CS480/680: Introduction to Machine Learning

Lecture 5: Hard-Margin Support Vector Machines

Hongyang Zhang

UNIVERSITY OF WATERLOO

Jan 25, 2024

Perceptron Revisited

- $\mathcal{Y} = \{-1, +1\}$; no padding trick today
- Assuming linearly separable
  - exist $\mathbf{w}$ and b such that

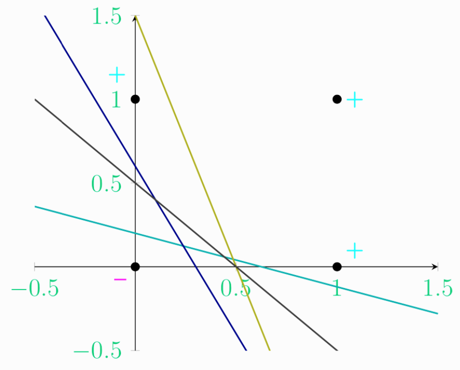
$$\forall i,\ \mathsf{y}_i \hat{y}_i > 0,\ \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b$$

- Perceptron: find any $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that:

$$\min_{\mathbf{w},b} 0, \ \text{s.t.} \quad \mathsf{y}_i \hat{y}_i > 0, \forall i$$

$$\overset{(\mathbf{w},b)\to(c\mathbf{w},cb)}{\Longleftrightarrow} \min_{\mathbf{w},b} 0, \ \text{s.t.} \quad \mathsf{y}_i \hat{y}_i \geq 1, \forall i$$

- The larger the margin is, the faster the Perceptron will converge

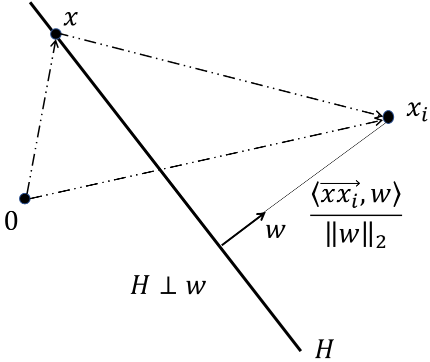


Hard-Margin SVM: Let's maximize margin (assume training data are linearly separable)

Margin: Distance from a Point to a Hyperplane

Let $H := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\}$. What is the distance from a point $\mathbf{x}_i$ to H?

- $\mathbf{w}$ is orthogonal to H (see Lecture 2)
- Let $\mathbf{x}$ be any vector in H. The distance = The length of the projection of $\mathbf{x}_i - \mathbf{x}$ onto $\mathbf{w}$
- Distance$(\mathbf{x}_i, H) = \frac{|\langle \mathbf{x}_i - \mathbf{x}, \mathbf{w} \rangle|}{\|\mathbf{w}\|_2} = \frac{|\langle \mathbf{x}_i, \mathbf{w} \rangle - \langle \mathbf{x}, \mathbf{w} \rangle|}{\|\mathbf{w}\|_2} \overset{\mathbf{x} \in H}{=} \frac{|\langle \mathbf{x}_i, \mathbf{w} \rangle + b|}{\|\mathbf{w}\|_2} \overset{\mathsf{y}_i \hat{y}_i > 0}{=} \frac{\mathsf{y}_i \hat{y}_i}{\|\mathbf{w}\|_2}$



Margin

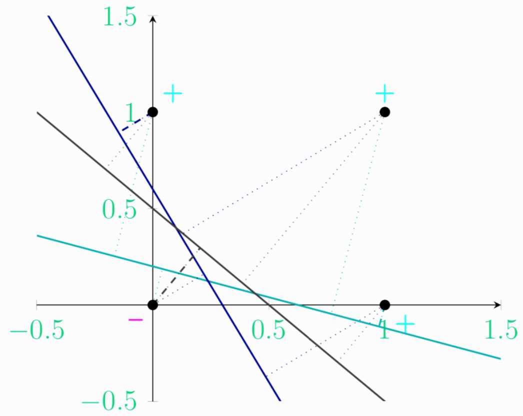
Define the smallest distance to a separating hyperplane H among all separable (training) data as the margin:

$$\min_i \frac{\mathsf{y}_i \hat{y}_i}{\|\mathbf{w}\|_2} = \min_i \frac{|\langle \mathbf{x}_i, \mathbf{w} \rangle + b|}{\|\mathbf{w}\|_2}$$

