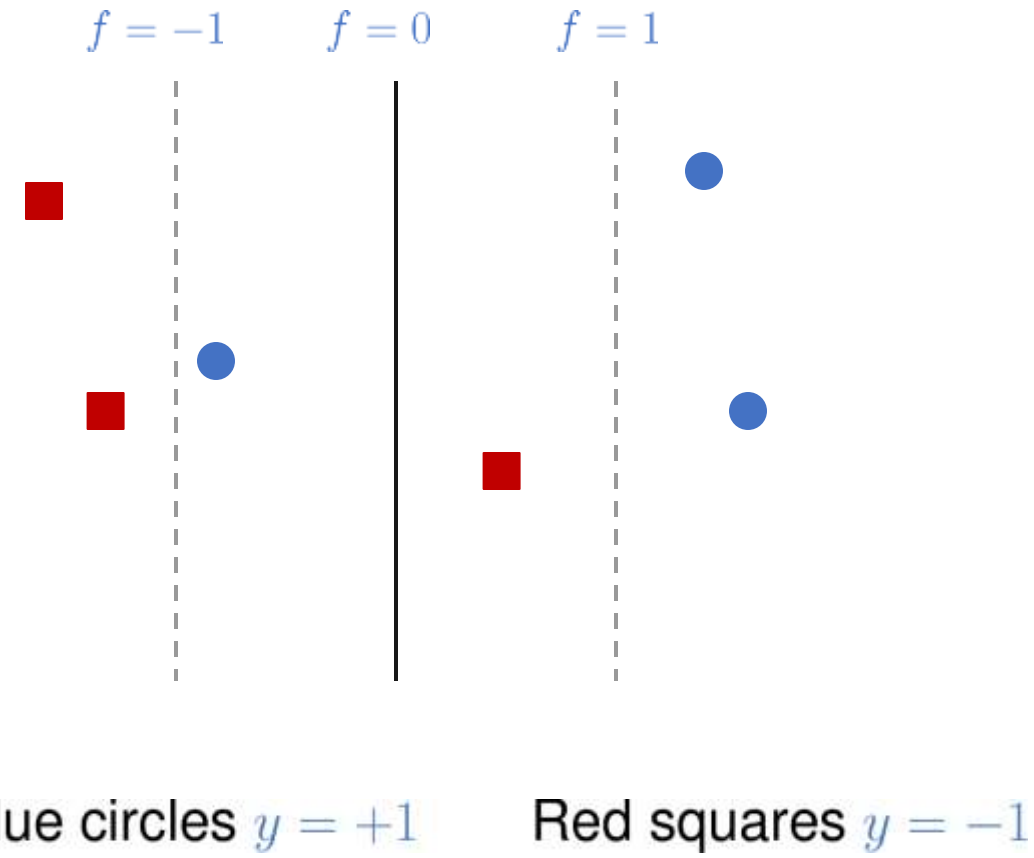
$$\min \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2$$

over $\boldsymbol{\beta}$ and β_0

subject to $y_i(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \beta_0) \geq 1$

Every training observation must lie on or beyond its class margin.

When hard margin is infeasible



Overlapping classes can make the hard-margin constraints infeasible.

Allow observations to violate the margin by a controlled amount.

Slack measures margin violation

$$y_i f(\mathbf{x}_i) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$$

At the optimum with a positive slack penalty

$$\zeta_i = \max\{0, 1 - y_i f(\mathbf{x}_i)\}$$

$\zeta_i = 0$	On or beyond the correct margin
$0 < \zeta_i < 1$	Correctly classified, inside the margin
$\zeta_i = 1$	On the decision boundary
$\zeta_i > 1$	Misclassified

Why slack needs a penalty

Without a penalty, $\beta = 0$, $\beta_0 = 0$ and $\zeta_i = 1$ satisfy every relaxed constraint.

$$\frac{1}{2}\|\beta\|^2 + \gamma \sum_i \zeta_i, \quad \gamma > 0$$

The first term encourages a wide margin.

