CS480/680: Introduction to Machine Learning

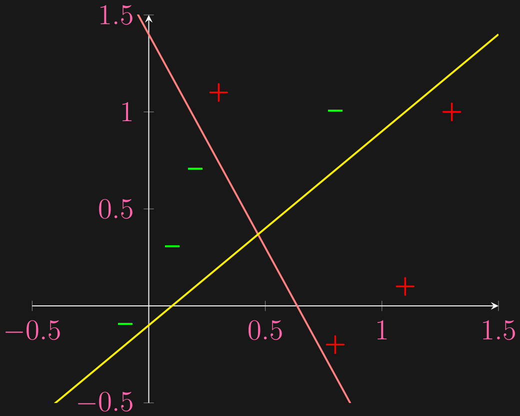
Lec 05: Soft-margin Support Vector Machines

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

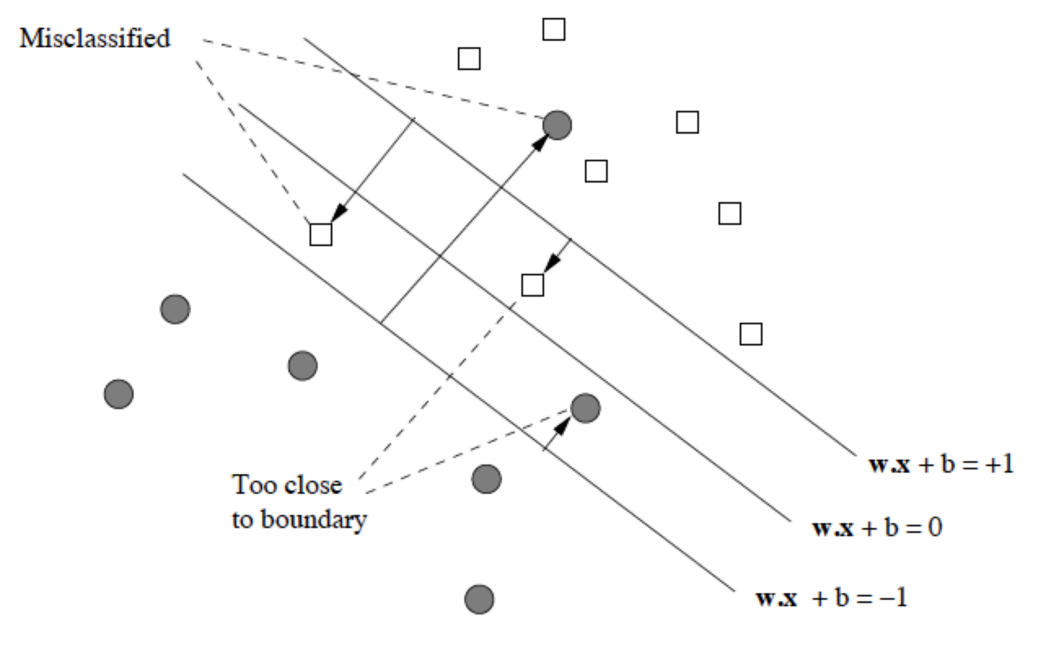
May 22, 2024

Beyond Separability



- Balancing between margin maximization and the soft-margin loss:

$$\min_{\mathbf{w},b} \ \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \textstyle\sum_i (1 - \mathsf{y}_i \hat{y}_i)^+, \quad \text{s.t.} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w} \rangle + b$$


Misclassified
Too close
to boundary
w.x + b = +1
w.x + b = 0
w.x + b = −1

Soft-margin SVM

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

- Hard constraint: must respect; "live or die"

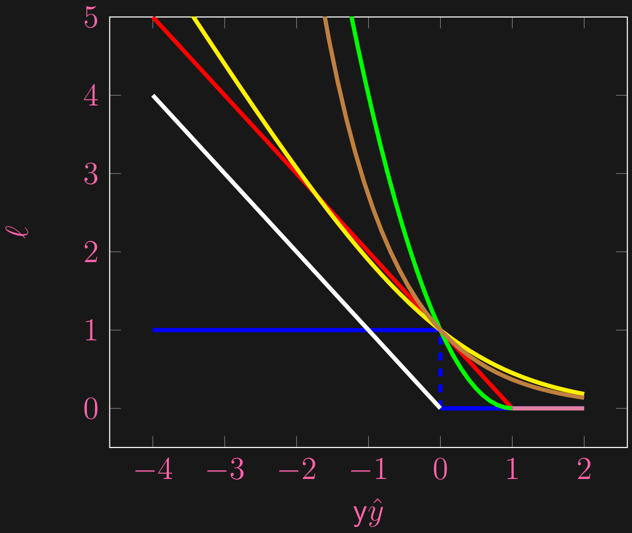
$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} (1 - \mathsf{y}_i \hat{y}_i)^+$$
$$\text{s.t.} \quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b, \forall i$$

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

- $\frac{1}{2}\|\mathbf{w}\|_2^2$: margin maximization
- $(1 - \mathsf{y}_i \hat{y}_i)^+$: i-th training error, 0 if $\mathsf{y}_i \hat{y}_i \geq 1$ and $1 - \mathsf{y}_i \hat{y}_i$ (grow linearly) otherwise
- C: hyper-parameter to control tradeoff

C. Cortes and V. Vapnik. "Support-vector networks". *Machine Learning*, vol. 20, no. 3 (1995), pp. 273–297.

The Hinge Loss



Zero-one Loss and Generalization Error

$$\Pr(\hat{\mathsf{Y}} \neq \mathsf{Y}) = \mathbb{E}[\![-\mathsf{Y} f(\mathsf{X}) \geq 0]\!], \quad \text{where} \quad \hat{\mathsf{Y}} = \mathrm{sign}(f(\mathsf{X}))$$

- $f : \mathcal{X} \to \mathbb{R}$ is our real-valued predictor, e.g., $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$
- Training error after sampling

$$\frac{1}{n} \sum_{i=1}^{n} [\![-\mathsf{Y}_i f(\mathsf{X}_i) \geq 0]\!]$$

- Even with linear predictors, minimizing the above training error is NP-hard

---

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

Classification Calibration

- Want to minimize the 0-1 loss, but often end up with minimizing something else
- Is this sensible?

Definition: Bayes rule

Let $\eta(\mathbf{x}) := \Pr(\mathsf{Y} = 1 | \mathsf{X} = \mathbf{x})$. The optimal Bayes classifier is $\mathrm{sign}(2\eta(\mathbf{x}) - 1)$.

Definition: Classification calibrated

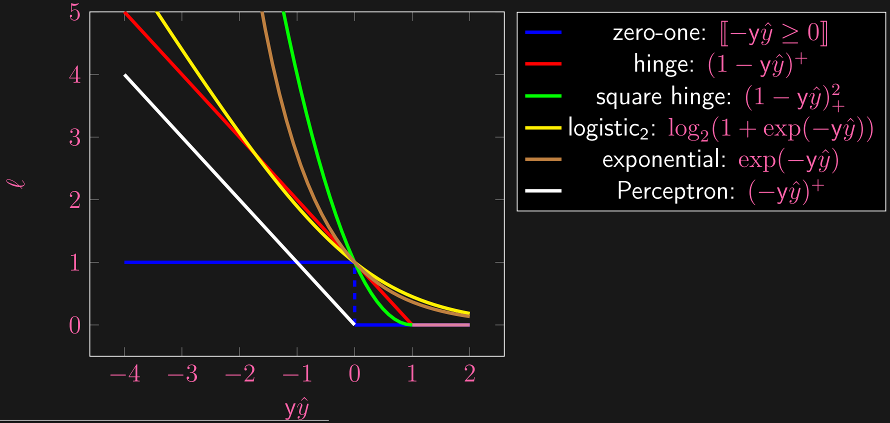
We say a (margin) loss $\ell(\mathsf{y}\hat{y})$ is classification calibrated iff

$$\hat{\mathsf{y}} = \hat{\mathsf{y}}(\mathbf{x}) := \operatorname*{argmin}_{y \in \mathbb{R}} \; \eta(\mathbf{x})\ell(y) + [1 - \eta(\mathbf{x})]\ell(-y) \quad \backslash\backslash = \; \mathbb{E}[\ell(y\mathsf{Y}) | \mathsf{X} = \mathbf{x}]$$

has the same sign as the Bayes rule.

