COGS 118A

Final Report

Zening Wang
A161414249
2023/12/16

1. Abstract

This report presents a investigation into the performance of various fundamental machine learning methods across diverse datasets. It aims to assess the effectiveness of these 5 methods: Linear Regression, Logistic Regression, SVM, Random Forest and KNN. The 4 datasets chosen are Adult, Wine, Heart, and Iris. By employing a range of datasets with distinct characteristics, the research provides insights into the generalizability and adaptability of each machine learning approach.

2. Introduction

Machine learning has emerged as a transformative force in the realm of data-driven decision-making, with various methods offering unique approaches to modeling and prediction tasks. The datasets chosen for analysis are sourced from the class designated repository, representing distinct domains, varying data sizes, and diverse prediction objectives.

In the training process, key methodologies such as cross-validation and hyperparameter tuning are employed to enhance model performance and robustness. Additionally, this study explores the effectiveness of different training and testing strategies, focusing on optimal approaches for achieving accurate and generalizable models. As we navigate through each dataset, our objective is to not only identify the best-performing model for each scenario but also to provide explanations for these outcomes.

3. Method

The basic idea of our methods is very thorough. For each dataset we are testing, after we clean the data, we would have 3 types of partitions of training and testing for the dataset (80%training-20%testing, 50%training-50%tesing and 30%training-70%testing) to examine the influence of limited training set size to different datasets and machine learning methods. To mitigate potential biases within individual datasets, we conduct three iterations of training and testing for each partition, ensuring that the dataset is completely shuffled before each iteration to counteract the influence of inherent data ordering and to address the impact of randomness. In terms of hyper parameter tuning, we would set the greatest n_estimators of Random Forest as my laptop could

afford because increasing the number of trees generally improves the model's performance, but it also increases computational cost. For the KNN method, we also set a parameter list for every dataset and find the optimal hyper parameter which is n-neighbors. This meticulous methodology ensures a comprehensive exploration of the parameter space, allowing us to draw meaningful conclusions about the performance of each method across diverse datasets and varying training set sizes.

4. Experiment

Notice that accuracy I provide is the average value of three trails which use different portions of dataset for training and testing. And the linear regression uses mean squared error to represent accuracy, so it is difficult to compare other machine learning methods.

Here is the accuracy report of Adult dataset (Notice that the Adult dataset has significantly larger data sizes than other 3 datasets. It has 47621 rows after cleaning while others most have hundreds)

Table for Adult dataset

Partitions/Methods	Linear Regression(mse)	Logistic Regression	SVM	Random Forest	KNN
80%training/20%testing	0.1167	0.7962	0.7964	0.8489	0.7857
50%training/50%tesing	0.1167	0.7962	0.7950	0.8339	0.7897
30%training/70%testing	0.1171	0.7963	0.7898	0.8335	0.7883

Here is the accuracy report of Heart dataset:

Table for Heart dataset

Partitions/Methods	Linear Regression(mse)	Logistic Regression	SVM	Random Forest	KNN
80%training/20%testing	0.6575	0.6667	0.6222	0.6333	0.5667
50%training/50%tesing	0.7966	0.5369	0.5794	0.5838	0.4876

30%training/70%testing	0.8237	0.5641	0.5512	0.5480	0.4983

Here is the accuracy report of Wine dataset:

Table for Wine dataset

Partitions/Methods	Linear Regression(mse)	Logistic Regression	SVM	Random Forest	KNN
80%training/20%testing	0.0786	0.9351	0.7407	0.9629	0.7778
50%training/50%tesing	0.0905	0.9138	0.7378	0.9588	0.7078
30%training/70%testing	0.0954	0.9093	0.7439	0.9199	0.6906

Here is the accuracy report of Iris dataset:

Table for Iris dataset

Partitions/Methods	Linear Regression(mse)	Logistic Regression	SVM	Random Forest	KNN
80%training/20%testing	0.0433	1.0	0.9888	1.0	0.9777
50%training/50%tesing	0.0522	0.9644	0.9466	0.9422	0.9467
30%training/70%testing	0.0513	0.9619	0.9460	0.9428	0.9523

Linear regression is generally different from other methods since I use mean squared error instead of accuracy to measure its correctness. And this indicates that the lower the mse is, the better it performs, contrary to the accuracy metric where higher values denote better performance. We can find that in 3 of 4 datasets, its error is generally small. And its performance is still good even if the data size decreases compared to the first dataset. For the Adult dataset, the model's performance is relatively consistent across different training-testing splits, which might be attributed to the larger data size providing a more robust training process. The reason why it has significantly larger error rate for Heart dataset might be that most of the features have nonlinear relationships and the result variable is non-binary. The lower MSE in Wine and Iris datasets indicates a better fit, possibly due to more linear relationships within the data.

Then I would compare all other methods together across datasets.:

1. **Adult Dataset**:
    - **Random Forest** emerges as the top performer, showcasing its strength in handling large datasets with complex features.
    - **Logistic Regression, SVM, and KNN** show similar performance, indicating their capability to handle large datasets but with some limitations compared to Random Forest.
2. **Heart Dataset**:
    - The performances of all methods are relatively moderate.
    - This dataset's complexity, possibly due to non-linear relationships and a non-binary target variable, challenges all methods, preventing any from standing out significantly.
3. **Wine Dataset**:
    - **Random Forest** again shows exceptional performance, underlining its adaptability to datasets with complex patterns.
    - **Logistic Regression** also performs well, suggesting its effectiveness in datasets with well-defined class separations.
4. **Iris Dataset**:
    - Exceptional performance by **Logistic Regression and Random Forest**, indicating their strong suitability for datasets with clear, distinct class separations.
    - **SVM and KNN** also perform well, but slightly lag behind the top performers.

Lastly, I also give some insight into method-specific:

1. **Logistic Regression**: Shows consistent performance across datasets. Particularly effective in scenarios with binary or well-defined target classifications.
2. **SVM**: Generally performs well, especially in datasets where margin maximization and kernel tricks are beneficial. However, its running time is substantially longer than any other methods, especially for Adult dataset, showcasing its deficit with large quantity of data.
3. **Random Forest**: Stands out in handling large and complex datasets with non-linear relationships. Its ensemble approach provides robustness against overfitting.
4. **KNN**: While effective, its performance depends heavily on the choice of 'k' and the dataset's features. It can struggle with high-dimensional data due to the limitation of dimensionality.

5. Conclusion

One of the conclusions is that each method has its strengths and weaknesses, and their performance varies across different datasets. The choice of the method should be guided by the nature of the dataset, the complexity of the features, and the type of problem being solved. So a deep understanding of the dataset is crucial for selecting the appropriate machine learning method for optimal performance.

Another point is that methods like Random Forest tend to be more adaptable and robust across various datasets, while others like SVM and Logistic Regression have specific scenarios where they excel. SVM and KNN can be highly effective but often require careful tuning and understanding of the dataset for better performance. Also, its simplicity and interpretability make Logistic Regression an excellent baseline model for binary classification tasks. While Linear Regression excels in datasets where relationships between variables are linear. It's particularly effective for continuous target variables, as seen in the lower MSE values for the Wine and Iris datasets.

6. Reference

Code link: https://github.com/Zening-W/Analysis-on-different-ML-methods

Fisher,R. A.. (1988). Iris. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Aeberhard,Stefan and Forina,M.. (1991). Wine. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.

https://en.wikipedia.org/wiki/Random_forest

https://en.wikipedia.org/wiki/Support_vector_machine

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm