An analysis of microchoice learning algorithms

John Langford

August 5, 1998

Consider a learning algorithm that proceeds by making a sequence of small choices which move the algorithm through the hypothesis space. If the true error for each (small) hypothesis space can be bounded, then we can bound the true error for all the choices that the learning algorithm makes by a triangle inequality. The proof of such a bound is straight forward, but the application to a particular algorithm may not be so straight forward.

1 related work

Pedro Domingos [1] works in essentially the same domain to derive a heuristic which attempts to approximate the generalization error. The difference in this work is that I want to derive a bound rather than a heuristic. Yoav Freund[2] worked on self-bounding learning algorithms which is a more general setting than I am considering here.

2 A theorem to bound the error of many microchoices

2.1 starting point

Take as a starting point one of the generalized PAC theorems by Haussler:

Theorem: Assume that H_I and the loss function, l, are such that H_I^l is permissible. Let P be any probability distribution on Z. Assume that $N \geq 1$, $\nu > 0$, and $0 < \alpha < 1$. Let z_N be generated by N independent draws from Z according to P. Then $P(\exists H_i \in H_I : d_\nu(E(H_i^l, P), \hat{E}(H_i^l, z_N)) > \alpha) \leq 4C(\frac{\alpha\nu}{8}, H_I)e^{\frac{-\alpha^2\nu N}{8}}$. This is essentially the normal PAC theorem generalized in several ways. For any nonobvious definitions, please look in: http://www.cs.cmu.edu/~jcl/ltol/staged.ps.

It's useful to put this equation into a simpler form. $E(H_i^l, P) > \hat{E}(H_i^l, z_N) + \epsilon$ implies $d_\nu(E(H_i^l, P), \hat{E}(H_i^l, z_N)) > \frac{\epsilon}{2+\nu} = \alpha$. Maximizing $\alpha^2\nu$ gives the substitutions: $\alpha = \frac{\epsilon}{3}, \nu = 1$, which implies:

$$P(\exists H_i \in H_I : E(H_i^l, P) > \hat{E}(H_i^l, z_N) + \epsilon) \leq 4C(\frac{\epsilon}{24}, H_I)e^{\frac{-\epsilon^2 N}{72}}$$

Letting, $4C(\frac{\epsilon}{24}, H_I)e^{\frac{-\epsilon^2 N}{72}} = \delta$ and solving for ϵ, we get: $\epsilon^2 = -\frac{72}{N}\ln\frac{\delta}{4C(\frac{\epsilon}{24}, H_I)}$ which implies $\epsilon = \sqrt{\frac{72}{N}(\ln\frac{1}{\delta} + \ln 4C(\frac{\epsilon}{24}, H_I))}$, which implies:

$$\overline{er}(H_i) \leq \hat{er}(H_i, S) + 6\sqrt{2}\sqrt{\frac{\ln C(\frac{\epsilon}{24}, H_I) + \ln\frac{1}{\delta}}{N}}$$

2.2 the plan

Now, if we have an algorithm (such as backprop) which makes a sequence of choices to move through a large hypothesis space, there are two ways to bound the error, one is by attempting to find the capacity, $C(\epsilon, H_I)$ of the space of all reachable hypothesis and make an a. priori bound. Unfortunately, this is not very good in almost all cases for several reasons:

1. The H_I used may just be a (loose) bound on the real $H_I =$ all hypothesis which the learning algorithm could possibly reach.
2. The learning algorithm will often terminate at some point, quite possibly well before it can explore the full H_I.

If we instead, look at the algorithm as making a sequence of choices in hypothesis spaces, $H_{I1}, H_{I2}, ..., H_{IM}$, any one of which may go wrong then it may be possible to give a better bound for the true error of the learning algorithm. If we apply the PAC theorem above to each individual hypothesis space with $\alpha \rightarrow \frac{\alpha}{M}$ and $\delta \rightarrow \frac{\delta}{M}$, then using the triangle inequality we get:

Simple theorem: If $\overline{er}(H_{im}) \leq \hat{er}(H_{im}, S) + 6\sqrt{2}\frac{1}{M}\sqrt{\frac{\ln C(\frac{\epsilon}{24M}, H_{Im}) + \ln\frac{1}{\delta M}}{N}}$ for each hypothesis space a choice is made in, H_{Im}, then the final output hypothesis has error bounded by $\overline{er}(H_{final}) \leq \hat{er}(H_{final}, S) + \frac{6\sqrt{2}}{M}\sum_{m=1}^{M}\sqrt{\frac{\ln C(\frac{\epsilon}{24M}, H_{Im}) + \ln\frac{1}{\delta M}}{N}}$ with probability $1-\delta$. Proof: triangle inequality applied $M-1$ times.

2.3 simple extensions

There are several simple extensions to this theorem.

1. The confidence can be divided nonuniformly amongst the individual choice spaces so that for space, H_{Im} the chance of a "bad" hypothesis is bounded by $a_m\delta_m$ and $\sum_m a_m = 1$.

2. The error can also be split up in a similar manner allocating $b_m \epsilon_m$ such that $\sum_m b_m = 1$.

3. Take as a starting point, Dave McAllester's PAC-Bayes theorem from COLT98. This is sometimes tighter than Haussler's theorem and allows the incorporation of a prior into the learning algorithm.

2.4 More complicated improvements

Because the PAC bounds hold for a range of ϵ, δ values simply adding together these two bounds on random variables is not optimal. Instead, a convulution could be done at each step to find a final distribution over errors at the final step, which could then be converted into a better (ϵ, δ) bound.

2.5 application to learning algorithms

These bounds may (or may not...) be meaningful for real machine learning algorithms. Consider standard backprop. If we instrument backprop so that at every update of the weights, the algorithm first tries to find the size of the hypothesis space from which it is choosing in the weight update, then we could give an alternative (to the normal, terribly loose) bound without interfering with the algorithm. Fortunately, it is not necessary to enumerate the full choice space - random draws from a space of unknown size should be able to determine the size of the space in $O(\sqrt{C(\epsilon, H_{Im})})$, which hopefully isn't too large. I'm planning to look into whether or not this process can be useful.

References

[1] Domingos, Pedro, "A Process-Oriented Heuristic for Model Selection", ICML 98, http://www.gia.ist.utl.pt/˜pedrod/mlc98.ps.gz

[2] Freund, Yoav, "Self bounding learning algorithms", COLT 98, http://www.research.att.com/˜yoav/papers/lsearch.ps.gz