Theorem: Characterization under convexity

Any convex (margin) loss ℓ is classification calibrated iff ℓ is differentiable at 0 and $\ell'(0) < 0$.



P. L. Bartlett et al. "Convexity, classification, and risk bounds". *Journal of the American Statistical Association*, vol. 101, no. 473 (2006), pp. 138–156.

A Simpler Way to Derive Lagrangian Dual

$$C \cdot (t)^+ := \max\{Ct, 0\} = \max_{0 \le \alpha \le C} \alpha t$$

- Apply above to each term:

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

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

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

- 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):

$$\max_{0 \leq \boldsymbol{\alpha} \leq C} \sum_i \alpha_i - \tfrac{1}{2} \|\textstyle\sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i\|_2^2$$

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

$$\min_{0 \leq \boldsymbol{\alpha} \leq C} \tfrac{1}{2} \textstyle\sum_i \sum_j \alpha_i \alpha_j \mathsf{y}_i \mathsf{y}_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle} - \sum_i \alpha_i$$

- What happens if $C \to \infty$?
- What happens if $C \to 0$?

Comparison

$$\min_{\mathbf{w},b} \tfrac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \iota_{1-\mathsf{y}_i\hat{y}_i\le 0}$$
$$\text{s.t.}\quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b, \forall i$$

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

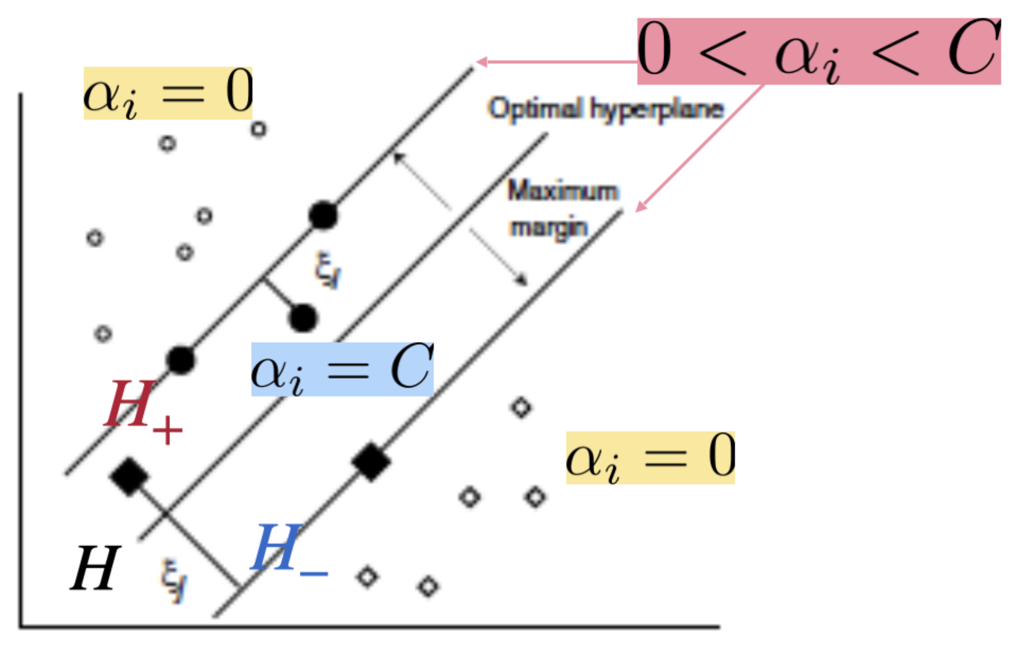
$$\min_{\mathbf{w},b} \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} (1-\mathsf{y}_i\hat{y}_i)^+$$
$$\text{s.t.}\quad \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b, \forall i$$

