Perceptron Convergence Theorem

David Helmbold

University of California, Santa Cruz
dph@soe.ucsc.edu

Fall '13 Modified W'14, Sp '15

Problem

Given a sequence of labeled examples, $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots$ where each $\mathbf{x}_i \in \Re^d$ and each $y_i \in \{+1, -1\}$, find a weight vector $\mathbf{w}$ and intercept b such that $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}_i + b) = y_i$ for all i.

Perceptron Algorithm (ignoring b):

1. initially $\mathbf{w}$ all zero's
2. for each $(\mathbf{x}_i, y_i)$ example in turn,
   if $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}_i) \neq y_i$ (a mistake) then $\mathbf{w}_{\mathrm{new}} = \mathbf{w}_{\mathrm{old}} + \eta y_i \mathbf{x}_i$.
3. repeat step 2 until convergence

(η is a learning rate, here $\eta = 1$ so $\eta y \mathbf{x}_i$ adds/subtracts $\mathbf{x}_i$)

Theorem: If the data is linearly separable then the Perceptron Algorithm converges to some hyperplane $\mathbf{w} \cdot \mathbf{x} + b$ that separates the positive and negative examples.

Proof Outline

1. Simplify,
2. Simplify,
3. Simplify,
4. Look at cosine between $\mathbf{w}$ and a good vector $\mathbf{u}$.
   Bound numerator and denominator in terms of $\#$ of mistakes and then solve for $\#$ of mistakes.

Simplification 1

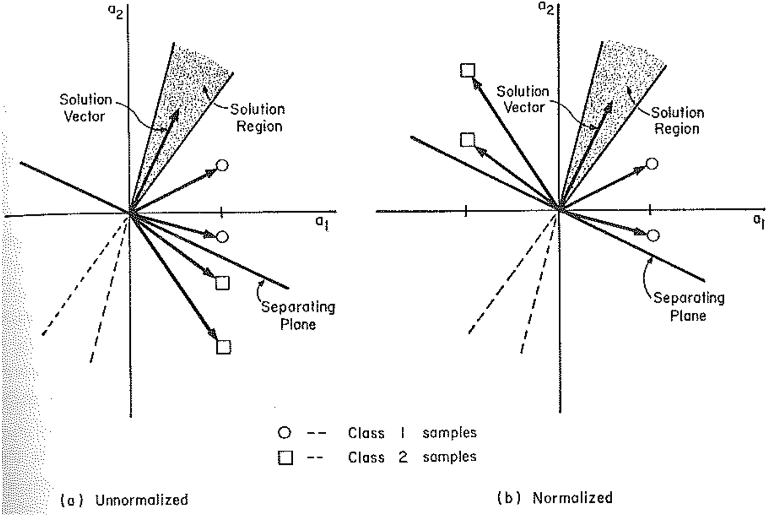
Eliminate intercept b.
Use the "add-a-dimension" trick, and solve the

$$\textit{find a } \mathbf{w} \textit{ such that } \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}) = y$$

problem.

Simplification 2

Avoid negative examples. Replace each $(\mathbf{x}, -1)$ example with $(-\mathbf{x}, +1)$.



from "Pattern Classification and Scene Analysis"', Duda and Hart, 1973

Simplification 3

Normalize lengths of $\mathbf{x}$'s.
Rescale instances to have length 1, so $\mathbf{x} \cdot \mathbf{x} = 1$ for all instances.
(note: rescaling $\mathbf{x}$ doesn't change sign of $\mathbf{w} \cdot \mathbf{x}$, does change x_0's)

Not done by Ng – he uses upper bound D on instance lengths in ...notes6.pdf

With simplifications and setting $\eta = 1$, algorithm becomes:

- initially $\mathbf{w}$ all zero's
- predict with $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$
- if $\mathbf{w} \cdot \mathbf{x} \leq 0$ (a mistake made) then $\mathbf{w}_{\mathrm{new}} = \mathbf{w}_{\mathrm{old}} + \mathbf{x}$.

