Machine Learning and Neural Networks CM3015

Table of Contents

1. Abstract

This project evaluates the performance of four machine learning models, namely, k-nearest neighbours, naive bayes, logistic regression, and decision tree on the scikit-learn iris dataset. The models are implemented using both the original dataset and a PCA-reduced version in order to assess the impact of dimensionality reduction. Results are compared across various metrics, including accuracy, precision, recall, and F1-score, with a focus on hyperparameter tuning and model complexity. Hyperparameter tuning and complexity analysis are performed to understand model behaviour and identify optimal configurations. This project aims to provide a comprehensive understanding of machine learning techniques and their practical applications on a well-established dataset while addressing the trade-offs between accuracy and model simplicity.

2. Introduction

The iris dataset is a well-known and commonly used dataset in machine learning, containing 150 samples of iris flowers divided into three species: setosa, versicolor, and virginica. Each sample includes measurements for sepal and petal length and width. The iris dataset is also often used as an introductory tool for machine learning model evaluation due to its simplicity and clear distinction between the 3 species.

This project aims to compare the performance and accuracy of the four machine learning models using both the original dataset and PCA-reduced data. Principal Component Analysis (PCA) is a dimensionality reduction technique which helps mitigate overfitting issues in high-dimensional data. Hyperparameter tuning and model complexity analysis are also applied to assess the strengths and weaknesses of each model. The findings are intended to highlight the impact of dimensionality reduction and model complexity on classification performance. This report discusses the methods used, results obtained, and possible insights derived from scoring metrics and visualisation.

3. Background

List of algorithms used:

1. k-Nearest Neighboburs (kNN)
   - The knn algorithm is a supervised machine learning algorithm commonly used in classification and regression tasks. It is known for being easy to use and implement, and also for its effectiveness in real-world applications. It works by using distance as a metric to identify the k-number of nearest neighbours and making predictions

based on similar data points in the dataset. "k", in this case, refers to any positive integer.

2. Naive Bayes (nb)
- The naive bayes algorithm acts on the basis that every pair of features being classified are independent of each other. It is most commonly used in text classification, for example spam filtering, rating classification or even sentiment detection. It is fast and making predictions are easy even with high dimensions of data. It works by using bayes' theorem, finding the probability of an event occurring given the probability of another event that has already occurred.

3. Logistic Regression
- The logistic regression algorithm is a supervised machine learning algorithm used for classification tasks. It works by transforming output from a linear regression function using a sigmoid function.

4. Decision Tree
- The decision tree algorithm is a machine learning algorithm used for both classification and regression tasks. It works by predicting the value of the target variable using simple decision rules based on the data features.

Additional techqniues:
1. Principal Component Analysis (PCA)
- a dimensionality reduction technique that transforms the data into a lower-dimensional space while preserving as much variance as possible.
- this transformation makes models more computationally efficient and reduces overfitting issues in high-dimensional data.
- in this project, PCA was used to reduce the iris dataset from 4 dimensions to 2 principal components.

4. Methodology

1. Dataset preparation
- The iris dataset will be split into training and testing sets in a ratio of 8:2 to evaluate model performance.
- PCA will be applied to the training set to create a reduced dataset with 2 principal components, which will then be applied to the test set using the same transformation.

2. Model implementation
- k-nearest neighbour, naive bayes, and PCA will be implemented from scratch as at least one of them has to be implemented from scratch following coursework requirements.

- Logistic regression and decision tree will be implemented using scikit-learn for making comparisons.

3. Evaluation metrics
- The models will be assessed based on accuracy, precision, recall and F1-score to capture their predictive performance.
- Hyperparameter tuning will be conducted for all 4 models, number of neighbours for kNN, gaussian distribution for nb, inverse of regularisation strength for logistic regression, and max depth for decision tree.

4. Analysis techniques
- Cross validation will be used to ensure the reliability of model evaluation by mitigating the effects of data variability and providing consistent results across different subsets or folds. Splitting the data into multiple 'k' folds and then averaging the results ensures that each part or subset of data is used for both training and testing. This is especially useful for small datasets like this one, the iris dataset which only has 150 samples, as the usual 8:2 train-test split might not be able to fullly capture variability in performance and could also lead to misleading results.

- Feature importance analysis helps to identify which features contribute to model predictions the most. This leads to better understanding behind model predictions, which can help with simplifying the model or even improving computational efficiency. Feature importance analysis will be performed on the decistion tree and logistic regression models. The other 2 models, kNN and nb, will not be analysed as they will be written from scratch.
    - For decision tree, importance should be determined by the contribution of each feature to the decision tree.
    - For logistic regression, importance should be determined by the magnitude of the coefficients of each feature.

- Model complexity will be analysed by varying key parameters to identify trends such as overfitting or underfitting if any. For example, increasing the range of number of neighbours in kNN or the depth of the decision tree might reveal insights about model flexibility and generalisation ability.