The second penalizes the total amount of violation.

Larger γ makes violations more expensive. Choose γ by validation.

The soft-margin primal problem

$$\min \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \zeta_i$$

over $\boldsymbol{\beta}$, β_0 and $\zeta_1, \ldots, \zeta_n$

$$y_i(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) - 1 + \zeta_i \geq 0$$

$$\zeta_i \geq 0, \quad i = 1, \ldots, n$$

Two families of inequalities require two families of multipliers.

Forming the Lagrangian

$\alpha_i \geq 0$ for the relaxed margin constraint.
$\lambda_i \geq 0$ for the nonnegative slack constraint.

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_i \zeta_i$$

$$- \sum_i \alpha_i [y_i(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{x}_i + \beta_0) - 1 + \zeta_i]$$

$$- \sum_i \lambda_i \zeta_i$$

For constraints written as $g \geq 0$, subtract multiplier $\times\, g$.

Stationarity

$$\nabla_{\boldsymbol{\beta}} \mathcal{L} = \boldsymbol{\beta} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$

$$\boldsymbol{\beta} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = -\sum_i \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \zeta_i} = \gamma - \alpha_i - \lambda_i = 0$$

Hence $\sum_i \alpha_i y_i = 0$ and $\alpha_i + \lambda_i = \gamma$.

The remaining KKT conditions

Primal feasibility

$$y_i f(\mathbf{x}_i) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$$

Dual feasibility

$$\alpha_i \geq 0, \quad \lambda_i \geq 0$$

Complementary slackness

$$\alpha_i [y_i f(\mathbf{x}_i) - 1 + \zeta_i] = 0$$

$$\lambda_i \zeta_i = 0$$

Substituting into the Lagrangian

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \boldsymbol{\beta}^\mathsf{T} \sum_i \alpha_i y_i \mathbf{x}_i$$
$$-\beta_0 \sum_i \alpha_i y_i + \sum_i \alpha_i$$
$$+ \sum_i (\gamma - \alpha_i - \lambda_i)\zeta_i$$

Substitute $\boldsymbol{\beta} = \sum_i \alpha_i y_i \mathbf{x}_i$. The β_0 and slack terms vanish.

$$q(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\mathsf{T} \mathbf{x}_j$$

The soft-margin dual

$$\max \quad q(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\mathsf{T} \mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \gamma$$

Why the upper bound?

$$\alpha_i + \lambda_i = \gamma, \quad \lambda_i \geq 0 \qquad \Longrightarrow \qquad \alpha_i \leq \gamma$$

The objective matches hard margin. Soft margin adds an upper bound.

Case 1 $\alpha_i = 0$

$\lambda_i = \gamma > 0$ and $\lambda_i \zeta_i = 0$

$\zeta_i = 0 \quad \Longrightarrow \quad y_i f(\mathbf{x}_i) \geq 1$

The observation is on or beyond its correct margin.

Its coefficient in $\boldsymbol{\beta} = \sum_j \alpha_j y_j \mathbf{x}_j$ is zero.

Equality is possible. $\alpha_i = 0$ does not imply a strict inequality.

Case 2 $0 < \alpha_i < \gamma$

$$\lambda_i = \gamma - \alpha_i > 0 \qquad \Longrightarrow \qquad \zeta_i = 0$$

$$\alpha_i > 0 \qquad \Longrightarrow \qquad y_i f(\mathbf{x}_i) - 1 + \zeta_i = 0$$

