Assignment Six
ECE 4200/5420

October 21, 2021

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 10/29/2021, 23.59.59 eastern time**.
- Submission rules are the same as previous assignments.

Problem 1. (10 points). Suppose AdaBoost is run on n training examples, and suppose on each round that the weighted training error ε_t of the tth weak hypothesis is at most $\frac{1}{2} - \gamma$, for some number $\gamma > 0$. Show that after $T > \frac{\ln n}{2\gamma^2}$ rounds of AdaBoost the final combined classifier has **zero training error**!

Problem 2. (10 points). Recall Bagging. Starting from a training set S of size n, we created R bootstrap training sets $S_1, \ldots, S_R$, each of size n each by sampling with replacement from S.

1. For the bootstrap sample S_1, what is the expected fraction of the training set that does not appear at all in S_1? As $n \to \infty$, what does this fraction approach?
2. What is the probability that the example (X_i, y_i) does not appear in **any of** the bootstrap samples $S_1, S_2, \ldots, S_R$. In other words, what is $\Pr\left((X_i, y_i) \notin S_1 \cup S_2 \cup \ldots \cup S_R\right)$?
3. Let $m > 2 \ln n$, and $n \to \infty$. Show that the expected number of training examples in S that appear in at least one S_i is more than $n - 1$.

Problem 4. (30 points). Please see attached notebook for details.