ECE/CS/ME 539 – Fall 2024 — Activity Solution6

Activity 6: Imputing Missing Data with Various Methods

In this activity, we will explore three methods of imputing missing data: Mean Imputation, Median Imputation, and k-Nearest Neighbors (kNN) Imputation.

Dataset

The given dataset with missing values is:

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

(a) Mean Imputation

To perform mean imputation, we replace the missing values with the mean of the non-missing values in the corresponding columns.

- **Column 1 (1, 3, NaN, 8):**
$$\text{Mean} = \frac{1+3+8}{3} = 4$$
- **Column 2 (2, 4, 6, 8):**
$$\text{Mean} = \frac{2+4+6+8}{4} = 5$$
- **Column 3 (NaN, 3, 5, 7):**
$$\text{Mean} = \frac{3+5+7}{3} = 5$$

The dataset after mean imputation is:

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

(b) Median Imputation

To perform median imputation, we replace the missing values with the median of the non-missing values in the corresponding columns.

- **Column 1 (1, 3, NaN, 8):**
$$\text{Median} = 3$$
- **Column 2 (2, 4, 6, 8):**
$$\text{Median} = 5$$
- **Column 3 (NaN, 3, 5, 7):**
$$\text{Median} = 5$$

The dataset after median imputation is:

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

(c) kNN Imputation

Using $k = 2$ neighbors, the kNN imputation method identifies the 2-nearest neighbors for each data point with missing values and uses their values to estimate the missing ones.

- For Column 1 (missing value at $X[3, 1]$): The nearest neighbors with available data are: - Data point 1: (8, 8, 7) - Data point 2: (3, 4, 3)
$$\text{Estimated value} = 5.5$$
- For Column 3 (missing value at $X[0, 3]$): Nearest neighbors with available data: - Data point 1: (3, 4, 3) - Data point 3: (8, 8, 7)
$$\text{Estimated value} = 5$$

(d) Comparison of Imputations

- **Mean Imputation:** Simple and quick to compute, but can be influenced heavily by outliers. Best used when missing data is minimal or randomly distributed.
- **Median Imputation:** Less sensitive to outliers but might not give a representative value if data is skewed.
- **kNN Imputation:** Can be better by adapting to local structure, but requires more computation. Ideal when missing data is not randomly missing.