Homework 4

Due: Thursday, February 20, 2020 at 12:00pm **(Noon)**

Written Assignment

Problem 1: Bias-Complexity Tradeoff

(10 points)

a. Explain why the No Free Lunch Theorem requires us to introduce bias into our machine learning models.

b. While training a machine learning model, Steve notices that the model's error remains large even after many training iterations. Steve determines that this error results from overfitting, and he proposes changing the size of the hypothesis class to reduce the error. Will changing the size of the hypothesis class successfully reduce the error? If so, how should the size be changed? If not, why not? Explain your reasoning.

c. After tinkering with his model, Steve eventually manages to prevent his model from overfitting. However, Steve notices that his model's error remains large because his model is now underfitting the data. To decrease underfitting, Steve decides to collect more training data. Will collecting more training data successfully reduce the error? Explain your reasoning.

Problem 2: Assumptions & Guarantees

(10 points)

a. In lecture, we made several assumptions before proving that an ERM solution is PAC learnable. What are these assumptions?

b. The HTAs want to produce a model to predict the final grade of a student based off of their midterm score. During the sampling process, the HTAs select the next example based on how similar it was to the previous example. Why does the ERM solution not perform well at test time? Explain in terms of the assumptions above.

c. Steve wants to produce a model to predict the weight of an animal based on its characteristics (fur, number of legs, and gender). He trains his model on a dataset of **baby** animals and then uses it to make predictions about **adult** animals. Why does the ERM solution not perform well at test time?

Problem 3: PAC Learning of Partial Orderings

(20 points)

Background Information: A *partial ordering* (denoted $\preceq$) is a binary relation over a set S that satisfies the following properties:

- *Reflexivity*: $\forall s \in S, \quad s \preceq s$
- *Antisymmetry*: $\forall s, t \in S, \quad s \preceq t, t \preceq s \implies s = t$
- *Transitivity*: $\forall r, s, t \in S, \quad r \preceq s, s \preceq t \implies r \preceq t$

A *total ordering* differs from a partial ordering in providing an additional guarantee that all elements within S can be compared in the relation. As an example, the "$\leq$" relation is a total ordering on the real numbers. Any two numbers a, b may be compared such that either $a \leq b$ or $b \leq a$. Additionally, the "$\leq$" relation satisfies the conditions of reflexivity, antisymmetry and transitivity stated above. *Note: All total orderings are also partial orderings.*

One example of a partial ordering that isn't a total ordering is the "$\subseteq$" relation. "$\subseteq$" satisfies reflexivity, transitivity and antisymmetry. However, given two sets A and B, there is no guarantee that either $A \subseteq B$ or $B \subseteq$ A. Now, consider a classification problem defined as follows:

Input Space: The set of all possible n-element permutations.
Hypothesis Space: The set of all hypotheses, where each hypothesis maps all n-element permutations to either 0 or 1. Each hypothesis is equivalent to a partial ordering.

Suppose that there is a hidden partial ordering that we do not have access to. We do have a sample of permutations, and a label for each sample indicating whether the permutation is consistent with the true hidden partial ordering. We are trying to learn the rule from this sample.

a. Show that $|H|$ is between $n!$ and $3^{(n^2)}$. That is, find the number of possible valid partial orders generated from n elements, and show that it is bounded by $n!$ and $3^{(n^2)}$. *Hint: Each $h \in H$ may be represented as a Directed Acyclic Graph (DAG). The nodes of the DAG can be thought of as adjacent elements in our input space. Think about how the relations between nodes in a DAG are similar to the relations between elements within a partial ordering. This question may also therefore be approached by thinking about the number of valid DAGs with n nodes.*

b. Show that the amount of data needed to ensure the ERM solution is PAC is polynomial in n.

c. Describe an algorithm you could use to efficiently find an ERM solution for a given sample of observations. Be sure to prove the correctness of the algorithm and analyze its runtime. *Hint: The true hypothesis is within our hypothesis space. Consider the order of each pair of elements in every permutation with label 1.*

Grading Breakdown

The grading breakdown for the assignment is as follows:

Problem 1	25%
Problem 2	25%
Problem 3	50%
Total	100%

Handing In

You will turn in your final handin via Gradescope, as detailed in the email sent to the course. If you have questions on how to set up or use Gradescope, ask on Piazza! For this assignment, you should have written answers for Questions 1, 2, and 3.

Anonymous Grading

You need to be graded anonymously, so do not write your name anywhere on your handin.

Obligatory Note on Academic Integrity

Plagiarism — don't do it.

As outlined in the Brown Academic Code, attempting to pass off another's work as your own can result in failing the assignment, failing this course, or even dismissal or expulsion from Brown. More than that, you will be missing out on the goal of your education, which is the cultivation of your own mind, thoughts, and abilities. Please review this course's collaboration policy and if you have any questions, please contact a member of the course staff.