Indian Institute of Technology Dharwad

CS201L: Artificial Intelligence Laboratory

Lab 4: Data Preprocessing: Outlier Detection, Correlation Analysis (Pearson and Spearman Correlation), Attribute Normalization and PCA-Practice Questions

1 Practice Problem Statement

A real-world medical dataset containing sales records is provided in a CSV file `sales data.csv`. Write a Python program to perform the following tasks on the following attributes: `Cost, Revenue`

1.1 Outlier Detection

1. Read the dataset into a pandas dataframe. Plot the boxplot of the attributes. Observe the number of outliers present in each attribute and list their values. Outliers are defined as the data values that do not satisfy the following condition: $(Q_1 - 1.5 \times IQR) < x < (Q_3 + 1.5 \times IQR)$ where $IQR = Q_3 - Q_1$, Q_1= Lower Quartile and Q_3= Upper Quartile.

1.2 Correlation

Degree of Correlation	**Positive Correlation**	**Negative Correlation**
Perfect Correlation	+1.0	−1.0
Very Strong Correlation	(0.5, 1.0)	(−1.0, −0.5)
Strong Correlation	(0.3, 0.5]	[−0.5, −0.3)
Moderate Correlation	(0.1, 0.3]	[−0.3, −0.1)
Weak Correlation	(0.0, 0.1]	[−0.1, 0.0)
Zero / No Correlation (Uncorrelated)	0.0	0.0

Table 1: Degree of Correlation

1. Load the outlier-corrected dataset `sales_data.csv` and compute the Pearson and Spearman correlation coefficients for every pair of attributes in the dataset.

1.3 Normalization and Standardization

1. Load the dataset and observe the range of the values in each attribute. Find the minimum and maximum values in each attribute.

2. Load the dataset and perform the Min-Max normalization of this data using **MinMaxScaler** to have the range of values between 0-1. After performing the operation save this file as `sales_data_normalisation.csv`.

3. Load the dataset and find the mean and standard deviation of the attributes. Standardize each attribute using the **StandardScaler** command. After performing the operation save this file as `sales_data_standardisation.csv`

For **MinMaxScaler:** `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`
For **StandardScaler:** `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

1.4 Principal Component Analysis (PCA)

1. Load the dataset and perform Principal Component Analysis (PCA) on the specified attributes. Fit the PCA model to the data and transform it accordingly. Analyze the resulting eigenvalues and eigenvectors by completing the following tasks:

   (a) Compute and display the eigenvalues and their corresponding eigenvectors of the covariance matrix.

   (b) Plot the eigenvalues in descending order.

For **Covariance:** `https://numpy.org/doc/2.1/reference/generated/numpy.cov.html`
For **Eigen values and vectors:** `https://numpy.org/devdocs/reference/generated/numpy.linalg.eig.html`
For **cumsum:** `https://numpy.org/devdocs/reference/generated/numpy.cumsum.html`
For **PCA:** `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`