CS480/680: Introduction to Machine Learning

Lec 04: Support Vector Machines

Yaoliang Yu

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

May 21, 2024

Perceptron Revisited

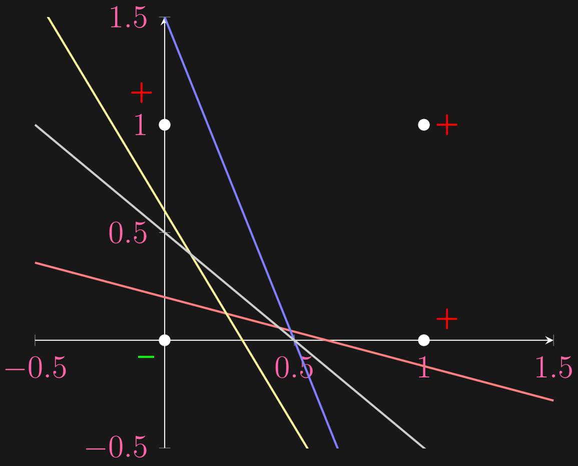
- Two classes: $\mathsf{y} \in \{\pm 1\}$
- Assuming linearly separable
  - exist $\mathbf{w}$ and b such that

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

- Perceptron: find any $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that for all i, $\mathsf{y}_i \hat{y}_i > 0$, i.e., the feasibility problem:

$$\min_{\mathbf{w}, b} 0$$

$$\text{s.t. } \mathsf{y}_i \hat{y}_i > 0, \forall i$$



Euclidean 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 $\mathbf{z}$ to H?

$$\min_{\mathbf{x} \in H} \|\mathbf{x} - \mathbf{z}\|_{\circ}$$

- Rotation does not change Euclidean distance
- Consider the axis defined by $\bar{\mathbf{w}} := \frac{\mathbf{w}}{\|\mathbf{w}\|_2}$ and the hyperplane $\langle \mathbf{x}, \bar{\mathbf{w}} \rangle + \bar{b} = 0$
- Projected onto $\bar{\mathbf{w}}$, $\mathbf{z}$ becomes $\langle \bar{\mathbf{w}}, \mathbf{z} \rangle \cdot \bar{\mathbf{w}}$
- Distance becomes $| -\bar{b} - \langle \bar{\mathbf{w}}, \mathbf{z} \rangle | = \frac{|\langle \mathbf{z}, \mathbf{w} \rangle + b|}{\|\mathbf{w}\|_2}$

Any Distance from a Point to a Hyperplane

$$\|\mathbf{x}-\mathbf{z}\|_2 \geq |\langle \bar{\mathbf{w}}, \mathbf{x}-\mathbf{z}\rangle| = \frac{|\langle \mathbf{z}, \mathbf{w}\rangle + b|}{\|\mathbf{w}\|_2}$$

- Equality is attained at $\mathbf{x}_\star - \mathbf{z} \propto \bar{\mathbf{w}}$, i.e., $\mathbf{x}_\star = \mathbf{z} - (\bar{b} + \langle \mathbf{z}, \bar{\mathbf{w}}\rangle)\bar{\mathbf{w}}$
- Indeed one can verify $\mathbf{x}_\star \in H$, i.e., $\langle \mathbf{x}_\star, \mathbf{w}\rangle + b = 0$
- Immediately extends to any distance defined by a norm:

$$\|\mathbf{x}-\mathbf{z}\|_\circ \geq \frac{|\langle \mathbf{w}, \mathbf{x}-\mathbf{z}\rangle|}{\|\mathbf{w}\|} = \frac{|\langle \mathbf{z}, \mathbf{w}\rangle + b|}{\|\mathbf{w}\|}$$

- Equality is attained at $\mathbf{x}_\star - \mathbf{z} \propto \bar{\mathbf{w}} \in \partial\|\mathbf{w}\|$, i.e., $\mathbf{x}_\star = \mathbf{z} - \frac{b+\langle \mathbf{z}, \mathbf{w}\rangle}{\|\mathbf{w}\|}\bar{\mathbf{w}}$

