MiniProject 1: Getting Started with Machine Learning

David Schrier 260827890

Theo Janson 260868223

Jacob McConnell 260706620

Abstract

Computer Algorithms have evolved over many decades of research and development which has lead to the creation of the machine learning field where with past experiences, computers can make predictions on future data. Machine learning has become a strong tool for applications in healthcare, engineering, language, etc. The goal of this project is to introduce the field of machine learning, so our team was tasked to implement two classical techniques and comparing their performance on two well-known datasets. These two classical techniques are K-Nearest Neighbours and Decision Trees. Our results have shown that K-Nearest Neighbours performed roughly equally to Decision Trees on the provided datasets.

Keywords: K-Nearest Neighbours, Decision Trees, K-Fold Cross-Validation

1. Introduction

Breast cancer has been recorded to be the second largest contributor of cancer diagnoses among all other cancers in the world [1]. In 2021 that there will be approximately 281,550 positive diagnoses and 43,600 deaths in the United States [1]. Similarly, viral hepatitis is a continual health crisis where the diagnostic rate has increased over the last few years [2]. 1 According to the World Health Organization (WHO), in 2015, viral hepatitis has resulted in 1.34 million deaths worldwide [2]. In both these cases accurate diagnosis of the underlying disease is very valuable. Machine Learning techniques for diagnosis are especially useful as healthcare field has an abundance of available data of patients' health history. There are many different methods for predicting disease, but in this report we will compare two well-known classifier algorithms: K-Nearest Neighbors and Decision Trees. The task is to develop algorithms to predict cases of breast cancer and hepatitis. The models' hyperparameters were tuned by the average accuracy of the validation sets. The datasets used in this report are the well-known UCI Breast Cancer Wisconsin Dataset and the UCI Hepatitis dataset. Results from other papers have shown that K-Nearest Neighbours perform much better compared to Decision Trees [3]. A detailed report of plots, histograms, and tables can be found in the associated Python Notebooks.

2. Datasets

Refer to the Data Processing Python Notebook for tables, histograms and plots.

2.1 Hepatitis Dataset

The dataset's features were not normalized, and many features contained missing examples. Some of the features' domains were binary, and others were discrete over a larger

interval. The dataset was investigated for skewness and the results showed that there were 32 examples of positive hepatitis and 123 examples for the negative case. This issue raises a concern as we did not want the removal of examples to further skew the dataset. We started by removing examples containing missing data, cutting the dataset by almost half. We also considered removing features containing many missing examples, such as 'protime', to conserve more data. This could pose a problem if these features are highly correlated with the class label. We chose to rank the features by the absolute value of their Pearson correlation coefficients, calculated against the class label. The aim was to drop features with too much missing data if its Pearson's coefficient was low enough. These results can be found in Section 1.1. The features 'protime' and 'alk phosphate', which contained many incomplete examples were moderately correlated to the label. Our conclusion was to create two datasets: a 'large' one with these features removed, but more examples, and a 'small' one with these features present, but with fewer examples. To visualize the data distributions and the correlations to the label, we plotted histograms for each feature and a heatmap. We also plotted a scatter plot of the three most highly correlated features. The histograms can be found in Section 1.2 of the Python notebook, and confirm our statistical findings using Pearson's coefficient. After investigating the dataset and creating a 'small' and 'large' dataset, we normalized the both datasets and scaled the features by their Pearson's correlation coefficient and a factor of 10. The most highly correlated features are scaled to a larger range, proportionally to their correlation. To assess the effect of scaling, we also kept non-scaled, but normalized datasets. Finally, we also created datasets in which we removed features with a Pearson's correlation coefficient lower than 0.3. This retains only the features which are most predictive and lowers the dimension. This had the effect of lowering the feature space's dimension of 20 to 7.

2.2 Breast Cancer Dataset

The cancer dataset was investigated through the same process as the hepatitis dataset. An initial look at the data shows two classes, that the data is not normalized and contains only 16 missing values, all associated with the 'bare nuclei' feature, as seen in Table 3. These incomplete examples were removed from the dataset. We also noted that one feature was labelled 'id', which we confirmed was not correlated to the class label. Many of the features in the cancer dataset are highly correlated to the class label. These findings can be found in Section 2.2. Visuals of the data distributions confirm these findings and can be found in Section 2.3. We addressed the issues of normalization and scaling. The dataset was first normalised, and the features were scaled by a their Pearson's correlation coefficient and a factor of 10, as we did with the hepatitis dataset. This results in two normalized datasets: a scaled version and a non-scaled version.

