

|  |
| --- |
| Introduction to Machine Learning |
| Assignment 4 |
| 7818 Mohamed Hussein  7721 Alaa Elkhouly |

**Assignment 4 Report**

**1. Principal Component Analysis (PCA)**

**1.1 What is PCA?**

Principal Component Analysis is a statistical technique used to reduce the dimensionality of a dataset while retaining as much variability as possible. High-dimensional data often contains redundant or irrelevant information. PCA simplifies the dataset by projecting it into a lower-dimensional space.

**1.2 How PCA Works**

PCA follows these steps:

1. **Standardization**: The features of the dataset are standardized to have a mean of 0 and a standard deviation of 1. This ensures that all features contribute equally to the analysis.
2. **Covariance Matrix Calculation**: A covariance matrix is computed to identify relationships between features.
3. **Eigenvalues and Eigenvectors**: Eigenvalues and eigenvectors of the covariance matrix are calculated. Eigenvalues represent the amount of variance captured, and eigenvectors define the direction of principal components.
4. **Sorting Principal Components**: Principal components are sorted based on their eigenvalues in descending order. The top components are selected based on the desired dimensionality.
5. **Projection**: The original data is projected onto the selected principal components to obtain the reduced dataset.

**1.3 Benefits of PCA**

* **Dimensionality Reduction**: Reduces the complexity of datasets, making them easier to visualize and analyze.
* **Noise Reduction**: Eliminates less significant variations that may be caused by noise.
* **Improved Performance**: Enhances the performance of machine learning algorithms by reducing redundant information.

**1.4 How PCA Affects Clustering**

**improves**

1. **Dimensionality Reduction**: By reducing the number of dimensions, PCA removes irrelevant features that may confuse clustering algorithms like K-Means.
2. **Improved Separation**: PCA can transform data into a space where clusters are more clearly separated, enhancing the accuracy of clustering.
3. **Reduced Noise**: Noise in the data can negatively impact clustering. PCA reduces noise by discarding less significant principal components.

**1.5 Visualizing Clusters After PCA**

Reducing data to 2 or 3 dimensions makes it possible to visualize the clustering results effectively. This aids in understanding the structure and relationships within the data.

**2. K-Means Clustering**

**2.1 What is K-Means?**

K-Means is an unsupervised machine learning algorithm used for clustering data into groups or clusters. It partitions data into k clusters such that each data point belongs to the cluster with the nearest mean (centroid).

**2.2 How K-Means Works**

The K-Means algorithm follows these steps:

1. **Initialization**: Select k initial centroids randomly or using methods like k-means+.
2. **Assignment**: Assign each data point to the nearest centroid based on a distance metric (usually Euclidean distance).
3. **Update**: Recalculate the centroids as the mean of all points assigned to each cluster.
4. **Iteration**: Repeat the assignment and update steps until centroids stabilize or a predefined number of iterations is reached.

**2.3 Challenges and Limitations**

* **Number of Clusters (k)**: Requires the user to specify k beforehand.
* **Sensitivity to Initialization**: Different initial centroids can lead to different results.
* **Outliers**: Sensitive to outliers, which can distort cluster boundaries.

**3. Clustering Mechanism**

**3.1 Clustering with K-Means**

K-Means groups data points based on similarity, measured as the distance from centroids. The algorithm aims to minimize the variance within clusters while maximizing the variance between clusters.

**3.2 The Role of Centroids**

Centroids act as representatives for each cluster. The assignment step groups points closest to each centroid, and the update step refines these centroids to better represent their clusters.

**3.3 Measuring Clustering Quality**

The quality of clustering can be evaluated using metrics like the **Sum of Squared Errors (SSE)**.

**4. Sum of Squared Errors (SSE)**

**4.1 Definition**

SSE measures the total squared distance between data points and their assigned cluster centroids. It is calculated as:

SSE=∑i=1n∑j=1k∣∣xi−cj∣∣2SSE = \sum\_{i=1}^{n} \sum\_{j=1}^{k} ||x\_i - c\_j||^2

Where:

* x\_i is a data point.
* c\_j is the centroid of the cluster to which x\_i belongs.

**4.2 Significance of SSE**

* **Lower SSE**: Indicates tighter clusters with less variance.
* **Elbow Method**: Used to determine the optimal number of clusters by plotting SSE against k. The "elbow" point on the graph suggests the best k.

**6. Experiments and conclusions in the Assignment**

**6.1 First Experiment: K-Means Without PCA**

In this experiment, K-Means is applied directly to the original dataset. The results are evaluated based on SSE and visualizations.

**6.2 Second Experiment: PCA Followed by K-Means**

In this experiment, PCA is applied to reduce the dataset’s dimensions before clustering. By experimenting with different numbers of principal components, we assess how dimensionality affects clustering performance.

**6.3 Comparison of Results, 7. Observations and Insights**

* **SSE Comparison**: Compare SSE values between the two experiments to evaluate the impact of PCA on clustering quality.
* **Visualization**: Visualize clusters in reduced dimensions and compare with original labels to understand clustering effectiveness.

1. **Effect of PCA**: PCA simplifies the dataset, improves computational efficiency, and enhances the separability of clusters.
2. **Optimal Clusters**: The elbow method identifies the best number of clusters, balancing underfitting and overfitting.
3. **Interpretability**: Reduced dimensions make clustering results easier to interpret and visualize.

challenges

* **Scalability**: Both PCA and K-Means can struggle with very large datasets.
* **Interpretation**: Principal components are linear combinations of original features, making them harder to interpret.

Principal Component Analysis and K-Means clustering are powerful tools for dimensionality reduction and unsupervised learning, respectively. Their synergy enables more effective clustering by addressing the challenges of high-dimensional data. By applying these techniques to the Breast Cancer Wisconsin dataset, we uncover patterns that help distinguish malignant and benign tumors, demonstrating the practical utility of PCA and K-Means in real-world applications.