Task 4 Report: Outlier Detection for Houston Weather Dataset

1. Introduction

In this task, a distance-based outlier detection technique was applied to the Houston Weather dataset (HW2021), which reports daily weather attributes including minimum temperature, rainfall, wind speed, humidity, and cloud cover for the year 2021. The goal was to identify unusual weather days using a custom distance function and calculate outlier scores (OLS) for different numbers of neighbors (k-values). These OLS values measure how dissimilar a day is compared to other days in the dataset.

2. Dataset Description

The dataset contains the following attributes:

- **min_temp**: Minimum temperature (°F) observed at 3 PM.
- **rainfall**: Amount of rainfall (inches) recorded for the day.
- **windspeed**: Wind speed (mph) at 3 PM.
- **humidity**: Humidity (%) recorded at 3 PM.
- **cloudcover**: Categorical variable with 17 unique categories describing cloud cover conditions.

3. Methodology

The outlier detection process involved the following steps:

1. **Preprocessing**:
   - Continuous attributes (min_temp, rainfall, windspeed, and humidity) were normalized using MinMaxScaler to ensure they were on the same scale.
   - Missing values in the `cloudcover` attribute were filled with the most frequent value (`mode`).
   - Ordinal values were assigned to the cloud cover categories based on their relative intensity, with "Fair" having the lowest value (1) and "T-Storm" having the highest value (17).
2. **Distance Calculation**:
   - A custom distance function was used, which combines the Euclidean distance for continuous attributes and the absolute difference for the ordinal `cloudcover` attribute.
   - The distance function was weighted, giving more significance to attributes like rainfall and wind speed, as they may better reflect unusual weather conditions.
3. **Outlier Score Calculation**:
   - The OLS value for each day was calculated as the mean distance to its `k` nearest neighbors (with `k=5`, `k=50`, and `k=100` being used as hyperparameters).
   - Higher OLS values indicate a greater likelihood that the day is an outlier.

4. Results

a. Top 4 Outliers

The following table shows the top 4 outliers for `k=50`, along with their OLS values:

Date	Min Temp	Rainfall	Wind Speed	Humidity	Cloud Cover	OLS
8/18/2021	0.00	0.10	0.48	0.81	17 (T-Storm)	13.00
8/6/2021	0.94	0.00	0.24	0.84	17 (T-Storm)	12.62
7/31/2021	0.96	0.00	0.34	0.68	16 (Thunder)	11.52
4/14/2021	0.00	0.24	0.24	0.96	15 (Heavy T-Storm)	11.11

These days were characterized by extreme weather events, including thunderstorms and heavy storms. Their high OLS values suggest that these days were significantly different from the rest of the year.

b. Most Normal Days

The following table shows the most normal days (lowest OLS values) for `k=50`:

Date	Min Temp	Rainfall	Wind Speed	Humidity	Cloud Cover	OLS
12/29/2021	0.87	0.00	0.41	0.53	5 (Mostly Cloudy)	2.47
12/25/2021	0.78	0.00	0.45	0.53	5 (Mostly Cloudy)	2.46

These days had typical weather conditions, such as moderate temperatures, no rainfall, and mostly cloudy skies, leading to low OLS values.

5. Discussion

- **Outliers**: Days with extreme weather conditions, particularly thunderstorms and heavy storms, were identified as outliers. This is expected because such events are rare compared to regular weather patterns in Houston.
- **Normal Days**: Days with consistent, moderate weather (e.g., mostly cloudy skies, moderate wind speeds, and no rainfall) were identified as the most normal. These days were very similar to many other days in the dataset, resulting in low OLS values.
- **Hyperparameter `k`**: The choice of `k` affected the detection of outliers. Smaller values of `k` (e.g., `k=5`) focus more on local neighborhood distances, while larger values (e.g., `k=100`) smooth out noise by considering a broader context. In this task, the results were consistent across different `k` values, with the same weather events being flagged as outliers.

6. Conclusion

The distance-based outlier detection technique successfully identified unusual weather days in the HW2021 dataset. Days with severe weather events, such as thunderstorms and heavy rain, were detected as outliers, while days with moderate, typical weather were classified as most normal. This method provides a simple yet effective way to analyze weather data and detect anomalies.