SML-Project 2023

Shrey, and Vinayak Sharma
IIIT-Delhi, shrey21562, vinayak21574@iiitd.ac.in

***Abstract* - This report caters to the theoretical backup behind our submissions for Fruit Classification SML-project 2023 (link).**

Introduction

This project aims to segregate the given samples among 20 categories of fruits. We are provided training data of 1216 entries, each with 4096 features. We are required to classify the testing data accordingly. This report summarizes the entire workflow in a theoretical yet practical observation manner.

Methodology

The workflow can be divided into six prominent stages.

A. Data Extraction

We used Python's pandas library to load data. By pre-examining the given data, we explicitly removed constant columns for all samples.

B. Splitting

A testing ratio of 0.2 was maintained to provide an optimal no of samples for both training and testing. The testing data was randomly selected for validation purposes.

C. Clustering

We needed a clustering algorithm to provide additional labels to support the further classification process. These labels had an increased physical significance.
We decided to use the K-means clustering algorithm since it guarantees convergence and adapts to different cluster shapes and sizes.
Over recursive validation, we found the optimum no of clusters as two. It indicated the two major fruit categories: raw and ripe.
Hence, we decided to keep no of clusters as two. These additional features are added in the enriching stage to skip the reduction stage.

D. Reduction stage

Since the number of components was approximately four times the number of samples, there was a need to reduce the dimensions. We used principal component analysis to remove features having low variance.
The optimal number of PCA components was selected as 370, considering best model performance.

E. Enriching

In this stage, the features were transformed, considering class labels.
Hence, we applied linear discriminant analysis to reduce the number of features to 19. These new features are appended to the class labels in the clustering stage.
These advanced features will be provided to the classification model for finer results.

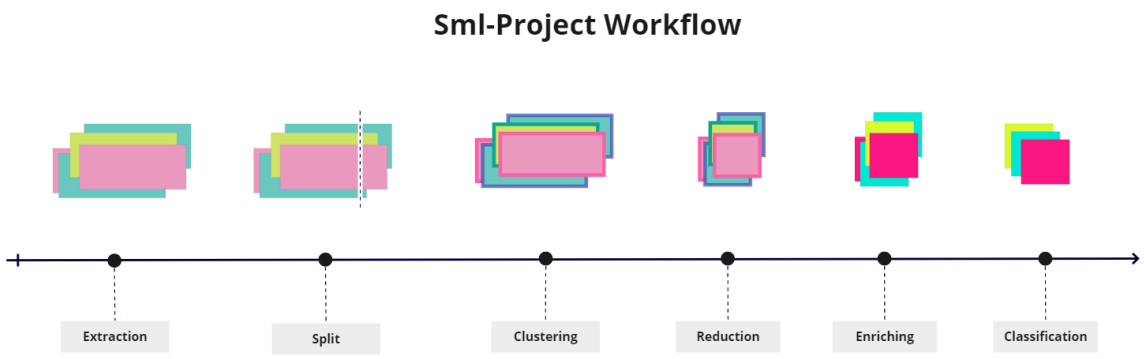
F. Classification

Given the refined features, a classification algorithm is required to segregate the samples into their appropriate labels.
We used logistic regression since it is less prone to overfitting in the case of lower dimensions.
Also, logistic regression acts as a benchmark for other models. For selecting a classification algorithm, logistic regression proved to be the best-performing one.

Observation

The observations on outlier detection.
Two algorithms are used to detect the outliers: Isolation Forest and LOF – Local Outlier Factor. Isolation Forest removed about 122 data-point on contamination=0.1. LOF was removing about 3-4 data-point on n_neighbors=20. It was observed that using an isolation forest removes too many data points, leading to decreasing the accuracy of the model. Also, it was observed that removing outliers through LOF does not affect the accuracy much, as it is not removing the data points. Hence no outliers are removed in the final model..

The observations on ensemble methods.
Many algorithms are used for ensembling, like Random Forest Classifier, XG-boosting, and gradient boosting. On validation, it was found that none gave good accuracy for the model, hence not using ensemble methods in the final model.



Applications

A. Application of PCA:

- **Image compression:** PCA can be used for image compression by reducing the dimensions of image data while getting as much of the original data as possible.
- **Signal processing:** PCA can help reduce noise and extract the most significant features from signals.

B. Application of LDA:

- **Face recognition:** LDA helps to identify the most significant features that separate face images belonging to different people.
- **Natural Language Processing (NLP):** LDA also helps to identify the most relevant topics present in a corpus of documents.

C. Application of K-Means:

- **Image segmentation:** K-means clustering helps to segment images into different regions based on their similarity.
- **Recommendation systems:** K-means clustering is used in recommendation systems to cluster users based on their preferences and behavior.

D. Application of Logistics Regression:

- **Medical diagnosis:** Logistic regression is used in medical diagnosis to predict the disease based on patient characteristics like age, gender, and medical history.
- **Fraud detection:** Logistic regression is used in fraud detection to predict fraudulent transactions based on transactional data like location, amount, and type of transaction.

Acknowledgment

We thank the college administration for giving us a significant chance to work on this project and aim to participate in more such projects. We also want to thank Dr. Koteswar Rao Jerripothula for his tremendous support and guidance during the project. The grand completion of the project would not have been possible without his help and insights. Lastly, we state that we created this project entirely and is not a forgery.

References

[1] Pedregosa, Varoquaux ,Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss,Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, Duchesnay, "Scikit-learn:" , Journal of Machine Learning Research, 2011.

[2] A. M. Martinez and A. C. Kak, "PCA versus LDA," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[3] Kalcheva, Neli, Maya Todorova, and Ginka Marinova. "Naive Bayes Classifier, Decision Tree and AdaBoost Ensemble Algorithm–Advantages and Disadvantages." KNOWLEDGE BASED SUSTAINABLE DEVELOPMENT,2020.

[4] Hamerly, Greg, and Charles Elkan. "Learning the k in k-means." Advances in neural information processing systems 16, 2003.

[5] Hilbe, Joseph M., "Logistic Regression Models", Chapman & Hall/CRC Press, 2009.