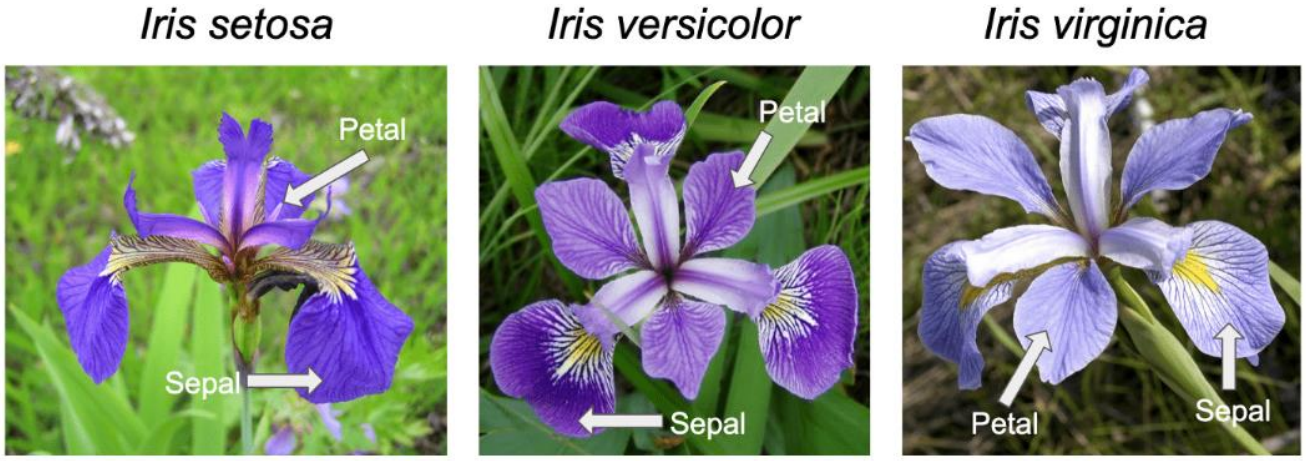
Report on k-Nearest Neighbors (k-NN) and k-means Clustering with the Iris Dataset

- **Mimoun Ayat Errahmane**
- **Aour Feriel**

Date: 08/11/2024

Objective

The goal of this work was to implement two selected machine learning algorithms: k-nearest neighbors (k-NN) for classification and k-means for clustering, using the famous Iris dataset. The purpose of this was to evaluate how well the algorithms were able to distinguishing among the three different species of iris flowers (Setosa, Versicolor and Virginica) and analyze their performance with the data provided.



Part I: k-Nearest Neighbors (k-NN) Algorithm

1. Objectives

The central aims for implementing the k-NN method were:

- To classify flowers into various species via a supervised learning approach.
- To explore data preprocessing methods such as normalization.
- To conduct modeling accuracy assessment and provide a comment on the model performance for different values of k.

2. Methodology

- **Data input and preparation:** The Iris dataset was loaded using seaborn, and then converted to a Pandas DataFrame to facilitate data manipulation in Python. The data were stored in an Excel file to facilitate access in case there was no network.
- **Exploration of data:** Simple statistics, data type, and some overviewed visuals were provided to reach an understanding of the structure of the dataset and to identify different species and features therein.
- **Data preprocessing:** We split the dataset into a training set (80%) and a test set (20%). StandardScaler was implemented to scale the features effectively for better k-NN functioning.
- **Implementation of k-NN:** k-NN was implemented with KNeighborsClassifier from sklearn with an initially assigned value of k=3.
- **Model performance evaluation:** We evaluated the performance of the model using accuracy on the test set.

3. Results

The results had been obtained after applying the k-NN algorithm.

- **High Accuracy**: The model achieved an impressive 100% accuracy for k=3.
- **Consistency:** The accuracy remained at 100% for values of k from 1 to 15.
- **Decrease in Accuracy:** Starting from k=15, the accuracy began to decrease.

4. Conclusions

The k-NN algorithm has exhibited good performance on the Iris dataset, achieving high accuracy for classification. As the algorithm is supervised, it used labeled data to predict, which then resulted in producing a reliable model for this dataset. The choice of the k value is crucial as it balances the model's sensitivity to noise and its generalization ability. An optimal k ensures high accuracy and reliable predictions.

Part II: k-means Clustering Algorithm

1. Objectives

The primary objectives of k-means clustering were as follow:

- To employ an unsupervised learning technique that forms clusters in the dataset without class labels.
- To visualize and analyze the resulting clusters.
- To compare the clustering result with the actual classes using evaluation metrics.

2. Methodology

- **Data Preparation:** In clustering, the Iris dataset was used without class labels to be a real-world unsupervised scenario.
- **Implementation of k-means:** Using KMeans from sklearn, n_clusters was set at 3, as there are three iris species in the dataset. The algorithm was run to assign each data point to a cluster.
- **Visualization of Clusters**: Scatter plots were devised to visually represent the clusters and juxtapose them against their actual classes.
- **Evaluation:** The clustering was evaluated basically through two methods:
  - ✓ **Confusion Matrix:** The clusters were compared to the actual classes, thus determining where the clusters completely matched the actual labels.
  - ✓ **Silhouette Score:** This would provide a metric for assessing compactness and separation of the clusters. The average cluster separation score was between 0.5 and 0.6, which essentially points towards moderate clustering quality itself.

3. Results

The clusters assigned by k-means clustering were seen to correspond somewhat to the actual classes in the Iris dataset:

- The clusters formed by k-means were mostly aligned with two of the classes (Iris Setosa and Iris Virginica), with some overlap between Versicolor and Virginica.
- Effects of Initialization: Observations of the fact that different starting positions of centroids generate different assignments of points to each cluster denote the sensitivity of k-means to initial conditions.

4. Conclusion

The k-means algorithm succeeded in clustering the data in ways that reflected the dataset structure. But many limitations were exposed, such as discriminating between the species Versicolor and Virginica, the former being significantly overlapped by features of the latter species. A dependence of the algorithm on the initialization point for the centroids was observed, thus marking another limitation of the k-means algorithm.

Comparison of k-NN and k-means

- **Supervised vs. Unsupervised:** The k-NN algorithm, being supervised, was highly accurate because it used labeled data. In contrast, k-means, an unsupervised algorithm, relied on feature similarity to form clusters, showing moderate alignment with the real classes but some misclassification.
- **Accuracy:** The k-NN classifier had higher accuracy, achieving approximately 100% on the test set. In comparison, k-means had a silhouette score indicating only moderate clustering quality.
- **Use Case:** k-NN is better suited for classification tasks where labeled data is available, while k-means is useful for exploratory data analysis when labels are absent.

Final Discussion

Both k-NN and k-means provided valuable insights into the Iris dataset:

- k-NN confirmed that the dataset's features could predict species labels accurately.
- k-means showed that clusters could form based on feature similarity, although some clusters overlapped due to feature distributions.

The link of our code: https://shorturl.at/VrN3L