ECE/CS/ME 539 – Fall 2024 — Activity 6

1. Imputing Missing Data with Various Methods

In many real-world datasets, some values might be missing. It is often necessary to fill in these missing values (a process known as "imputation") to work with the dataset effectively. Below, we explore three methods of imputing missing data: Mean Imputation, Median Imputation, and k-Nearest Neighbors (kNN) Imputation.

Mean Imputation

The mean imputation method fills in missing values with the mean of the non-missing values in the same feature (column). For example, if a column has the values $[1, 2, \text{NaN}, 4]$, the mean of the non-missing values $(1 + 2 + 4)/3 = 2.33$ will replace the missing value.

Median Imputation

The median imputation method fills in missing values with the median of the non-missing values in the same feature. The median is the middle value when the values are sorted. If a column has the values $[1, 2, \text{NaN}, 4]$, the median of the non-missing values 2 will replace the missing value.

k-Nearest Neighbors (kNN) Imputer Algorithm

The kNN imputer algorithm works by identifying the k-nearest neighbors of a data point with missing values, based on the available (non-missing) features. The missing values are then imputed based on the average (or median) of the corresponding feature values from the k-nearest neighbors.

1. For each data point with missing values, identify the k-nearest neighbors in the dataset based on the available features.
2. Compute the average (or median) of the corresponding feature values from the k-nearest neighbors.
3. Replace the missing value with this computed average (or median).

This method assumes that the missing values can be reasonably estimated by similar (neighboring) points in the dataset.

Now, consider the following dataset with missing values:

$$X = \begin{bmatrix} 1 & 2 & \text{NaN} \\ 3 & 4 & 3 \\ \text{NaN} & 6 & 5 \\ 8 & 8 & 7 \end{bmatrix}$$

(a) **Mean Imputation:** Impute the missing values in the dataset using the mean of the non-missing values for each column.

(b) **Median Imputation:** Impute the missing values in the dataset using the median of the non-missing values for each column.

(c) **kNN Imputation:** Using $k = 2$ neighbors, impute the missing values using the kNN imputation method.

(d) **Comparison:** Compare the results of the mean, median, and kNN imputations. Discuss the advantages and disadvantages of each method. In which scenarios would each method be preferable?