Adaptions of Yoav's Self bounding learning algorithms

John Langford jcl@cs.cmu.edu 8/10/98

August 12, 1998

1 Theorem 1

Theorem 1 states that $\forall\alpha > 0, \forall\epsilon > 0$ the generalization error of the final hypothesis is bounded by:

$$P_{S\sim D^m}[err(h^*) - \widehat{err}(h^*) > \epsilon] \leq 2|Q(T_A(D,\alpha))|e^{-2m\alpha^2} + |H(T_A(D,\alpha))|e^{-2m\epsilon^2}$$

The first term in the bound bounds the probability that some query will return an answer which is 'bad' enough to step off the tree, $T_A(D,\alpha)$. The second term bounds the (agnostic) difference between the empirical and true error of the hypothesis.

1.1 first generalization

We can generalize this theorem to state:

$$P_{S\sim D^m}[err(h^*) - \widehat{err}(h^*) > \sqrt{\frac{1}{2m}\log\frac{1}{\delta_{h^*}}}] \leq 2|Q(T_A(D,\alpha))|e^{-2m\alpha^2} + \delta$$

where $\delta = \sum_{h\in H(T_A(D,\alpha))}\delta_h$. One useable choice of δ_h is $\frac{\delta}{|Q_{1T}||Q_{2T}|\ldots|Q_{dT}|}$ where Q_{iT} = the branching factor (contained in T_A) at step i in the algorithm and the chain of queries leads to the choice of hypothesis, h.

1.2 second generalization

I believe (though I'm unsure not having read Kearns paper) that a similar trick can be played with the first term to apportion confidence amongst the various queries in the tree.

2 A slightly more general microchoice theorem

In http://www.cs.cmu.edu/˜jcl/microchoice/moremicro.ps, section 5, we gave a theorem that a bound on the generalization error could be given by:

$$P_{S\sim D^m}[err(h^*) - \widehat{err}(h^*) > \sqrt{\frac{1}{2m}\log\frac{1}{\delta_{h^*}}}] \leq \delta$$

with $\delta_h = \frac{\delta}{|H_{I1}|*|H_{I2}|*\ldots*|H_{IM}|}$.

This theorem can be made slightly more general by moving to a setting where a 'candidate hypothesis' is not required. Instead, identify every choice node in an algorithm - choice nodes are locations where choices which could affect the final hypothesis are made. For example, one choice node might rest on the output of a statistical query. Denote by C_i these choice nodes. Then we can replace the division of our confidence by, $\delta_h = \frac{\delta}{|C_1|*|C_2|*\ldots*|C_d|}$ where d is the last choice made resulting in the choice of hypothesis, h.

3 Comparison with Yoav

The more general microchoice theorem is similar to Yoav's work because it includes the choice nodes which allows application to algorithms which use statistical queries. The more generali microchoice theorem is tighter for cases where $T_A(D, \alpha)$ includes all possible choices along the path of the algorithm. Yoav's theorem is tighter for cases where

- The algorithm happens to choose a final hypothesis with an above average inverse measure. (doesn't hold for the first generalization above)
- The branching factor of the tree is smaller than the branching factor of the choice points.

The second case can be removed by changing the measure to reflect which choices are 'low probability' choices by using a 'luckiness function' across a set of choices.