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

Complementarity Slackness

$$C \cdot (t)^+ := \max\{Ct, 0\} = \max_{0 \le \alpha \le C} \alpha t$$

- $t > 0 \implies \alpha = C$ and $\alpha = C \implies t \geq 0$
- $t < 0 \implies \alpha = 0$ and $\alpha = 0 \implies t \leq 0$
- Apply to each term in soft-margin SVM:
  - $1 > \mathsf{y}_i \hat{y}_i \implies \alpha_i = C$ and $\alpha_i = C \implies 1 \geq \mathsf{y}_i \hat{y}_i$ (wrong side of $H_{\pm 1}$, correct/incorrect)
  - $1 < \mathsf{y}_i \hat{y}_i \implies \alpha_i = 0$ and $\alpha_i = 0 \implies 1 \leq \mathsf{y}_i \hat{y}_i$ (correctly classified, on/beyond $H_{\pm 1}$)
  - $1 = \mathsf{y}_i \hat{y}_i \implies 0 \geq \alpha_i \geq C$ and $0 < \alpha_i < C \implies 1 = \mathsf{y}_i \hat{y}_i$ (correctly classified, on $H_{\pm 1}$)


$0 < \alpha_i < C$
$\alpha_i = 0$
Optimal hyperplane
Maximum
margin
ξ_i
$\alpha_i = C$
H_+
$\alpha_i = 0$
H
ξ_i
H_-

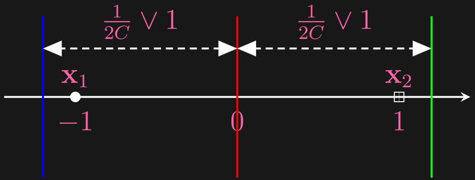
A Simple Example

$$\min_{w,b} \ \tfrac{1}{2}w^2 + C(1-w+b)^+ + C(1-w-b)^+$$

$$\min_{C \geq \boldsymbol{\alpha} \geq 0} \ \tfrac{1}{2}(\alpha_1+\alpha_2)^2 - \alpha_1 - \alpha_2$$

$$\text{s.t.} \quad \alpha_1 - \alpha_2 = 0$$

$$\alpha_1 = \alpha_2 = \tfrac{1}{2} \wedge C, \quad w = 1 \wedge (2C), \quad |b| \leq 1 - w$$



Recovering b

- W.l.o.g., there is always (at least) one data point sitting at one of $H_{\pm 1}$
  - suppose not, move the hyperplanes to the left / right until touching a data point
  - one of the directions must not increase the soft-margin loss
- This point can be used to recover b: $\mathrm{y}(\langle \mathbf{x}, \mathbf{w} \rangle + b) = 1$
  - can average if multiple points are (close to be) on $H_{\pm 1}$

A Word About Stochastic Gradient

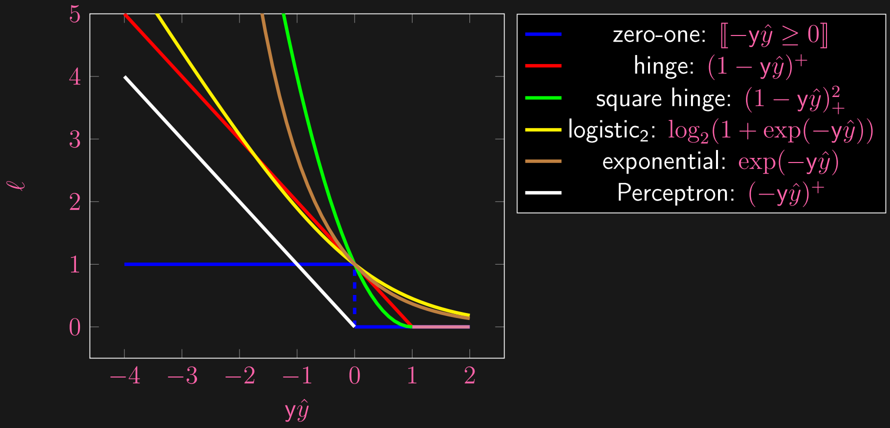
$$\min_{\mathbf{w},b} \frac{1}{2\lambda}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(\mathsf{y}_i \hat{y}_i)$$

