Topics in Machine Learning Theory

Avrim Blum

09/05/14

Lecture 4: The Perceptron Algorithm

Recap from last time

- Winnow algorithm for learning a disjunction of r out of n variables. eg $f(x)= x_3 \vee x_9 \vee x_{12}$
- h(x): predict pos iff $w_1x_1 + \ldots + w_nx_n \geq n$.
- Initialize $w_i = 1$ for all i.
  - Mistake on pos: $w_i \leftarrow 2w_i$ for all $x_i=1$.
  - Mistake on neg: $w_i \leftarrow 0$ for all $x_i=1$.
- Thm: Winnow makes at most O(r log n) mistakes.

Recap from last time

- Winnow algorithm for learning a k-of-r function: e.g., $x_3 + x_9 + x_{10} + x_{12} \geq 2$.
- h(x): predict pos iff $w_1x_1 + \ldots + w_nx_n \geq n$.
- Initialize $w_i = 1$ for all i.
  - Mistake on pos: $w_i \leftarrow w_i(1+\epsilon)$ for all $x_i=1$.
  - Mistake on neg: $w_i \leftarrow w_i/(1+\epsilon)$ for all $x_i=1$.
  - Use $\epsilon = 1/2k$.
- Thm: Winnow makes at most O(rk log n) mistakes.

Winnow for general LTFs

More generally, can show the following (you will do the analysis in class next week):

Suppose $\exists$ w* s.t.:

- $w^* \cdot x \geq c$ on positive x,
- $w^* \cdot x \leq c - \gamma$ on negative x.

Then mistake bound is

- $O((L_1(w^*)/\gamma)^2 \log n)$

Multiply by $L_\infty(X)$ if examples not in {0,1}

Perceptron algorithm

An even older and simpler algorithm, with a bound of a different form.

Suppose $\exists$ w* s.t.:

- $w^* \cdot x \geq \gamma$ on positive x,
- $w^* \cdot x \leq -\gamma$ on negative x.

Then mistake bound is

- $O(L_2(w^*)L_2(x)/\gamma^2)$

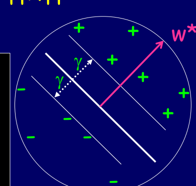
L_2 margin of examples

Perceptron algorithm

Thm: Suppose data is consistent with some LTF $w^* \cdot x > 0$, where $||w^*||=1$ and

$\gamma = \min_x |w^* \cdot x|/||x||$

Then # mistakes $\leq 1/\gamma^2$.



Algorithm:

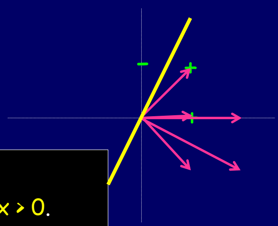
Initialize w=0. Use $w \cdot x > 0$.

- Mistake on pos: $w \leftarrow w+x$.
- Mistake on neg: $w \leftarrow w-x$.

(Pre-scale examples to be in unit ball)

Perceptron algorithm

Example: (0,1) –
(1,1) +
(1,0) +



Algorithm:
Initialize w=0. Use w · x > 0.
- Mistake on pos: w ← w+x.
- Mistake on neg: w ← w-x.

Analysis

Thm: Suppose data is consistent with some LTF w* · x > 0, where ||w*||=1 and

γ = $\min_x$ |w* · x| (after scaling so all ||x||=1)

Then # mistakes ≤ $1/\gamma^2$.

Proof: consider w · w* and ||w||
- Each mistake increases w · w* by at least γ.
  (w + x) · w* = w · w* + x · w* ≥ w · w* + γ.
- Each mistake increases w·w by at most 1.
  (w + x) · (w + x) = w·w + 2(w·x) + x·x ≤ w·w + 1.
- So, in M mistakes, γM ≤ w·w* ≤ ||w|| ≤ $M^{1/2}$.
- So, M ≤ $1/\gamma^2$.

Lower bound

It's not possible in general to get < $1/\gamma^2$ mistakes.
Proof: consider $1/\gamma^2$ coordinate vectors.

$$w^* = \pm\gamma x_1 \pm \gamma x_2 \pm \cdots \pm \gamma x_{1/\gamma^2}$$
$$||w^*|| = 1, |w^* \cdot x| = \gamma$$

Proof: consider w · w* and ||w||
- Each mistake increases w · w* by at least γ.
  (w + x) · w* = w · w* + x · w* ≥ w · w* + γ.
- Each mistake increases w·w by at most 1.
  (w + x) · (w + x) = w·w + 2(w·x) + x·x ≤ w·w + 1.
- So, in M mistakes, γM ≤ w·w* ≤ ||w|| ≤ $M^{1/2}$.
- So, M ≤ $1/\gamma^2$.

What if no perfect separator?

In this case, a mistake could cause |w · w*| to drop.
The γ-hinge-loss of w* = $\sum_x$ max[0, 1 – l(x)(x·w*)/γ]
(by how much, in units of γ, would you have to move the points to all be correct by γ)

Proof: consider w · w* and ||w||
- Each mistake increases w · w* by at least γ.
  (w + x) · w* = w · w* + x · w* ≥ w · w* + γ.
- Each mistake increases w·w by at most 1.
  (w + x) · (w + x) = w·w + 2(w·x) + x·x ≤ w·w + 1.
- So, in M mistakes, γM ≤ w·w* ≤ ||w|| ≤ $M^{1/2}$.
- So, M ≤ $1/\gamma^2$.

What if no perfect separator?

In this case, a mistake could cause |w · w*| to drop.
The γ-hinge-loss of w* = $\sum_x$ max[0, 1 – l(x)(x·w*)/γ]
Mistakes(perceptron) ≤ $1/\gamma^2$ + 2(γ-hinge-loss(w*))

Proof: consider w · w* and ||w||
- Each mistake increases w · w* by at least γ.
  (w + x) · w* = w · w* + x · w* ≥ w · w* + γ.
- Each mistake increases w·w by at most 1.
  (w + x) · (w + x) = w·w + 2(w·x) + x·x ≤ w·w + 1.
- So, in M mistakes, γM ≤ w·w* ≤ ||w|| ≤ $M^{1/2}$.
- So, M ≤ $1/\gamma^2$.

Kernel functions

See board…