

CS480/680: Introduction to Machine Learning

Lecture 6: Soft-Margin Support Vector Machines

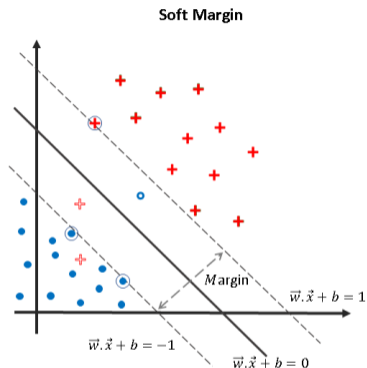
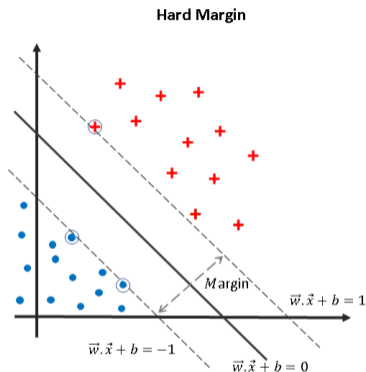
Hongyang Zhang



UNIVERSITY OF
WATERLOO

May 29&31, 2023

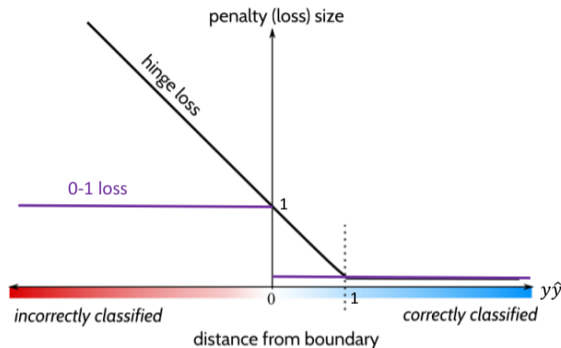
Hard-Margin SVM Recap



$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 \quad \text{s.t.} \quad y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \forall i$$

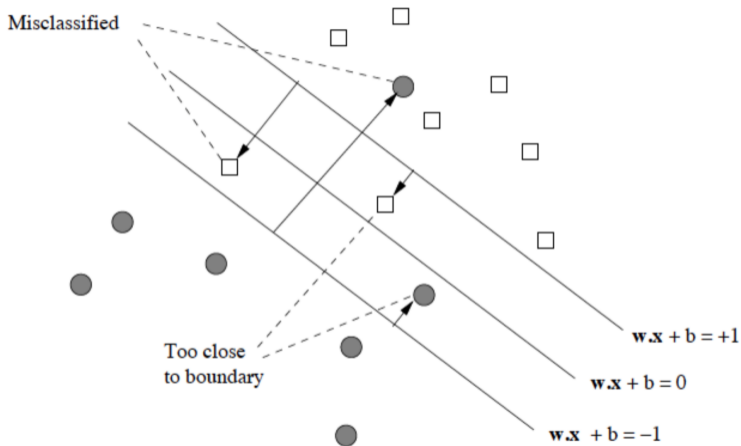
What if the data is not linearly separable? Penalize it in the loss!

The Hinge Loss

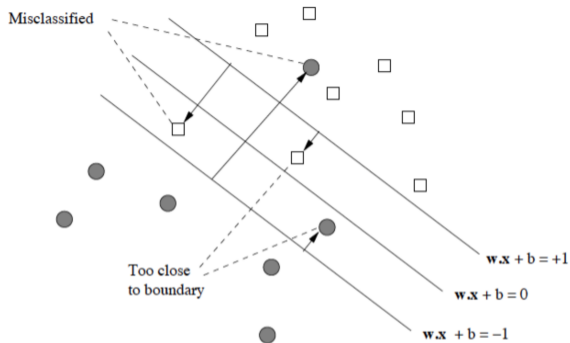


- Let $y \in \{-1, +1\}$ be the true label, $\hat{y} := \langle \mathbf{x}, \mathbf{w} \rangle + b$ be the predicted confidence
- We want to penalize $y(\langle \mathbf{x}, \mathbf{w} \rangle + b) < 1$, i.e., small or negative $y\hat{y}$
- Let's use $\ell_{\text{hinge}}(y\hat{y}) = (1 - y\hat{y})^+ = \begin{cases} 1 - y\hat{y}, & \text{if } y\hat{y} < 1; \\ 0, & \text{otherwise.} \end{cases}$

