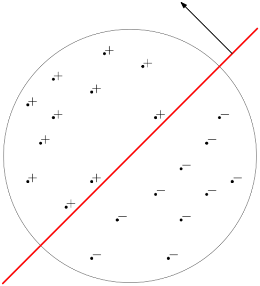
max planck institut informatik

Lecture 10: Learning halfpaces - Perceptron algorithm

Themis Gouleakis

June 22, 2021

Halfspaces



- Other names: Linear Threshold functions, Perceptrons, Linear separators, Threshold Gates, Weighted Voting Games, etc
- Extensively studied in ML since [Rosenblatt 58]

Definition: $f : \mathbb{R}^d \to \{\pm 1\}$, such that

$$f(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle - \theta), \|\mathbf{w}\|_2 = 1$$

where $\mathbf{x} \in \mathbb{R}^d, \theta \in \mathbb{R}$.

PAC Learning [Valiant 84]

$\mathcal{C}$: Known concept class of functions $f : \mathbb{R}^d \to \{\pm 1\}$.

- **Input:** Examples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \sim \mathcal{D} = (\mathcal{D}_\mathbf{x}, \mathcal{D}_y)$ supported in $\mathbb{R}^d \times \{\pm 1\}$, such that:

$$y^{(i)} = f(\mathbf{x}^{(i)})$$

  for some fixed unknown target concept $f \in \mathcal{C}$.
- **Goal:** Find a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ that minimizes $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$.

Mistake-bound model

- In each stage, the learning algorithm is given example $\mathbf{x}$ and asked to predict $f(\mathbf{x})$.
- No assumptions about the order.
- **Goal:** Bound the total number of mistakes.

> **Definition:** We say that a learner $\mathcal{L}$ learns class C with mistake bound M if $\mathcal{L}$ makes at most M mistakes on any sequence of examples consistent with some $f \in C$.

- **Note:** The sequence can have arbitrary length.
- A class C is learnable in the MB model if there exists a learner with mistake bound and running time (per stage) $poly(d, s)$, where s is the size of the smallest $f \in C$.

Example - Disjunctions

- We have boolean features $f : X \to \{0, 1\}$, where $X = \{0, 1\}^n$
- Target: OR function (e.g: $x_5 \vee x_8 \vee x_{11}$)

Can we learn with at most n mistakes in the MB model?

Example - Disjunctions

- We have boolean features $f : X \to \{0, 1\}$, where $X = \{0, 1\}^n$
- Target: OR function (e.g: $x_5 \vee x_8 \vee x_{11}$)

Can we learn with at most n mistakes in the MB model?

Example - Disjunctions

- What if most features are irrelevant? (i.e target disjunction has only r out of n variables)
- Can we do better?

Winnow algorithm:

1. Initialize: $\forall i \in [n] : w_i = 1$
2. $h(\mathbf{x})$: Predict 1 (positive) iff $w_1 x_1 + \cdots + w_n x_n \geq n$
3.
   - Mistake on positive: $w_i \leftarrow 2w_i$
   - Mistake on negative: $w_i \leftarrow 0$

Theorem: Winnow algorithm makes at most $O(r \log n)$ mistakes.

Winnow for general LTFs

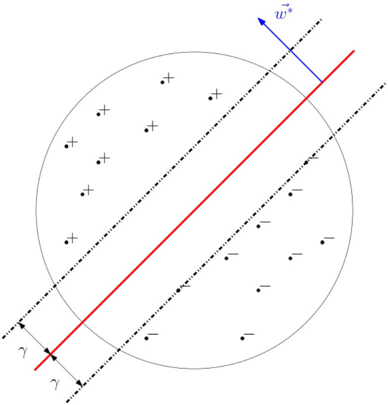
- Using similar ideas, we can learn halfspaces:
  (i.e $3x_5 + 5x_8 - 2x_{11} \geq 5$)

Theorem: Suppose that $\exists w^*$ s.t:

- $w^* \cdot \mathbf{x} \geq \gamma$ on positive $\mathbf{x}$
- $w^* \cdot \mathbf{x} \leq -\gamma$ on negative $\mathbf{x}$

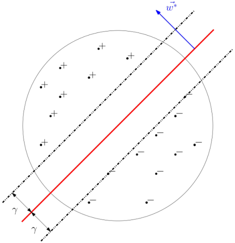
then the mistake bound is $M = O(L_1(w^*)/\gamma^2 \log n)$

Large margin assumption



- $f(\mathbf{x}^{(i)}) = sgn(\langle \mathbf{w}^*, \mathbf{x}^{(i)} \rangle)$.
- $|\langle \mathbf{w}^*, \mathbf{x}^{(i)} \rangle| > \gamma$.

Preceptron algorithm



Perceptron algorithm:

1. Initialize: $\mathbf{w} = 0$
2. $h(\mathbf{x})$: Predict 1 (positive) iff $\mathbf{w} \cdot \mathbf{x} > 0$
3.
   - Mistake on positive: $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{x}$
   - Mistake on positive: $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{x}$

Perceptron algorithm - Analysis

Perceptron algorithm - Lower bound

Perceptron algorithm - hinge loss