Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
9/23/05: **Lecture 13 aid – Conclusion of learning; beginning of information retrieval**

Agenda: concluding the discussion of our "one-sided" perceptron learning algorithm; we may commence information retrieval.

Follow-ups: Note that by showing last time that the "denominator" term D increases by at most one at each update, we actually finished our proof of the perceptron convergence theorem, because we completed filling in the proof outline given on the Lecture 11 aid (9/19/05).[1]

We note that our algorithm, theorem, and proof, although perhaps seemingly straightforward in retrospect, are actually quite clever (we will explore one aspect of this today, and you will explore another aspect in your homework). We do *not* expect you to come up with something similar on your own, but we *do* expect you to understand how the proof is put together and how it makes use of the restrictions on the oracle and the specific definition of the PLA.

I. Reminder: the restrictions we impose

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*: There is a $g > 0$ such that for all $\overrightarrow{x}^{(i)}$ and the $\overrightarrow{w}^\infty$ specified above, we have that $\overrightarrow{w}^\infty \cdot \overrightarrow{x}^{(i)} \geq g$.

II. Reminder: 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").

III. The "normalized" perceptron learning algorithm What happens if we don't use length information in our hypothesis vectors? After all, if we choose to have the threshold be 0, then in terms of where the dividing line (or hyperplane) lies, the weight vector's length is irrelevant.

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

---

[1] Actually, there is one minor additional argument we need to make: that after the first oracle instance $\overrightarrow{x}^{(1)}$ is presented, the learner never hypothesizes the all-zeroes vector as its weight vector (if it did, that would mean D= 0, which would ruin our upper bound on N/D). But, note that $\overrightarrow{w}^{(i)}$ for $i \geq 1$ is the sum of some of the instance vectors presented so far, each of which has positive inner product with $\overrightarrow{w}^\infty$ (why?). This implies by distributivity that $\overrightarrow{w}^{(i)}$ does too, and the all-zeroes vector doesn't have positive inner product with the unit-length vector $\overrightarrow{w}^\infty$.