The points on which the classifier is penalized



Soft-Margin SVM



- Balancing between margin maximization and the [hinge](#) loss:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_i \underbrace{(1 - y_i \hat{y}_i)^+}_{\text{penalize error and small margin}}, \quad \text{s.t.} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w} \rangle + b$$

Comparison

Hard-Margin SVM:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \forall i$$

- Hard constraint: must respect

Soft-Margin SVM

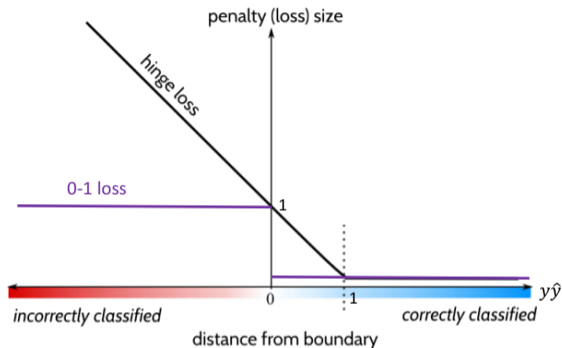
$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^n (1 - y_i \hat{y}_i)^+$$

$$\text{s.t. } \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$$

- Soft penalty: the more you deviate, the heavier the penalty

- $\frac{1}{2} \|\mathbf{w}\|_2^2$: margin maximization
- $(1 - y_i \hat{y}_i)^+$: error penalty
- C : hyper-parameter to control tradeoff

Why Hinge Loss?



Our goal: minimize over \mathbf{w}, b

$$\Pr(Y \neq \text{sign}(\hat{Y})) = \Pr(Y\hat{Y} \leq 0) = \mathbb{E} \underbrace{\mathbb{I}[Y\hat{Y} \leq 0]}_{\text{indicator function}} := \mathbb{E} \ell_{0-1}(Y\hat{Y}),$$

where $\hat{Y} = \langle \mathbf{X}, \mathbf{w} \rangle + b$, $Y = -1$ or $+1$

Why Hinge Loss? — Cont'

- Our goal: minimize $\mathbb{E}l_{0-1}(Y\hat{Y})$
- Even with linear predictors, minimizing the above 0-1 error is NP-hard
 - ▶ The loss is not continuous at 0
 - ▶ The gradient of the loss is 0 almost surely
- Therefore, we need to consider a surrogate loss, e.g., the hinge loss

A. L. Blum and R. L. Rivest (1992). "Training a 3-node neural network is NP-complete". *Neural Networks*, vol. 5, no. 1, pp. 117–127; S. Ben-David et al. (2003). "On the difficulty of approximately maximizing agreements". *Journal of Computer and System Sciences*, vol. 66, no. 3, pp. 496–514.

Why Hinge Loss? — Cont'

Definition: Bayes rule

Given an instance \mathbf{x} , the Bayes rule is given by $\eta(\mathbf{x}) := \operatorname{argmin}_{\hat{y} \in \mathbb{R}} \mathbb{E}[\ell_{0-1}(Y\hat{y}) | \mathbf{X} = \mathbf{x}]$.

Definition: Classification calibrated

We say a loss $\ell(y\hat{y})$ is classification-calibrated, iff for all \mathbf{x} ,

$$\hat{y}(\mathbf{x}) := \operatorname{argmin}_{\hat{y} \in \mathbb{R}} \mathbb{E}[\ell(Y\hat{y}) | \mathbf{X} = \mathbf{x}]$$

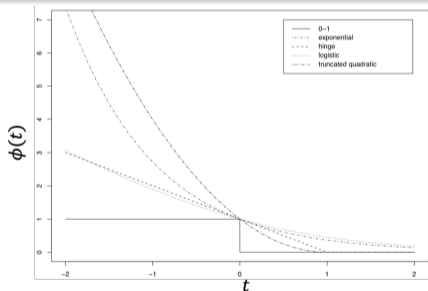
has the same sign as the Bayes rule $\eta(\mathbf{x}) = \operatorname{argmin}_{\hat{y} \in \mathbb{R}} \mathbb{E}[\ell_{0-1}(Y\hat{y}) | \mathbf{X} = \mathbf{x}]$.

