MCS 549 – Mathematical Foundations of Data Science
Fall 2022
Problem Set 3

Lev Reyzin

Due: 12/2/22 at the beginning of class

Instructions: Atop your problem set, please write your name and list your collaborators.

Problems

1. Let the domain be $\mathcal{R}$ and the concept class $\mathcal{C}_s$ (assume the learner knows s) be the class of concepts defined by unions of s intervals: i.e. c is defined by $a_1 \leq a_2 \leq \ldots \leq a_{2s-1} \leq a_{2s} \in \mathcal{R}$ and $c(x) = 1$ if $x \in [a_1, a_2] \cup [a_3, a_4] \cup \ldots \cup [a_{2s-1}, a_{2s}]$. Show that C_s is efficiently PAC learnable.

2. Consider modifying the definition of PAC learning by getting rid of the δ parameter and letting ϵ serve as a bound on both the approximation error and the failure probability. In essence, the learner would be asked to produce a hypothesis h_S such that

$$\Pr_{S \sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \epsilon$$

using a sample size polynomial in $1/\epsilon$, and the dependence on the other parameters would remain unchanged. Does this redefinition change which classes of functions are PAC learnable?

3. Give a streaming algorithm to select symbol i with probability proportional to a_i^2, where each a_i is in $\{1, \ldots, m\}$.

4. Give an example of a set H of hash functions such that $h(x)$ is equally likely to be any element of $\{0, ..., M-1\}$ but H is not 2-universal. Prove your answer correct.

5. For the k-median and the k-means objectives, prove upper bounds on the ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points (sometimes called "Steiner points") to be centers.

6. Fill in the details for the dynamic programming algorithm for clustering n points on the line using k clusters. Let $\mathrm{OPT}(\ell, i)$ be the optimal clustering for points $a_1, \ldots, a_i$ using $\ell \leq k$ clusters for $i \leq n$. As part of your answer, make sure to write this as a function of "smaller" values of ℓ and i. Use this to derive the complexity of finding $\mathrm{OPT}(k, n)$.