CS 203: Software Tools & Techniques for AI

IIT Gandhinagar

Sem-II - 2024-25

Lab Assignment 10

Total: 10 Marks
Submission deadline: Wednesday, 16/04/2025 11:59:59 PM
Submission guidelines:

1. Implement the code in the jupyter notebook and share it during submission.
2. Add all screenshots of the results in a pdf file and share.
3. The deadline will not be extended at any cost.

Note: Submitting this assignment solution confirms that you will follow the IITGN's honor code. We shall strictly penalize the submissions containing plagiarized text/code.

Objective

Using real-world data, this assignment will introduce students to key concepts in A/B testing and Covariate Shift Detection. Students will perform hypothesis testing using the scipy library and identify distributional shifts in datasets using classification-based techniques.

Dataset

Get the Ad click prediction and Air_quality dataset.

Part 1: A/B Testing using Ad Click Prediction

1. Load the dataset into a pandas DataFrame.
2. Perform necessary data cleaning and preprocessing: **[10 points]**
    a. Handle missing values
    b. Convert categorical columns (e.g., gender, ad_position)

3. Split the dataset into two groups: **[10 points]**
    a. Group A: Users with ad_position = 0 (**Top**)
    b. Group B: Users with ad_position = 1 (**Bottom**)
4. Use the statsmodel's proportions_ztest function to perform an independent two-sample z-test between Group A and Group B.
5. Print the following:
    a. The z-score **[10 points]**
    b. The p-value **[10 points]**
6. Interpret the result: Is there a statistically significant difference in click-through rates between the two groups? Justify your answer. **[10 points]**

Part 2: Covariate Shift Detection Using Air Quality Data

1. You are provided with 3 datasets via this Google Drive link:
    a. train.csv
    b. test1.csv
    c. test2.csv
2. Load all three datasets using pandas. **[10 points]**
3. For each test dataset (test1.csv and test2.csv), compare it with train.csv using the **Kolmogorov–Smirnov test** (scipy.stats.ks_2samp).
   Perform the KS test on the **NO2(GT)** column to identify whether there are any distributional differences. **[20 points]**
4. Report the KS statistic and p-value for each feature. **[10 points]**
5. Determine which of the two test datasets (test1.csv or test2.csv) exhibits a covariate shift relative to the training dataset (train.csv). Use the results of the Kolmogorov–Smirnov test to support your answer. **[10 points]**

*****Best of Luck*****