- Note that $\eta(\mathbf{x})$ and $\hat{y}(\mathbf{x})$ provide the confidence, but not the prediction. Their sign operation provides the prediction.

Why Hinge Loss? — Cont'

Theorem: Characterization under convexity

Any **convex** loss ℓ is classification calibrated iff ℓ is differentiable at 0 and $\ell'(0) < 0$. Therefore, the classifier that minimizes the expected **hinge loss** minimizes the expected **zero-one** loss.



Lagrangian Dual

- Recall soft-margin SVM: $\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^n (1 - y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b))^+$
- Apply $C \cdot (t)^+ := \max\{Ct, 0\} = \max_{0 \leq \alpha \leq C} \alpha t$:

$$\min_{\mathbf{w}, b} \max_{0 \leq \alpha \leq C} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i \alpha_i [1 - y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b)]$$

- Swap min with max (strong duality holds due to convexity):

$$\boxed{\max_{0 \leq \alpha \leq C} \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i \alpha_i [1 - y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b)]}$$

- Solving the inner unconstrained problem by setting derivative to 0:

$$\frac{\partial}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = \mathbf{0}, \quad \frac{\partial}{\partial b} = - \sum_i \alpha_i y_i = 0$$

Lagrangian Dual — Cont'

- Plug in back to eliminate the inner problem (of \mathbf{w} and b):

$$\max_{0 \leq \alpha \leq C} \sum_i \alpha_i - \frac{1}{2} \left\| \sum_i \alpha_i y_i \mathbf{x}_i \right\|_2^2 \quad \text{s.t.} \quad \sum_i \alpha_i y_i = 0$$

- Changing max to min and expanding the norm:

$$\min_{0 \leq \alpha \leq C} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle} - \sum_i \alpha_i \quad \text{s.t.} \quad \sum_i \alpha_i y_i = 0$$

- What happens if $C \rightarrow \infty$? **Hard-margin SVM!**
- What happens if $C \rightarrow 0$? **A constant classifier!**

Comparison

Hard-margin SVM:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } \langle \mathbf{x}_i, \mathbf{w} \rangle + b \geq 1, \forall i$$

$$\min_{\alpha \geq 0} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\text{s.t. } \sum_i \alpha_i y_i = 0$$

Soft-margin SVM:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (1 - y_i \hat{y}_i)^+$$

$$\text{s.t. } \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$$

$$\min_{C \geq \alpha \geq 0} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\text{s.t. } \sum_i \alpha_i y_i = 0$$

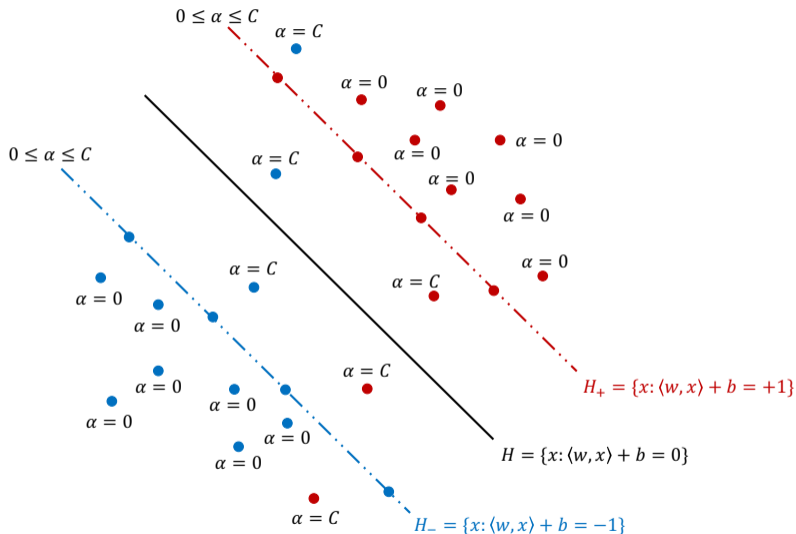
Complementarity Slackness

We have used the following relation to introduce the dual variables:

$$C \cdot (t)^+ \stackrel{C \geq 0}{=} \max\{Ct, 0\} = \max_{0 \leq \alpha \leq C} \alpha t =: \alpha^* t$$