Margin

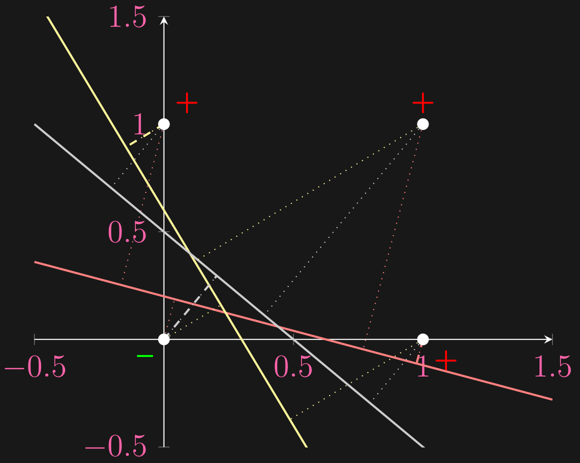
The margin of a (separable) dataset $\mathcal{D} := \{\mathbf{x}_i, \mathsf{y}_i\}_{i=1}^n$ w.r.t. a separating hyperplane $H := \{\mathbf{x} : \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\}$ is:

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

- H separates the data points
- Margin w.r.t. a separating hyperplane is the minimum distance to every point
- Margin of a (separable) dataset is the maximum among all hyperplanes:

$$\gamma_2(\mathcal{D}) := \max_{\mathbf{w}, b} \min_i \frac{\mathsf{y}_i \hat{y}_i}{\|\mathbf{w}\|_2}$$

Margin Maximization



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

Transforming to the Standard Form

$$\max_{\mathbf{w},b} \min_i \frac{\mathsf{y}_i \hat{y}_i}{\|\mathbf{w}\|_2}$$

- Both numerator and denominator are homogeneous in $(\mathbf{w}, b)$
- 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 $\rightarrow$ min, and squaring for convenience:

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

Comparison to Perceptron

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

- Quadratic programming
- Unique solution
- Margin maximizing

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

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

Support Vectors

- Support vectors: points lie on the supporting hyperplanes
  - $\mathcal{D}_+ := \{\mathbf{x}_i : \mathsf{y}_i = +1,\ \mathsf{y}_i \hat{y}_i = 1\}$
  - $\mathcal{D}_- := \{\mathbf{x}_i : \mathsf{y}_i = -1,\ \mathsf{y}_i \hat{y}_i = 1\}$

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

$$\text{s.t.} \quad \mathsf{y}_i \hat{y}_i \geq 1, \forall i$$

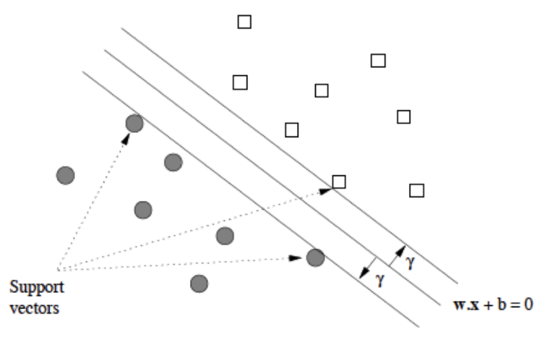
- Usually only a handful, but decisive

- Three parallel hyperplanes:

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

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

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



Lagrangian Dual

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

- Introducing Lagrangian multipliers, a.k.a. dual variables $\boldsymbol{\alpha}$:

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

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

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

- No more constraint on $\mathbf{w}, b$

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 in back to eliminate the inner problem:

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

- Change to minimization and expand the norm:

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

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

Primal vs. Dual

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

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

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

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

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

- dual variables: $\boldsymbol{\alpha} \in \mathbb{R}^n$
- dual inequalities: n
- dual equalities: 1 (coming from $\frac{\partial}{\partial b} = 0$)

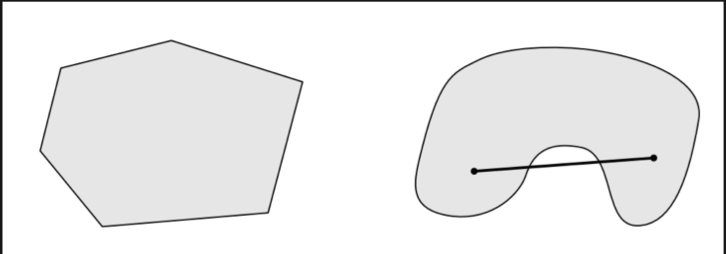
$$\boxed{\mathbf{w} = \sum_i \alpha_i \mathsf{y}_i \mathbf{x}_i}$$

