

15-859(B) Machine Learning Theory

Homework # 6

Due: April 24, 2007

Groundrules: Same as before. You should work on the exercises by yourself but may work with others on the problems (just write down who you worked with). Also if you use material from outside sources, say where you got it.

Exercises:

1. **DFAs.** A *distinguishing sequence* for a DFA is a sequence of actions such that the observations produced from these actions uniquely determines the starting state. I.e., a sequence h such that if $q \neq q'$ then $\text{obs}(q, h) \neq \text{obs}(q', h)$.
 - (a) Describe a strongly-connected DFA that has no distinguishing sequence. Note that the definition of “ $q \neq q'$ ” is that there must exist a sequence $h_{qq'}$ such that $\text{obs}(q, h_{qq'}) \neq \text{obs}(q', h_{qq'})$, it's just that no single h works for all pairs.
 - (b) Give a homing sequence for your DFA.
2. **Mistake-bound odds and ends.** A mistake-bound algorithm is called “conservative” if it only updates its hypothesis when it makes a mistake. Show that without loss of generality we can consider only conservative algorithms when examining worst-case mistake bounds. That is, if there exists any algorithm for learning class C with mistake-bound M then there must exist a conservative algorithm for learning C with mistake-bound M .

Problems:

3. **Sample complexity bounds.** For some learning algorithms, the hypothesis produced can be uniquely described by a small subset of k of the training examples. E.g., if you are learning an interval on the line using the simple algorithm “take the smallest interval that encloses all the positive examples,” then the hypothesis can be reconstructed from just the outermost positive examples, so $k = 2$. For a conservative Mistake-Bound learning algorithm, you can reconstruct the hypothesis by just looking at the examples on which a mistake was made, so $k \leq M$, where M is the algorithm's mistake-bound. (In this case, you may also care about the *order* in which those examples arrived.)

Prove a PAC guarantee based on k . Specifically, fixing a description language (reconstruction procedure), so for a given set S' of examples we have a well-defined hypothesis $h_{S'}$, show that

$$\Pr_{S \sim D^n} \left(\exists S' \subseteq S, |S'| = k, \text{ such that } h_{S'} \text{ has 0 error on } S - S' \text{ but true error } > \epsilon \right) \leq \delta,$$

so long as

$$n \geq \frac{1}{\epsilon} \left(k \ln n + \epsilon k + \ln \frac{1}{\delta} \right).$$

Hint: Think of S' as a subset of indices, and imagine drawing points in S by drawing those in S' first.

Note the similarity of the form of this bound to VC-dimension and other bounds we have seen. These are often called “compression bounds”.