University of Michigan

College of Engineering

Department of Climate and Space Sciences and Engineering

Fall Semester 2024

CLIMATE 405

Machine Learning for Earth and Environmental Sciences

Homework 1

Due by 11:59 PM on Friday (09/06/2024)

Dr. Mohammed Ombadi

General Guidelines:

- Please type your answers to the questions in this document and upload it in a PDF or Word document.
- Include a copy of the code you developed to answer the questions.
- Ensure your code is annotated for clarity.

1. In Lecture 1, we discussed the Kurtosis, the 4th moment of a probability distribution, and examined the difference between Kurtosis and Excess Kurtosis. Using the Python function *scipy.stats.kurtosis* from the Lecture 1 slides, follow these steps:

    - Generate samples from a standard normal distribution (mean = 0, standard deviation = 1) with the following sample sizes: 100, 1000, and 10,000 points, what are the values of Kurtosis for each sample size? Plot the histogram associated with each sample size.

    - Observe how the Kurtosis value changes as the sample size increases. Try increasing the sample size to 100,000 data points. What is your main observation? Is this consistent with our theoretical understanding?

    - Determine whether the obtained values correspond to Kurtosis or Excess Kurtosis.

    - If you are using a programming language other than Python, utilize the built-in functions available in that language and address the questions above.

2. The following file "*Ann-Arbor-Temp.csv*", contains average daily temperature data (in units of °C) for the city of Ann Arbor for the period (01/01/2023 – 12/31/2023).

    - Use a persistence model to predict temperature at 1 day lead time. *NOTE: A persistence model assumes that tomorrow's temperature will be the same as today's temperature. That is, T(t+1) = T(t).* Plot a line plot (predictions) and a scatter plot (observations).

    - What are the MSE, RMSE, MAE, Bias, Relative Bias and Correlation Coefficient for the predictions of the persistence model?

    - Is the persistence model consistently underestimating or overestimating? Why?

    - Using a temperature threshold of 15 °C, transform both the observations and the predictions to categorical variables with values of Warm ($T \geq 15$ °C) and Cold ($T < 15$ °C).

- Use the confusion matrix to calculate the accuracy of the persistence model. *NOTE: The accuracy is the proportion of total correctly classified instances out of all instances.*

- *What do you think about the persistence model? Is it a good model? Do you think we can build a machine learning model that will beat the persistence model?*

3. Using the temperature data in problem.2 and the same threshold of 15 °C:

    - Estimate the transition probabilities associated with a Marko Chain to model the data (for example, estimate the probability of transitioning from Warm to Cold, Cold to Warm, and so on). Show your results in a transition probability matrix.

    - Using the transition probabilities estimated above, start with the first observation in the year and generate a sequence of 30 values. Repeat this step 5 times. Are all the simulations similar to one another? Are they identical to observations? Discuss the reasons.

    - Estimate the transition probabilities using two parts of the data, (March to August) and (August to December). Compare the transition probabilities for each case and summarize your observations?

4. Which one of the following metrics is more robust to outliers, and why? If they are equally robust, please indicate so.

    - Mean vs Median vs Mode.
    - Bias vs Relative Bias.
    - Root Mean Squared Error (RMSE) vs Mean Absolute Error (MAE).

5. **Extra Credit:** A typical example of a stochastic system is Random Walk. The simplest type of a random walk is a one-dimensional walk over the real line, starting at 0. At each step, you can move +1 or -1 steps with equal probabilities.

    - Write a function to model this Random Walk.
    - Generate a sequence of n= 1,000 values.
    - Plot the histogram of the data. Describe the properties of this distribution.