MCS 549 – Foundations of Data Science
Fall 2026
Problem Set 3

Lev Reyzin

Problems

1. In boosting, suppose

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \qquad \alpha_t \geq 0, \qquad \sum_{t=1}^{T} \alpha_t = 1,$$

where each h_t takes values in $\{-1, 1\}$. Assume every training example satisfies $y_i F(x_i) \geq \gamma > 0$. Some classifiers are replaced by arbitrary $\{-1, 1\}$-valued functions, with their weights unchanged. Prove that all training examples remain correctly classified if the total weight of the replaced classifiers is less than $\gamma/2$. Show that the constant $1/2$ cannot be improved.

2. Consider thresholds on $(0, 1)$:

$$h_\theta(x) = \begin{cases} +1, & x \geq \theta, \\ -1, & x < \theta, \end{cases} \qquad \theta \in (0, 1).$$

Show that, for every deterministic online learner and every positive integer T, there is a sequence of T labeled examples, consistent with a single threshold, on which the learner makes T mistakes. Why does this not contradict the fact that thresholds have VC dimension 1 and are PAC learnable?

3. Construct a distribution on graphs with $n \geq 3$ vertices s.t. every possible edge is present with probability $1/2$, the indicators of any two distinct edges are independent, and the graph contains no triangle with probability 1. In contrast, prove that $G(n, 1/2)$ contains a triangle with probability tending to 1 as $n \to \infty$.

4. Consider the random graph model with a prescribed bounded degree distribution. In the *subcritical* regime, all connected components have size $O(\log n)$ with high probability. In the *supercritical* regime, a connected component of size $\Theta(n)$ exists with high probability. Use the Molloy–Reed:

1. Suppose a fraction a of the vertices have degree 1, and the rest have degree 3. Determine the strictly subcritical and strictly supercritical ranges of a.
2. Give two degree distributions with the same expected degree, one strictly subcritical and the other strictly supercritical.

5. Assign independent uniform hash values in $(0, 1)$ to distinct items, reusing an item's value whenever it reappears. Suppose a stream contains $d \geq 2$ distinct items. Let R_1 and R_2 be the smallest and second-smallest hash values among those distinct items.

1. Show that $\mathbb{E}[1/R_1] = \infty$.
2. Show that $1/R_2$ is an unbiased estimator of d. Explain how to maintain R_2 in one pass while handling repeated items correctly.

6. A stream is divided into blocks. On each block, run the majority-vote streaming algorithm, retaining its final candidate and counter. A counter of zero represents an empty summary.

1. Give a rule for combining two such summaries into one summary. Prove that repeatedly applying your rule, in any order of pairwise combinations, returns the global majority item whenever one exists. A global majority item occurs more than half the time in the entire stream.

2. Show that simply taking a majority vote of the block candidates can fail, even when all blocks have equal length and each block has a strict majority.

Hint: Interpret each summary as the items remaining after cancellations of pairs of unequal items.