CS482 Midterm Exam - Spring 2022

Please write your name and ID in all your submitted pages.

Please do not write on these pages - use your own and submit those only.

Problem Set 1 (take at home - 35 points)

You joined the data science team of a electricity provider and you are told to develop a model that can detect pricing jumps in the electricity market.

Using Google Colab. **Numpy and/or scipy are the only numerical libraries you can use**:

A. (10 points) Dataset

Develop a synthetic dataset that captures what happened during geopolitical events that suddenly occurred at time t_s. Each event reduced the supply and this resulted in a sudden jump in the price x the providers had to pay. The synthetic dataset is that of price vs. time. The dataset must be compliant to the following spec.

1. The total duration is of the dataset is $T = 365$ time units.
2. The price jump $x_e = 2 \times x_s$.
3. The price jump is a step function that at time t_s causes the price to jump from x_s to x_e and at time t_e causes the price to go back to x_s
4. Gaussian noise with zero mean and variance equal to 20% of p_s is added to all prices in the dataset.
5. Price jump events have duration of $t_e - t_s = 10$ time units and there are $A = 2$ such events in the dataset located randomly across the dataset duration.
6. x_s is 50 dollars per MWh.

Plot the dataset of x vs t.

B. (25 points) Model

Your manager wants you to implement the following approach that will predict all price jump events.

1. Randomly sample the dataset you synthesized in step A, creating $N < T$ samples and feed the selected samples to a binary tree.
2. Define a hyperparameter D_{max} that represents the max depth of the tree.
3. Define a variable d that represent the current depth of the tree.

4. In each node of the tree, randomly choose a threshold between the min and max price values in the input to the tree samples to split the feature x.
5. Continue the splits until you have only one sample at the leaf nodes or you have reached the depth D_{max}.
6. Create a list that contains K trees as developed in the above steps 1-5.

7. Write a function `evaluate_forest` that will return the number of edges (path length) that each each price x trasverses from the root note to the leaf node across the trees of the forest.
8. Plot the average path length that each sample in the dataset. What can you notice about the average path lengths of prices that correspond to price jumps events vs the rest ? Select a threshold that can best detect the price jump events.
9. Tune N, K and D_{max} to get a good detection performance of the price jump events.

Submit your **notebook URL** after sharing it with the grader and the professor to the Canvas URL that will be given to you at the date of the exam. No PDFs will be accepted as an answer to Problem Set 1 - all explanations must be inserted / typed in the notebook itself.

This is a exam question. If you are found you communicated or collaborated with anyone to submit your answer, all parties involved will receive a grade of 0 for the whole midterm exam and referred to the Dean of Students.