CEE 554: Data Mining in Transportation

Homework Assignment 2
Due date: 02/01, 11:59pm on Canvas

- **You can either type your answers or handwrite them and then scan the document. If you choose to handwrite, ensure your writing is legible and the scan quality is clear. Illegible submissions will not be awarded points.**
- **For the multiple choice problems, make sure to show your work.**
- **Add Appendix A at the end of each homework assignment, specifying whether you used Generative AI in developing your solution, the Generative AI engine you used, and the command you used for each problem.**
- **Add Appendix B at the end of the assignment, including the code you used for each problem.**

1. Use the datasets from your homework 1.
   (a) We saw in class that linear regression can be used as a classification tool. Use linear regression alone to classify points in dataset A. What is the resulting W? What is the in-sample error? Draw the final hypothesis. **(15 points)**
   (b) Instead of generating a random weight vector to initialize PLA (as you did in homework 1), use the linear regression line from 1(a) as the initial weight vector for PLA. Report the number of iterations it takes for PLA to converge using (i) linear regression as the initial weight, and (ii) a random weight vector as the initial weight (you can use your results from HW #1 for (ii)). **(20 points)**

2. In this problem, we again apply Linear Regression for classification. Consider the target function:

$$f(x_1, x_2) = \text{sign}(x_1^2 + x_2^2 - 0.6)$$

Generate a training set of $N = 1000$ points on $\mathcal{X} = [-1,1]\times[-1,1]$ with uniform probability of picking each $x \in \mathcal{X}$. Generate simulated noise by flipping the sign of the output f in a randomly selected 10% subset of the generated training set.

(a) Carry out linear regression without transformation, i.e., with input vector: $(1, x_1, x_2)$, to find the weight W. What is the classification in-sample error, E_{in}? (Run the experiment 1000 times and take the average E_{in} to reduce variation in your results). **(15 points)**

(b) Now, transform the N=1000 training data into the following nonlinear feature vector:

$$(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

.

Find the vector W that corresponds to the solution of linear regression. Which of the following hypotheses is closest to the one you find? Closest here means agrees the most with your hypothesis (Average over a few runs to make sure your answer is stable.) **(20 points)**

i. $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 1.5x_2^2)$
ii. $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 15x_2^2)$
iii. $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 15x_1^2 + 1.5x_2^2)$
iv. $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 0.05x_2^2)$
v. $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 0.15x_1^2 + 0.15x_2^2)$

3. Run a computer simulation to flip 1,000 virtual fair coins. Flip each coin independently 10 times. Focus on 3 coins as follows: c_1 is the first coin flipped, c_{rand} is a coin chosen randomly from the 1,000, and c_{min} is the coin which had the minimum frequency of heads (pick the earlier one in case of a tie). Let v_1, v_{rand}, and v_{min} be the fraction of heads obtained for the 3 respective coins out of the 10 tosses. Run the experiment 100,000 times in order to get a full distribution of v_1, v_{rand}, and v_{min} (note that c_{rand} and c_{min} will change from run to run).

(a) The average value of v_{min} is closest to **(15 points)**

i. 0
ii. 0.01
iii. 0.1
iv. 0.5
v. 0.67

(b) Which coin(s) has a distribution of v that satisfies the (single-bin) Hoeffding Inequality, and why? **(15 points)**

i. c_1 only

ii. c_{rand} only

iii. $c_{\min}$ only

iv. c_1 and c_{rand}

v. $c_{\min}$ and c_{rand}