

MCS 549 – Mathematical Foundations of Data Science

Fall 2024

Problem Set 3

Lev Reyzin

Due: 12/6/24 at the beginning of class

Problems

1. 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?

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

3. 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.

4. Fill in the details for the dynamic programming algorithm for clustering n points on the line using k clusters. Let $\text{OPT}(\ell, i)$ be the optimal clustering for points a_1, \dots, 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 $\text{OPT}(k, n)$.