3. Results

We used test sets comprised of 20% of the data to test the models. The model's hyper-parameters were tuned by average performance of the validation set in 5-fold Cross Validation on both datasets. For the breast cancer dataset, we tested both the scaled and non scaled datasets to determine which preprocessing method worked best for each algorithm. As for the hepatitis dataset, we first tested our 'small' and 'large' datasets using Euclidean distance. We then tested scaled and non-scaled datasets. Finally, the datasets which produced the best results were also tested using Manhattan distance and tested after removing

features with low correlations to the class label. We assessed the datasets' and the models' performances using accuracy, precision, recall, F1, and loss. These results can be found in the code. Accuracy was used to tune hyper-parameters.

3.1 K-Nearest Neighbours

To determine the best hyper-parameter K, we obtained the average and standard deviation of the model's accuracy for various K values, ranging from 1 to 20.

3.1.1 Breast Cancer Dataset

	Dataset	Distance	Highest Accuracy (Val)	Highest Accuracy (Test)	Best K (Val)	Best K (Test)
0	cancer scaled	Euclidean	0.970642	0.970588	5.0	5.0
1	cancer non-scaled	Euclidean	0.968807	0.985294	6.0	5.0
2	cancer non-scaled	Manhattan	0.970642	0.985294	3.0	3.0
3	cancer scaled	Manhattan	0.976147	0.970588	3.0	3.0

Figure 1: KNN Cancer Table

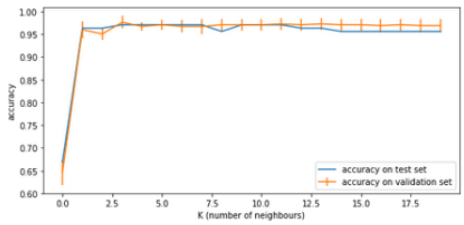

Figure 2: KNN Cancer Scaled Manhattan - Accuracy

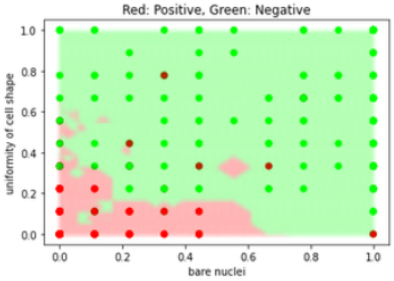

Figure 3: KNN Cancer Scaled Manhattan - Decision Boundary

3.1.2 Hepatitis Dataset

We first tested the 'large' and 'small' datasets for the scaled an non scaled cases, using both distances for the former. The 'small' one performed better for both cases. We also tested the 'small' and 'large' datasets with poorly correlated features removed.

	Dataset	Distance	Highest Accuracy (Val)	Highest Accuracy (Test)	Best K (Val)	Best K (Test)
0	Large Scaled Hepatitis	Euclidean	0.840000	0.9600	4.0	15.0
1	Small Scaled Hepatitis	Euclidean	0.933333	0.9375	6.0	19.0
2	Small Non-Scaled Hepatitis	Euclidean	0.883333	0.9375	9.0	1.0
3	Large Non-Scaled Hepatitis	Euclidean	0.870000	0.8400	7.0	8.0
4	Small Scaled Hepatitis (features removed)	Euclidean	0.916667	0.8750	9.0	17.0
5	Large Scaled Hepatitis (features removed)	Euclidean	0.830000	1.0000	13.0	3.0
8	Small Scaled Hepatitis	Manhattan	0.883333	0.8750	3.0	3.0
9	Large Scaled Hepatitis	Manhattan	0.840000	1.0000	11.0	3.0

Figure 4: KNN Hepatitis Table

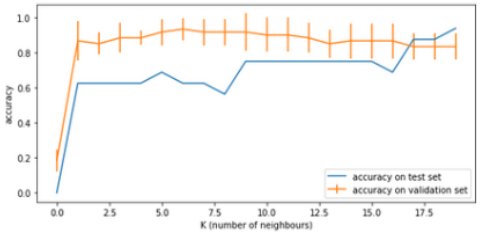

Figure 5: KNN Small Scaled Hepatitis Euclidean - Accuracy

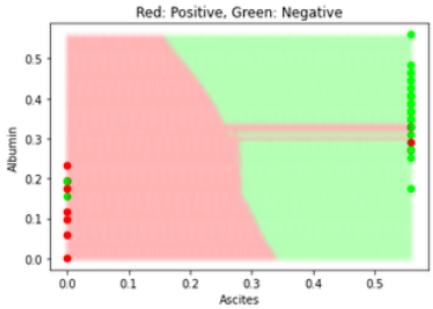

Figure 6: KNN Small Scaled Hepatitis Euclidean - Decision Boundary

3.2 Decision Trees

To determine the ideal max depth of the tree, we obtained the average and standard deviation of the model's accuracy for various K's, ranging from 1 to 20.

