Homework 5 Written

10-605/10-805: Machine Learning with Large Datasets

Due Wednesday, November 20th at 11:59 PM Eastern Time

Submit your solutions via Gradescope, **with your solution to each subproblem on a separate page**, i.e., following the template below.

IMPORTANT: Please use the provided template. If you do not follow the template (i.e. there is some misalignment), your assignment may not be graded correctly by our AI assisted grader and there will be a **2% penalty** (e.g., if the homework is out of 100 points, 2 points will be deducted from your final score).

Note that Homework 5 consists of three parts: this written assignment and a two programming assignment, each with their own PDF handout. Remember to fill out the collaboration section found at the end of this homework as per the course policy.

All students are required to complete **all** sections of the homework. Your homework score will be calculated as a percentage over the maximum points you can earn, which is 100. The grading breakdown is as follows:

1. Written Section *[40 points]*
2. Programming Part 1 *[50 points]*
3. Programming Part 2 *[10 points]*

1 Written Section [40 Points]

1.1 Curse of Dimensionality (20 pts)

We define the ϵ-cover of a search space S as a subset $E \subset S$:

$$E = \{x \in \mathbb{R}^n \mid \forall y \in S, \exists x \; s.t. \; \|y - x\|_\infty \leq \epsilon\}$$

In words, every coordinate of every point in our search space S is within ϵ of a point in our set E. The ϵ-covering number is the size of the smallest such set that provides an ϵ-cover of S.

$$N(\epsilon, S) = \min_E \{|E| : E \text{ is an } \epsilon\text{-cover of } S\}$$

Let's work through a concrete example to make the concept clearer: let's say we want to tune the learning rate for gradient descent on a logistic regression model by considering learning rates in the range of $[1, 2]$. We want to have $\epsilon = .05$-coverage of our search space $S = [1, 2]$. Then $E = \{1.05, 1.15, 1.25, 1.35, 1.45, 1.55, 1.65, 1.75, 1.85, 1.95\}$. We see that any point in $[1, 2]$ is within 0.05 of a point in E. So E provides ϵ-coverage of S. The ϵ-covering number is 10. The questions below will ask you to generalize this idea.

(a) *[6 points]* Find the size of a set needed to provide ϵ-coverage (ϵ-covering number) of $S = [0, 1] \times [0, 2]$, the Cartesian product of two intervals. Note that we want this to hold for a generic ϵ.

(b) *[6 points]* Find the ϵ-covering number of $S = [a_1, b_1] \times \ldots \times [a_d, b_d]$, the Cartesian product of d arbitrary closed intervals. Specifically, write down an expression $N(\epsilon, S)$ as a function of a_i, b_i, and ϵ.

(c) *[8 points]* With respect to your answer in 1.2 (b), intuitively, how is the covering number related to the volume of S? Moreover, on what order does the covering number grow with respect to the dimension d? What does this say about the volume of S as a function of dimensionality?

1.2 Grid versus Random Search (20 pts)

There are two basic strategies commonly employed in hyperparameter search: grid search and random search. Grid search is defined as choosing an independent set of values to try for each hyperparameter and the configurations are the Cartesian product of these sets. Random search chooses a random value for each hyperparameter at each configuration. Imagine we are in the (extreme) case where we have d hyperparameters all in the range $[0, 1]$, but only *one* (h_1) of them has any impact on the model, while the rest have *no* effect on model performance. For simplicity, assume that for a fixed value of h_1 that our training procedure will return the exact same trained model regardless of the values of the other hyperparameters.

(a) *[8 points]* Imagine that we consider q hyperparameter configurations, which are enough to provide ϵ-coverage for grid search. How many distinct models will this training procedure produce? What is this as a fraction of the total number of configurations?

(b) *[5 points]* Alternatively, if we consider q random configurations as part of random search, then what percent of the configurations will cause a change to the model? Why?

(c) *[7 points]* Now let's change our point of view: instead of desiring ϵ-coverage, we have a fixed budget of $B \ll |E|$ configurations to try (where $|E|$ is an epsilon cover of S). Which search strategy—grid or random search, should we employ? Why?

2 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

2. (a) Did you give any help whatsoever to anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

3. (a) Did you find or come across code that implements any part of this assignment?

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).