- $t > 0 \implies \alpha^* = C, \quad \alpha^* = C \implies t \geq 0$
- $t < 0 \implies \alpha^* = 0, \quad \alpha^* = 0 \implies t \leq 0$
- Consider $t = 1 - y_i \hat{y}_i$:
 - ▶ $1 > y_i \hat{y}_i \implies \alpha_i^* = C, \quad \alpha_i^* = C \implies 1 \geq y_i \hat{y}_i$ (margin area or wrong area)
 - ▶ $1 < y_i \hat{y}_i \implies \alpha_i^* = 0, \quad \alpha_i^* = 0 \implies 1 \leq y_i \hat{y}_i$ (correctly classified with good confidence)
 - ▶ $1 = y_i \hat{y}_i \implies 0 \leq \alpha_i^* \leq C, \quad 0 < \alpha_i^* < C \implies 1 = y_i \hat{y}_i$ (correctly classified, on $H_{\pm 1}$)

Using α to determine the locations of points



Recovering \mathbf{w} and b from α

Primal Problem:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (1 - y_i \hat{y}_i)^+ \\ \text{s.t. } \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$$

Dual Problem:

$$\min_{C \geq \alpha \geq 0} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{s.t. } \sum_i \alpha_i y_i = 0$$

- Recovering $\mathbf{w}^* := \sum_i \alpha_i^* y_i \mathbf{x}_i$
- Normally, we set C large enough such that there is (at least) one data point sitting at one of $H_{\pm 1}$, i.e., $y_i \hat{y}_i = 1$
- This point can be used to recover b^* : $y(\langle \mathbf{x}, \mathbf{w}^* \rangle + b^*) = 1 \implies b^* = y - \langle \mathbf{x}, \mathbf{w}^* \rangle$
- Given a test data \mathbf{x} , prediction: $\hat{y} = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle + b^*)$

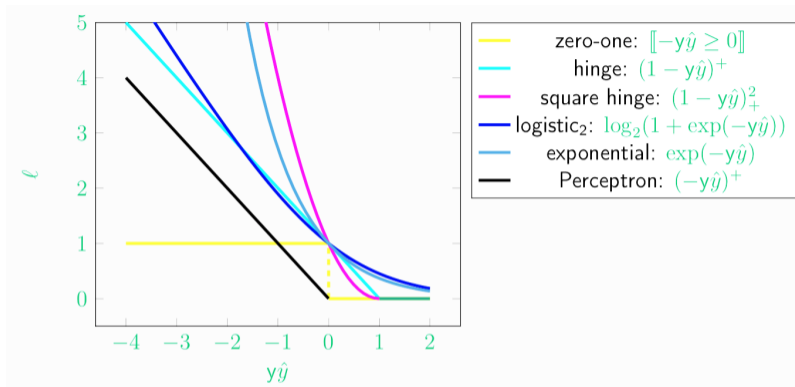
Training by Gradient Descent

$$\min_{\mathbf{w}, b} \frac{1}{2\lambda} \|\mathbf{w}\|_2^2 + \frac{1}{n} \sum_{i=1}^n \ell(y_i \hat{y}_i)$$

- Gradient descent:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \left[\frac{\mathbf{w}}{\lambda} + \frac{1}{n} \sum_{i=1}^n \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i \right]$$

Gradient of Hinge Loss



- $\ell'_{\text{hinge}}(t) = \begin{cases} -1, & t < 1 \\ 0, & t > 1 \\ [-1, 0], & t = 1 \end{cases}$ while we choose $\ell'_{\text{Perceptron}}(t) = \begin{cases} -1, & t \leq 0 \\ 0, & t > 0 \end{cases}$
- All other losses are differentiable

Questions

?

?

Answers

?