CS480/680: Introduction to Machine Learning

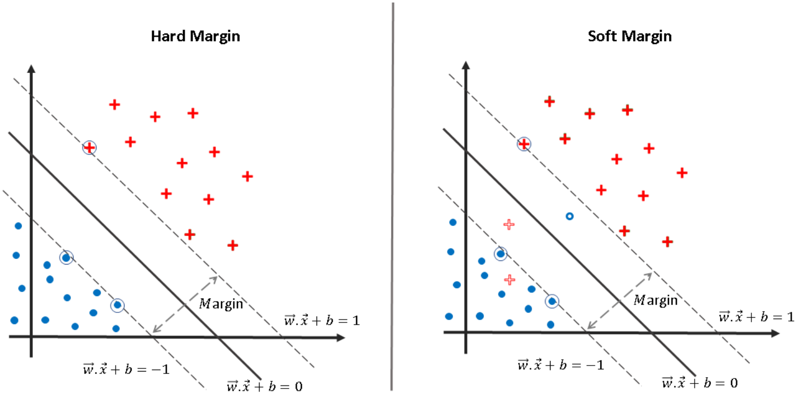
Lecture 6: Soft-Margin Support Vector Machines

Hongyang Zhang

UNIVERSITY OF WATERLOO

Jan 30, 2024

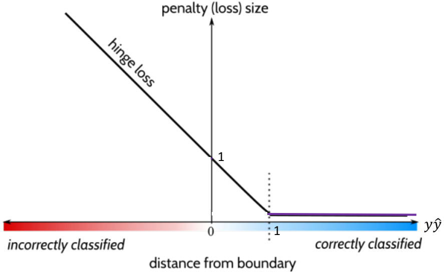
Hard-Margin SVM Recap



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

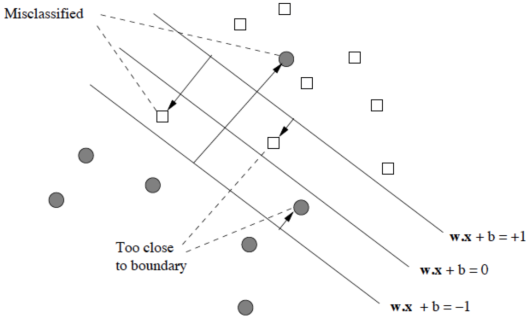
What if the data is not linearly separable? Penalize it in the loss!

The Hinge Loss



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

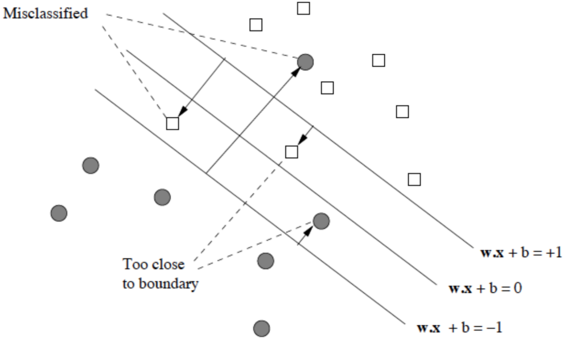
The points on which the classifier is penalized



Penalize $y\hat{y} < 1$ with two cases:

- $0 \leq y\hat{y} < 1$: points that are classified correctly but close to boundary
- $y\hat{y} < 0$: mis-classified points

Soft-Margin SVM



- Balancing between margin maximization and the hinge loss:

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

Comparison

Hard-Margin SVM:

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

- Hard constraint: must respect
- A special case of soft-margin SVM when $C = +\infty$

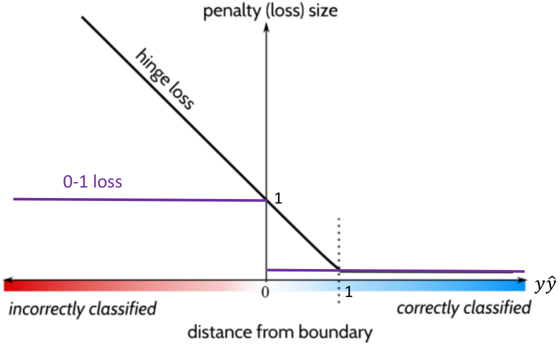
Soft-Margin SVM

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

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

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

Why Hinge Loss?



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

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

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

Why Hinge Loss? — Cont'

- Our goal: $\text{minimize}_{\hat{\mathsf{Y}}:\mathcal{X}\to\mathbb{R}}\ \mathbb{E}\ell_{0-1}(\mathsf{Y}\hat{\mathsf{Y}}) = \mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{Y}|\mathsf{X}}[\ell_{0-1}(\mathsf{Y}\hat{\mathsf{Y}})] \qquad (1)$
- 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

Definition: Bayes rule

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

$\hat{\mathsf{Y}} = \eta(\mathsf{X})$ minimizes (1), as it minimizes the inner expectation in (1).

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 $\eta(\mathbf{x}) := \text{argmin}_{\hat{y}\in\mathbb{R}} \mathbb{E}_{\mathsf{Y}|\mathsf{X}=\mathbf{x}}[\ell_{0-1}(\mathsf{Y}\hat{y})]$.

Definition: Classification calibrated

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

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

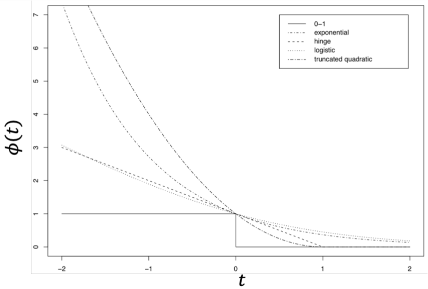
has the same sign as the Bayes rule $\eta(\mathbf{x}) = \text{argmin}_{\hat{y}\in\mathbb{R}} \mathbb{E}_{\mathsf{Y}|\mathsf{X}=\mathbf{x}}[\ell_{0-1}(\mathsf{Y}\hat{y})]$.

