Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
9/19/05: **Lecture 11 aid – The perceptron convergence theorem**

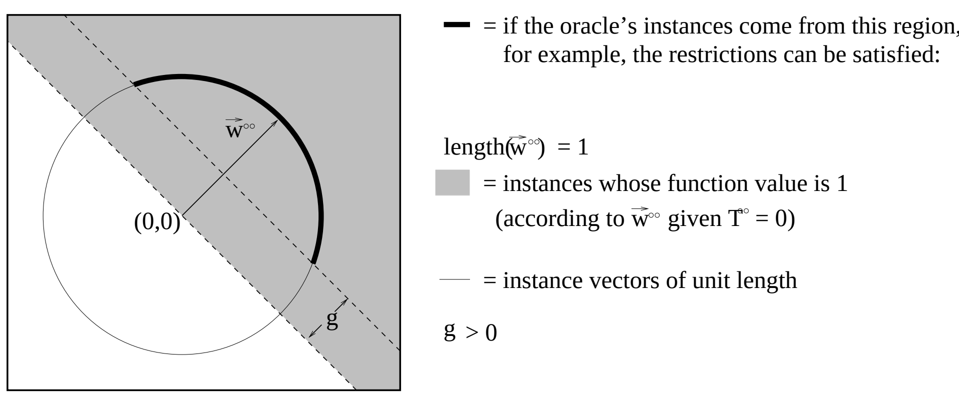
Agenda: Restrictions on the on-line perceptron-learning setting; a "one-sided" version of Rosenblatt's perceptron learning algorithm; a corresponding version of the perceptron convergence theorem.

I. Reminders Recall from the lecture 10 aid of 9/16/05 that we use $\overrightarrow{x}^{(i)}$ to denote the i^{th} example presented by the oracle, and that *positive* instances are those with label +1, whereas *negative* instances are those with label -1. Recall from the lecture 9 aid (9/14/05) that: the inner product distributes in the expected way; the length of a vector $\overrightarrow{v}$ can be computed as $\sqrt{\overrightarrow{v} \cdot \overrightarrow{v}}$; vector addition and subtraction is component-wise; and that the cosine between two vectors is the inner product of the two divided by their respective lengths.

II. Restrictions and/or simplifications we impose More general versions of these restrictions can also be considered.

1. The *one-zero consistency* condition: All the labeled examples turn out to be consistent with some perceptron function $f_{\overrightarrow{w}^{\infty}, T^{\infty}}$ where length $(\overrightarrow{w}^{\infty}) = 1, T^{\infty} = 0$.
2. The *length restriction*: For all i, length $\left(\overrightarrow{x}^{(i)}\right) = 1$.
3. The *gap condition*[1]: There is a $g > 0$ such that for all $\overrightarrow{x}^{(i)}$ and the $\overrightarrow{w}^{\infty}$ specified above, we have that $\overrightarrow{w}^{\infty} \cdot \overrightarrow{x}^{(i)} \geq g$.

Here is what all this looks like in two dimensions:



[1]The real but (slightly) harder to work with version of this condition is "double-sided", requiring only that $|\overrightarrow{w}^{\infty} \cdot \overrightarrow{x}^{(i)}| \geq g$. This corresponds to having a gap, or *margin*, between the positive and negative examples, and eliminates "cheating" solutions (such as setting the weight vector to all-zeroes always) on the part of the learner.

III. The perceptron learning algorithm This is a "one-sided" version of the algorithm Rosenblatt proposed.

1) Set $\overrightarrow{w}^{(0)}$ to all zeroes.
2) For each example $\overrightarrow{x}^{(i)}$ (i increasing from 1 on),
3) If $\overrightarrow{w}^{(i-1)} \cdot \overrightarrow{x}^{(i)} \leq 0$,
4) set $\overrightarrow{w}^{(i)}$ to $\overrightarrow{w}^{(i-1)} + \overrightarrow{x}^{(i)}$ ("update");
5) otherwise, set $\overrightarrow{w}^{(i)}$ to $\overrightarrow{w}^{(i-1)}$ ("no change").

IV. Outline of the proof of (our version of) the perceptron convergence theorem Given all the constraints we have about the oracle and learner,

- Use the cosine function to measure how "close" successive hypothesis vectors are to $\overrightarrow{w}^{\infty}$. Observe that it takes the form N/D (numerator over denominator), and that we can think of it as "starting" at $0 = 0/\sqrt{1}$.
- Show that at each *update* of the perceptron learning algorithm, i.e., where $\overrightarrow{w}^{(i)}$ is different from $\overrightarrow{w}^{(i-1)}$, the cosine increases by a non-negligible amount:
  - N increases by *at least* g, the *gap* quantity.
  - The square of D increases by *at most* 1.

  Hence, after t updates, the cosine must be at least $\sqrt{t}g$.
- Since cosines can't get bigger than one, we get that t can be at most $1/g^2$, which, since $g > 0$, implies only a finite number of updates, and hence a finite number of mistakes, gets made.