| | |
|---|---|
| **Assigned: 16 Dec 2024**<br>**Due:**<br>**10 Jan 2025** | **KRR HW No.2** |

## Part I (7p)
----------------

The most important use of the Expectation-Maximization (EM) algorithm is in handling datasets with incomplete or missing data. In this homework will be applying EM algorithm for imputing missing values in a dataset.

**1.** Consider a dataset with two variables X and Y, where some values of Y are missing. Explain how the EM algorithm can be used to estimate the mean and variance of Y and the correlation between X and Y. Why is the log-likelihood expected to increase after each iteration?

**2.** Implement the EM algorithm to handle missing data in a dataset. Your implementation should:

- Accept a dataset with missing values – generate a synthetic dataset then eliminate some values to obtain a set with missing values;
- Randomly initialize the missing values;
- Iteratively perform the E-step (estimating the missing values) and M-step (recomputing the parameters);
- Output the imputed dataset and the final parameter estimates (e.g., means, variances, correlations).

Test your implementation on the synthetic dataset:

- Visualize the original data with missing values and the imputed dataset;
- Compare your imputed values to the true values;
- Plot the log-likelihood values at each iteration to demonstrate convergence.

**3.** Choose a real-world dataset with missing values. e.g.,

- Air Quality Dataset (UCI Machine Learning Repository) - includes sensor data with missing entries;
- Housing Prices Dataset (Kaggle) - contains features like house prices, lot size, and number of rooms, with missing values in some columns.

You are free to choose another real-world dataset if you wish.

Preprocess the data as necessary (e.g., handle categorical variables, normalize numerical features).

- Use your EM algorithm to impute the missing values;
- Visualize the dataset before and after imputation;
- Analyze how the imputed values compare with domain knowledge;
- Evaluate the overall impact of imputation on the dataset’s structure (e.g., correlation matrix).

You will submit a PDF file combining your answer for 1-3 with the obtained plots and data set visualization and a separate file with your program.

## Part II (3p)

----------------

Critically analyze different embedding techniques used in machine learning across various types of data (e.g., numerical, textual, and categorical) in an essay between 1,000-1,500 words.

The essay must have the following structure:

1. Introduction to Embeddings
   - Discuss the general concept of embedding high-dimensional data into a lower-dimensional space, highlighting why dimensionality reduction or feature transformation is useful
   - Provide examples of embedding techniques across different data types
2. Analysis of Different Types of Embeddings
   Compare and contrast at least three embedding techniques across different data types. For each technique, explain:
   - How it works
   - Type of data
   - Applications: real-world applications where the embedding is commonly used
   - Advantages and limitations
3. Based on your understanding, provide a real-world scenario where you compare the performance of at least two embedding techniques.
4. Summarize your findings and reflect on the following:
   - How does the choice of embedding technique affect the downstream tasks in machine learning?
   - Are embeddings universal, or do they need to be customized for each domain?

Submit as a PDF document, include at least 3 references.

Therefore, you must submit 2 PDF files and the program.

**Nota Bene**

- You should explicitly indicate whether you used any generative model in providing your answers or writing your code. In case you did so, you should clearly explain which is your contribution. Be aware that answers will be checked with adequate tools.
- You will have to defend your homework, both parts, during the last laboratory, and answer questions both from the written part and from the code.