- We have assumed H separates the data points ($\mathsf{y}_i \hat{y}_i > 0$ for all i)
- Margin w.r.t. a separating hyperplane is the minimum distance to every point
- Our goal is to maximize the margin among all hyperplanes:

$$\max_{\mathbf{w},b} \min_i \frac{\mathsf{y}_i \hat{y}_i}{\|\mathbf{w}\|_2}, \quad \text{s.t.} \quad \mathsf{y}_i \hat{y}_i > 0 \text{ for all } i.$$

Margin Maximization



$$\max_{\mathbf{w},b:\forall i,\mathrm{y}_i\hat{y}_i>0} \min_{i=1,\ldots,n} \frac{\mathrm{y}_i\hat{y}_i}{\|\mathbf{w}\|_2}, \quad \text{where} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w}\rangle + b$$

Transforming to the Standard Form

$$\max_{\mathbf{w},b:\forall i,\mathsf{y}_i\hat{y}_i>0} \min_i \frac{\mathsf{y}_i\hat{y}_i}{\|\mathbf{w}\|_2}, \quad \text{where} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w}\rangle + b$$

- Both numerator and denominator are homogeneous in $(\mathbf{w}, b)$
  - Meaning that $(\mathbf{w}, b)$ and $(c\mathbf{w}, cb)$ will have the same loss for $c > 0$
  - Varying $c(> 0)$ will not break the condition $\mathsf{y}_i\hat{y}_i > 0$ (same decision boundary)
  - Varying $c(> 0)$ can change the numerator arbitrarily
- Can fix the numerator arbitrarily, say to 1

$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2} \quad \text{s.t.} \quad \min_i \ \mathsf{y}_i\hat{y}_i = 1$$

- Max $\to$ min, and squaring for convenience:

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

Comparison to Perceptron

Hard-Margin SVM:

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

- Quadratic programming
- Fewer solutions
- Margin maximizing

Perceptron:

$$\min_{\mathbf{w},b} 0$$
$$\text{s.t.} \quad \mathsf{y}_i \hat{y}_i \geq 1, \forall i$$

- Linear programming
- Infinitely many solutions
- Convergence rate depends on maximum margin

Support Vectors

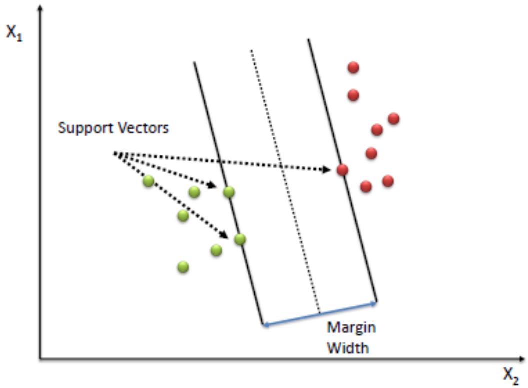
$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|_2^2 \ \text{ s.t. } \ \hat{y}_i \geq 1, \forall i : \mathrm{y}_i = +1$$

$$\hat{y}_i \leq -1, \forall i : \mathrm{y}_i = -1$$

- Three parallel hyperplanes:

$$H := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\}$$
$$H_+ := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle + b = 1\}$$
$$H_- := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle + b = -1\}$$

- Support vectors: points lie on the supporting hyperplanes
  - Usually only a few, but decisive
  - decisive because these points reach the boundary of constraint



