More microchoice discussion

Avrim Blum and John Langford

August 7, 1998

1 A quick counter example

Here's a quick counter example to the proof I gave earlier: Consider a two stage algorithm which makes a choice in one continuous space, H_{I1}, then a choice in a second space, H_{I2}, determined by the first choice. In particular, this algorithm will choose a hypothesis, H_{i1}, which encodes all of the examples in the first space. The attached hypothesis space, H_{I2}, at this hypothesis just contains a hypothesis which has delta functions at every positive instance. Here we have a hypothesis with 0 training error which generalizes very poorly.

2 A simple discrete case

If a sequence of discrete hypothesis spaces, H_{I1}, H_{I2}, ..., H_{IM} are fixed in advance then only the confidence, δ, needs to be partitioned over each hypothesis space. This is essentially structural risk minimization and was first pointed out by Shaw-Taylor and others and again by McAllester.

3 Partitioning the examples

If the sequence of hypothesis spaces, H_{I1}, H_{I2}, ..., H_{IM} are not fixed in advance but instead the examples are partitioned with $\frac{N}{M}$ new examples used in each. Then just use the standard bound with $N \to \frac{N}{M}$ and $\delta \to \frac{\delta}{M}$ will apply for the final hypothesis chosen from the final space. Proof: the new set of examples will be an unbiased estimator on the space so the standard bounds must apply assuming none of the δ_i confidences have been violated.

There are several interesting things about this example. First of all, it is the setting of online algorithms. Second, this bound is actually better than the earlier one I claimed. To see this, note that partitioning $\epsilon \to \frac{\epsilon}{2}$ means $\sqrt{\frac{\ln C(\frac{\epsilon}{24}, H_I) + \ln\frac{1}{\delta}}{N}} \to \sqrt{\frac{\ln C(\frac{\epsilon}{24}, H_I) + \ln\frac{1}{\delta}}{4N}}$ which implies that $N \to 4N$. With the new bound, in order to maintain the same confidence we need to merely do $N \to 2N$. This discrepancy disappears in the realizable case.

4 same size child spaces

The sequence of discrete hypothesis spaces are not fixed in advance, but instead we know that all possible H_{Im} are of the same size for a particular $m \in [1, ..., M]$. Then we can use a bound based on $|H_{I1}|*|H_{I2}|*...*|H_{IM}|$ = the total number of possible hypothesis. proof: we have bounded the number of possible hypothesis so the normal bound applies.

5 different size child spaces

There is a sequence of discrete hypothesis spaces which are not fixed in advance and of varying sizes. In other words, the algorithm proceeds by making a choice from H_{I1} which implies an attached (yum, fibre bundles in machine learning!), H_{I2}, from which a choice is made leading to an attached H_{I3}, and so on.... The bound still uses $|H_{I1}| * |H_{I2}| * ... * |H_{IM}|$. Proof: we imagine an a. priori assignment of measure/probability/weight (I'm not sure what to call this, "measure" seems most correct, but McAllester used "probability") to every every final hypothesis, H_{final} which is: $M(H_{final}) = \frac{1}{|H_{I1}|*|H_{I2}|*...*|H_{IM}|}$. All of these assignments will sum to one, $\sum_{H_{final}} M(H_{final}) = 1$. When taking the union bound in the standard discrete learning theory, we can unevenly weight the (normally uniform) confidence assigned to each hypothesis.

This bound is interesting. We do not need to actually calculate $M(H_{final})$ for every H_{final}. Instead, we can calculate $M(H_{final})$ by the following process: Start with an estimate of $e := 1$. Every time a choice is made from a hypothesis space, H_{Im}, update the estimate by: $e := \frac{e}{|H_{Im}|}$. When the algorithm finishes it will have a bound on the true error which is dependent on the data. Furthermore, this bound is calculated using only the sizes of spaces which the algorithm actually makes a choice in.

6 Some interesting cases to think about

6.1 random then algorithm

Consider an algorithm which works with the following, admittedly odd, method. Before looking at the data, a true random source is used to make a choice between several hypothesis space. Then the algorithm learns on the space it has chosen, outputing both it's final choice and an indication of which hypothesis space it learned on. What is the bound on the error for this algorithm? It should be the same bound as an algorithm which simply learns on the randomly chosen hypothesis space.

6.2 algorithm then random

Now switch the order of events. An algorithm learns on a particular hypothesis space to get a hypothesis. Then an attached hypothesis space has a choice made uniformly on it (without looking at the data). I believe this will have the same error bound as an algorithm which doesn't make the same random choice. There is a theorem in decision theory related to this problem.

7 Extensions to the continuous case

Clearly we need to induce a metric between hypothesis spaces. This is a bit difficult because hypothesis spaces because two hypothesis spaces can have varying size. Perhaps the following will do:

$$< H_I, H_J >= \frac{1}{2}(\int_{H_J} min_{H_I} < H_i, H_j > dP(H_j) \ + \int_{H_I} min_{H_J} < H_i, H_j > dP(H_i))$$

This says that the distance between two hypothesis spaces is found by averaging over the minimal distance between two two elements of the hypothesis space weighted by a measure $P(H_i)$ and $P(H_j)$ over the hypothesis spaces. This metric isn't entirely satisfactory to me because if $H_i = H_j$ but the measures are different the metric will still give 0. Any ideas for a better metric?