Learning Halfspaces

Wednesday, October 6, 2021 8:05 AM

yuhu

w, b : $\{x : w^T x \geq b\}$.

- disjunctions
- conjunctions
- decision lists

How many halfspaces? infinite?

Effectively $< 2^{n^2}$ over $\{0,1\}^n$. Why?

What about over Reals / Rationals? Later...

---

we can <u>assume $b = 0$</u>. $w^T x - b \geq 0$

$$(w, b)^T \begin{pmatrix} x \\ -1 \end{pmatrix} \geq 0.$$

and <u>assume</u> $\|w\|_2 = 1$

and $\|x\| \leq 1$ (or $= 1$).

Perceptron

Perceptron

1. $w = 0$
2. On next example x, Predict $\mathrm{sign}(w \cdot x)$

   if mistake, $w \leftarrow w + \ell(x)\, x$

Thm. Given data classifiable by a halfspace w^*, $\|w^*\| = 1$, with margin γ $\left(\min_x |w^* \cdot x|\right)$, Perceptron makes at most $\frac{1}{\gamma^2}$ mistakes.

Pf. consider $\cos(w, w^*) = \dfrac{w^T w^*}{\|w\|}$

starts at 0.

At each mistake

$$w \leftarrow w + \ell(x)\, x$$

$$w \cdot w^* \leftarrow w \cdot w^* + \underbrace{\ell(x)\; w^* \cdot x}_{\text{same sign}}$$

$w \cdot w^*$ goes up by $\geq \gamma$.

After t mistakes $\quad w \cdot w^* \geq \gamma t$.

$$\|w\|^2 \leftarrow (w + \ell(x)\, x)^T (w + \ell(x)\, x)$$

$\|w\|^2 \leftarrow (w+\ell(x)x)^T(w+\ell(x)x)$

$$= \|w\|^2 + \|x\|^2 + 2\,\underbrace{\ell(x)\,w^Tx}_{\text{opp. sign}}$$

$$\leq \|w\|^2 + 1.$$

After t steps $\|w\|^2 \leq t$.

$\therefore \cos(w, x) \geq \dfrac{\gamma t}{\sqrt{t}}$

Since $\cos \leq 1$, $\quad \gamma\sqrt{t} \leq 1 \Rightarrow t \leq \dfrac{1}{\gamma^2}$.

More generally, # mistakes $\leq \dfrac{\|w^*\|^2 \|x\|^2}{\gamma^2}$.

What if γ is super tiny?

- Modified Perceptron: correctly classifies all x with $|w^{*} \cdot x| > \sigma$, makes $O\left(\frac{\log n}{\sigma^2}\right)$ mistakes.
- Using Linear Programming, can get $\log\frac{1}{\gamma}$ dependence.

Kernels.

mapping $\phi : \mathbb{R}^n \to \mathbb{R}^m$ nonlinear.

$K(X,Y) = \langle \phi(X), \phi(Y) \rangle = \phi(X)^T \phi(Y)$.

legal kernel if such mapping exists.

i.e. $K \succeq 0$. PSD.

e.g. $\phi(x) = x$. $K(X,Y) = X^T Y$

$$\left.\begin{aligned} K(X,Y) &= \langle X, Y \rangle^d \\ K(X,Y) &= (1 + \langle X, Y \rangle)^d \\ K(X,Y) &= e^{-\|X-Y\|^2} \end{aligned}\right\} \text{legal } \checkmark.$$

Suppose $\nexists$ halfspace matching $\ell(x)$

but $\exists$ halfspace in some nonlinear map.

Then if you know the map $X \to \phi(x)$

apply Perceptron to $\phi(x)$ ....

apply Perceptron to [...]

What if you have access to $K(x, y)$

or ϕ is in very high (or infinite) dimension?

We maintain $w = \sum_i l(x^{(i)})\, x^{(i)}$

$$w \cdot x = \sum_i l(x^{(i)}) (x^{(i)} \cdot x)$$

Instead, Kernel Perceptron:

$$w = \sum_i l(x)\, \phi(x^{(i)}) \qquad \text{implicit}$$

$$w \cdot x = \sum_i l(x^{(i)})\, K(x^{(i)}, x)$$

Keep all examples on which a mistake was made.

Output as prediction on next x,

$$\text{sign}\left(\sum_i l(x^{(i)})\, K(x^{(i)}, x)\right).$$

Modified Perceptron

Modified Perceptron

$w \leftarrow$ random unit vector

On mistake, if $|w \cdot x| > \sigma \|w\|$

$w \leftarrow w + \ell(x)(w \cdot x)\, x$

Thm # mistakes $= O\left(\frac{\log n}{\sigma^2}\right)$.

Pf. assume $\ell(x) = +$ (else flip $x \to -x$)

$w \leftarrow w - (w \cdot x)\, x$

$w^* \cdot w_{\text{initial}} \geq \frac{1}{\sqrt{n}}$ with Prob $\geq \frac{1}{8}$.

$w^* \cdot w \geq w^* \cdot w - \underbrace{(w \cdot x)(w^* \cdot x)}_{\text{opp}} \geq w^* \cdot w \geq \frac{1}{\sqrt{n}}$.

$w \cdot w \geq w \cdot w - (w \cdot x)^2 \geq \|w\|^2 (1 - \sigma^2)$

after t steps $\cos(w, w^*) \geq \frac{\frac{1}{\sqrt{n}}}{(1-\sigma^2)^{t/2}}$

$\Rightarrow t \leq \frac{\ln n}{\sigma^2}$.