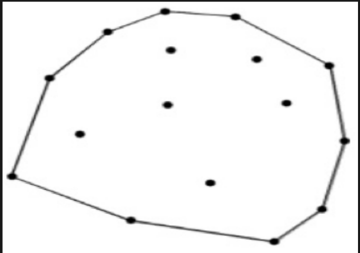
Convex Sets and Convex Hull

Convex set. A point set $C \subseteq \mathbb{R}^d$ is convex if any line segment $[\mathbf{w}, \mathbf{z}]$ connecting two points $\mathbf{w}, \mathbf{z} \in C$ lies entirely in C.

Convex hull. The smallest convex set containing C:

$$\operatorname{conv} C := \left\{ \sum_i \alpha_i \mathbf{w}_i : \mathbf{w}_i \in C,\ \alpha_i \geq 0,\ \sum_i \alpha_i = 1 \right\}$$





Separating Scale from Direction

Setting $\boldsymbol{\alpha} = r \cdot \bar{\boldsymbol{\alpha}}$ for some $r > 0$ and $\bar{\boldsymbol{\alpha}} \in 2\Delta := \{\beta \geq 0 : \sum_i \beta_i = 2\}$:

$$\min_{r \geq 0} \min_{\bar{\boldsymbol{\alpha}} \in 2\Delta} \; -2r + \frac{r^2}{2} \|\sum_i \bar{\alpha}_i \mathsf{y}_i \mathbf{x}_i\|_2^2$$
$$\text{s.t.} \quad \sum_i \bar{\alpha}_i \mathsf{y}_i = 0$$

- Solving r by setting derivative to 0:

$$r = \frac{2}{\|\sum_i \bar{\alpha}_i \mathsf{y}_i \mathbf{x}_i\|_2^2}$$

- Plug in to eliminate r:

$$\min_{\bar{\boldsymbol{\alpha}} \in 2\Delta, \, \sum_i \bar{\alpha}_i \mathsf{y}_i = 0} -\frac{2}{\|\sum_i \bar{\alpha}_i \mathsf{y}_i \mathbf{x}_i\|_2^2}$$

Split

$$\min_{\bar{\boldsymbol{\alpha}}\in 2\Delta,\ \sum_i \bar{\alpha}_i \mathsf{y}_i = 0} \|\sum_i \bar{\alpha}_i \mathsf{y}_i \mathbf{x}_i\|_2^2$$

- Positive set $P := \{i : \mathsf{y}_i = +1\}$
- Negative set $N := \{i : \mathsf{y}_i = -1\}$
- Split $\bar{\alpha}$ into $[\boldsymbol{\mu} : i \in P$ and $\boldsymbol{\nu} : i \in N]$
- $\bar{\boldsymbol{\alpha}} \in 2\Delta,\ \sum_i \bar{\alpha}_i \mathsf{y}_i = 0 \iff \boldsymbol{\mu} \in \Delta_+, \boldsymbol{\nu} \in \Delta_-$

$$\begin{aligned}\min_{\bar{\boldsymbol{\alpha}}\in 2\Delta,\ \sum_i \bar{\alpha}_i \mathsf{y}_i = 0} \|\sum_i \bar{\alpha}_i \mathsf{y}_i \mathbf{x}_i\|_2^2 &= \min_{\boldsymbol{\mu}\in\Delta_+, \boldsymbol{\nu}\in\Delta_-} \|\sum_{i\in P} \mu_i \mathbf{x}_i - \sum_{j\in N} \nu_j \mathbf{x}_j\|_2^2 \\ &= \min_{\mathbf{x}_+\in\mathcal{D}_+} \min_{\mathbf{x}_-\in\mathcal{D}_-} \|\mathbf{x}_+ - \mathbf{x}_-\|_2^2\end{aligned}$$

Support Vector Machines: Dual