5. Results

Model	Dataset	Accuracy	Precision	Recall	F1 Score
kNN	Original	1	1	1	1
kNN	PCA-Reduced	0.93	0.93	0.93	0.93
Naive Bayes	Original	1	1	1	1
Naive Bayes	PCA-Reduced	0.97	0.97	0.97	0.97

Logistic Regression	Original	1	1	1	1
Logistic Regression	PCA-Reduced	0.90	0.90	0.90	0.90
Decision Tree	Original	1	1	1	1
Decision Tree	PCA-Reduced	0.87	0.87	0.87	0.87

Feature importance analysis

- Decision tree:

Petal length showed the most impact on predictability with an importance score of 0.94, followed by petal width with an importance score of 0.06, while sepal length and width score 0. These findings strongly correlate with the class separability of the dataset.

- Logistic Regression:

Similar to the decision tree model's feature importance analysis, the petal length feature has the highest magnitude in coefficient scores with a score of 1.81, followed by petal width with a score of 1.69. However, unlike decision tree, the coefficient scores for sepal length and width are not completely negligible with scores of 1 and 1.14. This shows that applying PCA might cause a decrease in performance as features are reduced.

- PCA Analysis:

By adding up the values of both principal components variance scores, we can tell that the PCA-reduced dataset retained approximately 97% of the variance, making it an efficient representation of the data.

Models performed slightly worse on the PCA-reduced dataset, for example the kNN model got a score of 0.93 across all 4 metrics on the PCA-reduced data, a drop from a score of 1 across all metrics. The same for the rest of the models, dropped from 1 to 0.97 for the nb model, 1 to 0.9 for logistic regression model and 1 to 0.87 for decision tree model. From this, we can tell that PCA affects logistic regression and decision tree more than the other 2 models.

6. Evaluation

Model Complexity Analysis

- kNN:

The analysis of the number of neighbours, k, showed that a value of 2 led to overfitting, as the model captured noise in the training data. As k increases, training

accuracy generally decreases, this is expected as a larger number of neighbours would make the model a lot more generalised. Tuned k is valued at 5, which is to be expected as the iris dataset only has 150 samples, which is very small. The number of features is even lesser, at only 4, and having too large a number of neighbours would only make the model overly generalised.

- Logistic Regression:

The analysis of regularization strength (C) showed that smaller values of C, which means stronger regularization, simplified the model, reducing overfitting, while larger values increased model complexity but could lead to overfitting.

- Decision Tree:

The analysis of the max depth for the decision tree model showed that max depths of under 2 led to underfitting, indicated by both train and test accuracy being low. While max depths of 6 and above led to over fitting, as training accuracy scored perfectly. Tuned max depth is 4, this is expected as there are only 3 types to classify.

Impact of PCA

- PCA helped simplify the dataset while retaining most of the variance. While this reduced dimensionality improved computational efficiency, it resulted in slightly lower model performance as critical feature-specific information was lost.

- Models like kNN and naive bayes were more robust to PCA-reduced data, while logistic regression and decision tree models showed slightly reduced accuracy due to their reliance on individual feature contributions.

7. Conclusions

This report highlights the performance of 4 machine learning models, k-nearest neighbour, naive bayes, logistic regression, and decision tree on both original and PCA-reduced datasets.

First of the key findings, is that all 4 models performed perfectly on the original dataset. This shows how the iris dataset's linear separability and simplicity allowed the models to classify them perfectly.

Second of the key findings, is that kNN and naive bayes were more robust to the PCA-reduced data achieving high accuracy scores of 0.93 and 0.97, as compared to the logistic regression and decision tree models that did not perform as well, with accuracy scores of 0.9 and 0.87.

Third of the key findings, is that performing PCA on the iris dataset is effective. Although reducing accuracy slightly for all 4 models, having only 2 principal components to compute reduces processing time. This is especially true for models like kNN, where lesser features meant lesser or faster calculations of distance.

Fourth of the key findings, is that feature importance analysis revealed that the petal-related features, petal length and width, contributed the most towards predictability.

Fifth and last of the key findings, is that model complexity analysis highlighted the importance of tuning hyperparameters to balance overfitting and underfitting. For example, the tuned parameters for kNN and decision tree were k=5 and max depth=5. These tuned parameters provied the best balance between training and test accuracy.

In conclusion, this project showed the effectiveness of machine learning algorithms and dimensionality reduction techniques in combination to achieve computational efficiency and robust classification performance. The insights further emphasise the importance of proper EDA, hyperparameter tuning, and feature analysis in building models that are reliable and easy to understand.

8. References

1. https://www.datastax.com/guides/what-is-k-nearest-neighbors-knn-algorithm
2. https://www.geeksforgeeks.org/naive-bayes-classifiers/
3. https://www.geeksforgeeks.org/understanding-logistic-regression/
4. https://www.geeksforgeeks.org/decision-tree-algorithms/