15-850(C) Machine Learning Theory

Assignment 1 Due: September 28, 1995

Ground rules: If you get stuck trying to solve the problems by yourself, its fine to work together, but in that case acknowledge the help you received.

1. **The difficulty of 2-term DNF.** Prove that the consistency problem for 2-term DNF formulas is NP-hard (this was shown by Pitt and Valiant [1]).

   Hint 1: reduce from the NP-complete Set-Splitting problem (also called the hypergraph 2-colorability problem), which is the following. You are given m subsets $S_1, S_2, \ldots, S_m$ of the set $\{1, \ldots, n\}$. You want to know if there is a coloring of the numbers $1, \ldots, n$ in which each number is colored either red or blue, so that none of the sets S_i has all its elements the same color.[1] For instance, if the sets are $\{1,2,3\}$, $\{2,3\}$, and $\{1,2\}$, a legal coloring might give the numbers 1, 2, and 3 the colors red, blue, and red respectively.

   Hint 2: Think about what it means if examples 011111 and 101111 are positive but 001111 is negative.

2. **Getting rid of δ.** Suppose we have an algorithm $\mathcal{A}$ for learning concept class C in the PAC model, but it only succeeds for $\delta = 1/2$. In other words, for some polynomial p, after seeing $p(n, 1/\epsilon, size(c))$ examples, the algorithm has at least a 50% probability of producing a hypothesis with error at most ϵ. Show how you can use such an algorithm to construct a PAC-learning algorithm for C in the usual sense (it works for arbitrary $\delta > 0$).

References

[1] Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, 1988.

---

[1] It's not needed for the question, but if you want to know why *this* problem is NP-hard, there's an easy reduction from 3-SAT in which the two colors correspond to True and False, there is a set $\{x_i, \bar{x}_i\}$ for each variable x_i, and there is a set $c_j \cup \{F\}$ for each clause c_j, where "F" is an element whose color represents false.