Analysis Setup

- Let $\mathbf{u}$ be any good (correct) weight vector with $||\mathbf{u}||_2 = 1$.
- Define the gap δ be $\min_i \mathbf{u} \cdot \mathbf{x}_i$.
  (after rescaling $\mathbf{x}$'s; "better" $\mathbf{u}$ have bigger gaps)
- Since $\mathbf{u}$ correct, $\delta > 0$.
- Consider $\cos(\mathbf{u}, \mathbf{w}) = \dfrac{\mathbf{u} \cdot \mathbf{w}}{||\mathbf{w}||_2}$.
- Cosine always ≤ 1.

- Each mistake, $\mathbf{w}_{\text{new}} := \mathbf{w}_{\text{old}} + \mathbf{x}$.
- Note $\mathbf{u} \cdot \mathbf{w}_{\text{new}} = \mathbf{u} \cdot (\mathbf{w}_{\text{old}} + \mathbf{x}) \geq \mathbf{u} \cdot \mathbf{w}_{\text{old}} + \delta$,
- so $\boxed{\mathbf{u} \cdot \mathbf{w}_{\text{new}} \geq \delta \times (\#\ \text{mistakes so far})}$.

- Now bound $||\mathbf{w}||_2$ by considering $||\mathbf{w}||_2^2$ on a mistake,

$$\begin{aligned} ||\mathbf{w}_{\mathrm{new}}||_2^2 &= \mathbf{w}_{\mathrm{new}} \cdot \mathbf{w}_{\mathrm{new}} && (1) \\ &= (\mathbf{w}_{\mathrm{old}} + \mathbf{x}) \cdot (\mathbf{w}_{\mathrm{old}} + \mathbf{x}) && (2) \\ &= \mathbf{w}_{\mathrm{old}} \cdot \mathbf{w}_{\mathrm{old}} + \underbrace{\mathbf{x} \cdot \mathbf{x}}_{=1} + 2 \underbrace{\mathbf{w}_{\mathrm{old}} \cdot \mathbf{x}}_{\text{negative}} && (3) \\ &\leq ||\mathbf{w}_{\mathrm{old}}||_2^2 + 1 && (4) \end{aligned}$$

- Therefore, $||\mathbf{w}_{\mathrm{new}}||_2^2 \leq (\#\ \text{mistakes})$, and

$$\boxed{||\mathbf{w}_{\mathrm{new}}||_2 \leq \sqrt{(\#\ \text{mistakes})}}$$

Ng uses $\mathbf{x} \cdot \mathbf{x} \leq D^2$, so $||\mathbf{w}_{\mathrm{new}}||_2 \leq \sqrt{D^2(\#\ \text{mistakes})}$

Finishing up

Thus we always have the inequalities:

$$1 \geq \frac{\mathbf{u} \cdot \mathbf{w}}{||\mathbf{w}||_2} \geq \frac{\delta(\#\text{ mistakes})}{\sqrt{(\#\text{ mistakes})}}$$

and solving for $(\#$ mistakes$)$ gives

$$(\#\text{ mistakes}) \leq \frac{1}{\delta^2}$$

Why does it work?

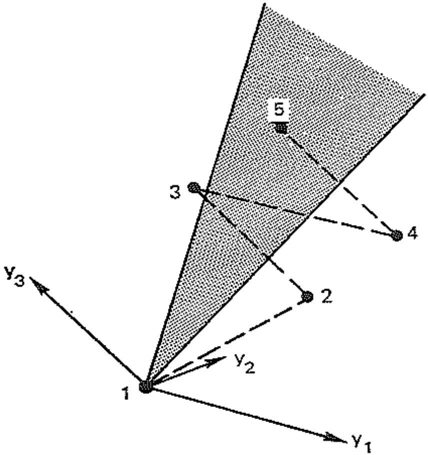


FIGURE 5.9. Finding a solution region by a gradient search.

from “Pattern Classification and Scene Analysis”’, Duda and Hart, 1973