- Gradient descent costs $O(nd)$:

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

- A random sample suffices:

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



- $\ell'_{\text{hinge}}(t) = \begin{cases} -1, & t < 1 \\ 0, & t > 1 \\ [-1, 0], & t = 1 \end{cases}$ while we *choose* $\ell'_{\text{Perceptron}}(t) = \begin{cases} -1, & t \leq 0 \\ 0, & t > 0 \end{cases}$
- What about the zero-one loss? Other losses?

Multi-class

$$\forall i, \quad \hat{\mathbf{y}}_i = W\mathbf{x}_i + \mathbf{b} \in \mathbb{R}^c,$$

$$\min_{W,\mathbf{b}} \frac{1}{2}\|W\|_\mathsf{F}^2$$

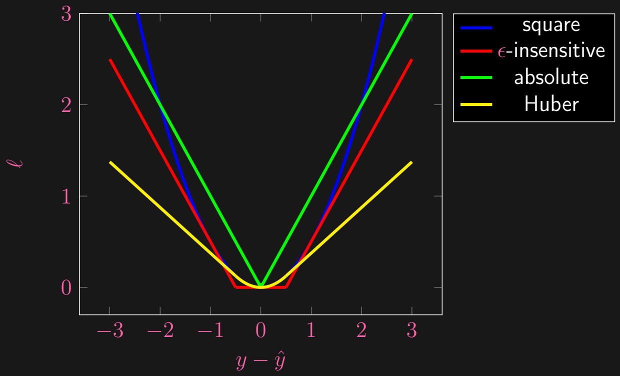
$$\text{s.t.} \quad \hat{y}_{\mathsf{y}_i,i} \geq [\![k \neq \mathsf{y}_i]\!] + \hat{y}_{k,i}, \quad \forall i, \forall k = 1, \ldots, c$$

$$\min_{W,\mathbf{b}} \frac{1}{2}\|W\|_\mathsf{F}^2 + C\sum_{i=1}^{n} \max_{k=1,\ldots,c} \{[\![k \neq \mathsf{y}_i]\!] + \hat{y}_{k,i} - \hat{y}_{\mathsf{y}_i,i}\}$$

K. Crammer and Y. Singer. "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines". *Journal of Machine Learning Research*, vol. 2 (2001), pp. 265–292.

Regression

$$\min_{W} \frac{1}{2}\|W\|_{\mathsf{F}}^2 + C\sum_{i=1}^{n}(\|\mathbf{y} - \hat{\mathbf{y}}_i\| - \epsilon)^+$$



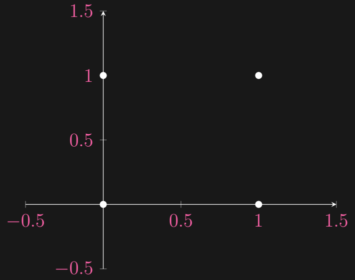
H. Drucker et al. "Support Vector Regression Machines". In: *Advances in Neural Information Processing Systems 9*. 1996.

Clustering

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

$$\text{s.t.}\quad \text{label balance,}\quad \text{e.g.,}\quad |\langle \mathbf{1}, \mathbf{y}\rangle| \le t$$

- No longer a convex program due to the bilinear term $y_i\hat{y}_i$



L. Xu et al. "Maximum Margin Clustering". In: *Advances in Neural Information Processing Systems 17*. 2004.