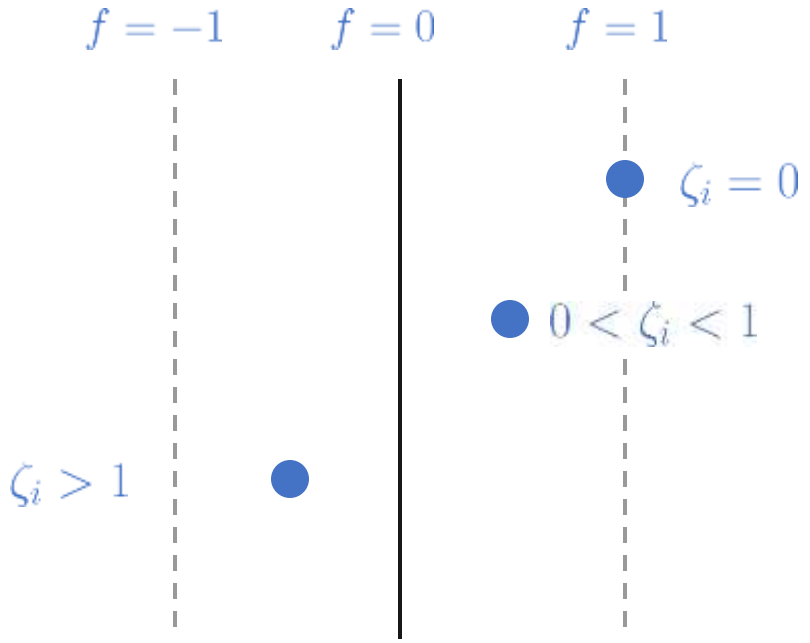
$$y_i f(\mathbf{x}_i) = 1$$

This is a support vector exactly on its class margin.

We can use this equality to compute β_0.

Case 3 $\alpha_i = \gamma$

$\lambda_i = 0$ and $y_i f(\mathbf{x}_i) = 1 - \zeta_i \leq 1$



A bounded support vector may lie
on the margin, inside it,
or on the wrong side.

$\zeta_i > 0$ implies $\alpha_i = \gamma$.
The converse need not hold.

Blue circles represent $y = +1$. The vertical lines mark $f = -1, 0, 1$.

The dual with a kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{x}_j)$$

$$\max \quad \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \gamma$$

The constraints stay the same. Training uses only kernel evaluations.

Prediction without computing ϕ explicitly

$$\boldsymbol{\beta} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)$$

$$\boldsymbol{\beta}^\mathsf{T} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)^\mathsf{T}$$

$$f(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T} \phi(\mathbf{x}) + \beta_0$$

$$= \sum_i \alpha_i y_i \phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{x}) + \beta_0$$

$$= \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + \beta_0$$

Computing the intercept β_0

Choose a support vector $\mathbf{x}_s$ with $0 < \alpha_s < \gamma$.

$$y_s f(\mathbf{x}_s) = 1 \qquad \Longrightarrow \qquad f(\mathbf{x}_s) = y_s$$

$$\sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_s) + \beta_0 = y_s$$

$$\beta_0 = y_s - \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_s)$$

In practice, average this value over all such support vectors.

If no multiplier is strictly between 0 and γ

Let $g_i = \sum_j \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}_i)$, so $f(\mathbf{x}_i) = g_i + \beta_0$.

Multiplier	Label	KKT bound on β_0
$\alpha_i = 0$	$y_i = +1$	$\beta_0 \geq 1 - g_i$
$\alpha_i = 0$	$y_i = -1$	$\beta_0 \leq -1 - g_i$
$\alpha_i = \gamma$	$y_i = +1$	$\beta_0 \leq 1 - g_i$
$\alpha_i = \gamma$	$y_i = -1$	$\beta_0 \geq -1 - g_i$

Take the largest lower bound and smallest upper bound.

Choose β_0 in that interval, for example its midpoint when finite.

The complete classifier

1. Choose γ and a kernel using validation.

2. Solve the dual for $\alpha_1, \ldots, \alpha_n$.

3. Compute β_0 from the KKT conditions.

$$f(\mathbf{x}) = \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + \beta_0$$

$h(\mathbf{x}) = +1$ if $f(\mathbf{x}) \geq 0$, and -1 otherwise