- Note that $\eta(\mathbf{x})$ and $\hat{\mathsf{y}}(\mathbf{x})$ provide the score, 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 $\ell'(0) < 0$. So, the classifier that minimizes the expected hinge loss minimizes the expected 0-1 loss.



$\ell_{\text{perceptron}}(y\hat{y}) = -\min\{y\hat{y}, 0\}$ is NOT classification-calibrated; non-differentiable at 0.

P. L. Bartlett et al. (2006). "Convexity, Classification, and Risk Bounds". Journal of the American Statistical Association, vol. 101, no. 473, pp. 138–156.

Lagrangian Dual

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

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

- Swap $\min$ with $\max$:

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

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

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

Lagrangian Dual — Cont'

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

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

- Changing $\max$ to $\min$ and expanding the norm:

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

- What happens if $C \to \infty$? Hard-margin SVM! (Soft $\to$ Hard Constraint)
- What happens if $C \to 0$? A constant classifier! (Dual: $\boldsymbol{\alpha}^* = \mathbf{0}$; Primal: $\mathbf{w}^* = \mathbf{0}$)

Comparison

Hard-margin SVM:

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

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

Soft-margin SVM:

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

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

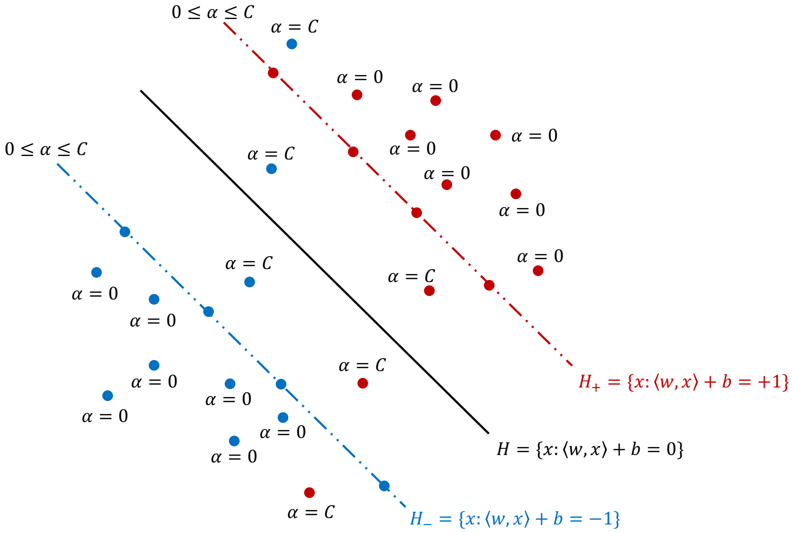
Complementary Slackness

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

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

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

Using the locations of points to determine α



Recovering $\mathbf{w}$ and b from α

Primal Problem:

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

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

Dual Problem:

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

$$\text{s.t.} \quad \sum_i \alpha_i \mathsf{y}_i = 0$$

- Recovering $\mathbf{w}^* := \sum_i \alpha_i^* \mathsf{y}_i \mathbf{x}_i$
- Normally, C is large enough such that there is (at least) one data point sitting at one of $H_{\pm 1}$, i.e., $\mathsf{y}\hat{y} = 1$; Otherwise, your choice of C might be too small and allow too many mistakes.
- This point can be used to recover b^*: $\mathsf{y}(\langle \mathbf{x}, \mathbf{w}^*\rangle + b^*) = 1 \implies b^* = \mathsf{y} - \langle \mathbf{x}, \mathbf{w}^*\rangle$
- Given a test data $\mathbf{x}$, prediction: $\hat{y} = \text{sign}(\langle \mathbf{w}^*, \mathbf{x}\rangle + b^*)$

Training by Gradient Descent

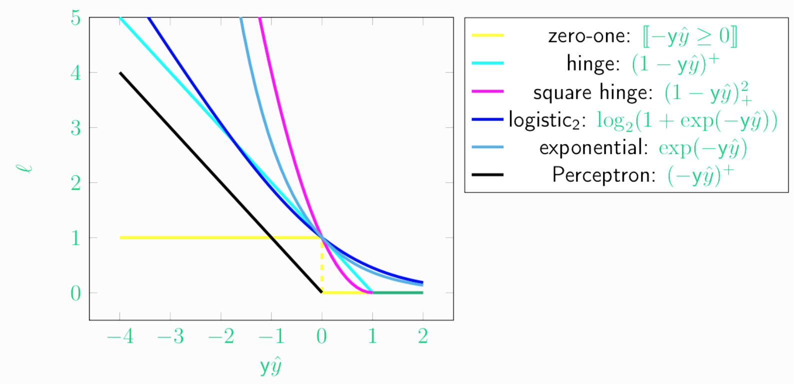
$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\ell(\mathsf{y}_i\hat{y}_i), \text{ where } \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b$$

- Gradient descent with step size η:

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

$$\boxed{b \leftarrow b - \eta\left[C\sum_{i=1}^{n}\ell'(\mathsf{y}_i\hat{y}_i)\mathsf{y}_i\right]}$$

Gradient of Hinge Loss



- $\ell'_{\text{hinge}}(t) = \begin{cases} -1, & t \leq 1 \\ 0, & t > 1 \end{cases}$ 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
?