Explanation from Dual Perspective

Lagrangian Dual

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

- Introducing Lagrangian multipliers, a.k.a. dual variables $\boldsymbol{\alpha} \in \mathbb{R}^n$:

$$\min_{\mathbf{w},b} \ \max_{\boldsymbol{\alpha}\geq \mathbf{0}} \ \tfrac{1}{2}\|\mathbf{w}\|_2^2 - \sum_i \alpha_i[\mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1]$$
$$= \min_{\mathbf{w},b} \begin{cases} +\infty, & \text{if } \exists i, \mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) < 1 \text{ (set } \alpha_i \text{ as } +\infty) \\ \tfrac{1}{2}\|\mathbf{w}\|_2^2, & \text{if } \forall i, \mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 \text{ (set } \alpha_i \text{ as } 0 \text{ for all } i) \end{cases}$$
$$= \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad \mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1, \forall i$$

- P.S.: transfer a constrained optimization to a unconstrained one on $\mathbf{w}, b$

Lagrangian Dual — Cont'

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

- We have proved that it is equivalent to:

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

- Swapping $\min$ with $\max$:

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

Lagrangian Dual — Cont'

- Solving inner unconstrained problem by setting derivative to 0:

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

- Plug $\mathbf{w}$ in the loss to eliminate the inner problem:

$$\begin{aligned} \text{Loss}(\boldsymbol{\alpha}) &= \min_{\mathbf{w},b} \ \tfrac{1}{2}\|\mathbf{w}\|_2^2 - \sum_i \alpha_i[\mathsf{y}_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) - 1] \\ &= \frac{1}{2}\|\sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i\|_2^2 - \langle \sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i, \sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i \rangle - b \sum_i \alpha_i \mathsf{y}_i + \sum_i \alpha_i \end{aligned}$$

That is, $\max_{\alpha \geq \mathbf{0}} \sum_i \alpha_i - \frac{1}{2}\|\sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i\|_2^2 \qquad \text{s.t.} \quad \sum_i \alpha_i \mathsf{y}_i = 0$

- Change to minimization and expand the norm:

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

Primal vs. Dual

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

- primal variables: $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$
- primal inequalities: n
- primal equalities: 0

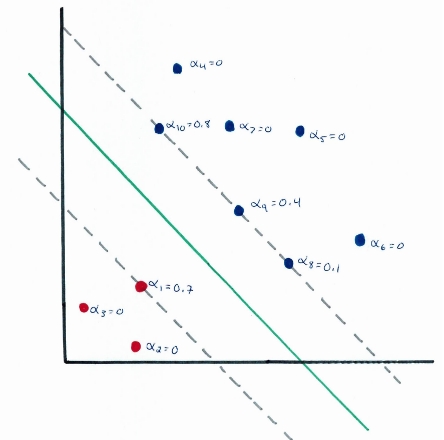
$$\min_{\alpha \geq \mathbf{0}} \ -\sum_i \alpha_i + \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j \mathsf{y}_i \mathsf{y}_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}$$
$$\text{s.t.} \quad \sum_i \alpha_i \mathsf{y}_i = 0$$

- dual variables: $\boldsymbol{\alpha} \in \mathbb{R}^n$
- each α_i corresponds to a data sample
- dual inequalities: n
- dual equalities: 1

Support Vectors and Dual Variables

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i.$$

The data with $\alpha_i = 0$ does not contribute to the decision boundary (non-support vector).



The reason why the dual form is of interest

Sometimes, data might not be linearly separable

Better idea: use a non-linear mapping ϕ to map the data; but ϕ is unknown?

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

$\Downarrow$ an unknown non-linear mapping ϕ

$$\min_{\alpha \geq \mathbf{0}} -\sum_i \alpha_i + \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j \mathsf{y}_i \mathsf{y}_j \underbrace{\boxed{\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle}}_{\text{has a closed form w.r.t. } \mathbf{x}_i \text{ and } \mathbf{x}_j} \quad \text{s.t.} \quad \sum_i \alpha_i \mathsf{y}_i = 0$$

- This is also known as kernel; we don't need to know ϕ explicitly
  - Example: $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^{10}$ (no ϕ appears in the RHS)
  - Only inner product between data has this nice property
- We will see more details in Lecture 7


Questions
?
?
Answers
?