3.2.1 Breast Cancer Dataset

	Dataset	Highest Accuracy (Val)	Highest Accuracy (Test)	Best Max Depth (Val)	Best Max Depth (Test)
0	Scaled Breast Cancer	0.946789	0.970588	2.0	4.0
0	Non-Scaled Breast Cancer	0.952294	0.948529	4.0	2.0

Figure 7: DT Cancer Table

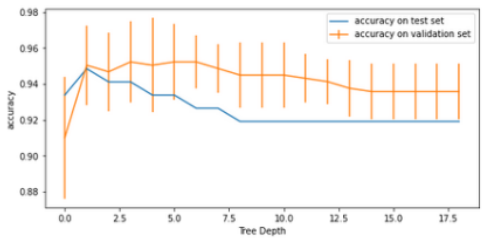

Figure 8: DT Non-Scaled Breast Cancer - Accuracy

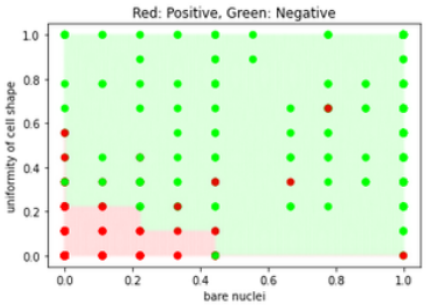

Figure 9: DT Non-Scaled Breast Cancer - Decision Boundary

3.2.2 Hepatitis Dataset

	Dataset	Highest Accuracy (Val)	Highest Accuracy (Test)	Best Max Depth (Val)	Best Max Depth (Test)
0	Small Scaled Hepatitis	0.850000	0.875	3.0	1.0
1	Large Scaled Hepatitis	0.800000	0.800	5.0	2.0
2	Large Non-Scaled Hepatitis	0.840000	0.840	8.0	2.0
5	Large Non-Scaled Hepatitis with Features Removed	0.870000	0.800	1.0	2.0
8	Small Non-Scaled Hepatitis	0.816667	0.875	1.0	4.0

Figure 10: DT Hepatitis Table

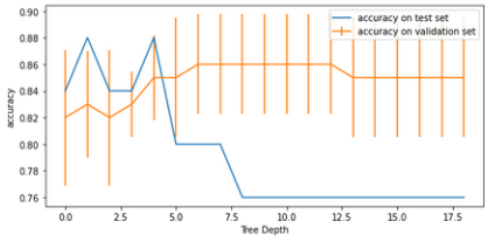

Figure 11: DT Hepatitis Larger Non-Scaled Removed - Accuracy

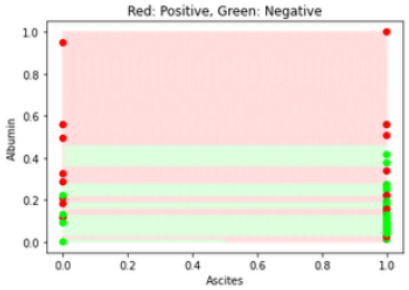

Figure 12: DT Hepatitis Larger Non-Scaled Removed - Decision Boundary

4. Discussion and Conclusion

The breast cancer dataset that showed the best performance for KNN was the scaled one, but it performed minimally better than the others, which suggests that as most of the features are highly correlated to the data, feature scaling was unnecessary. Moreover, there was a negligible difference in performance between using Manhattan and Euclidean distances. The ideal number of neighbours was determined to be 5. KNN performed slightly better than decision trees, which also showed best performance for the scaled dataset. The ideal max tree depth was determined to be 2. For both hyper-parameters, the plots of the decision boundary and accuracy in figures 2 and 3 suggest that the models were not over-fitting the data.

The 'small' and scaled hepatitis dataset performed best for KNN, and notably, performed better than the larger and non-scaled datasets, which suggests that feature scaling is very important when the dataset contains poorly correlated features. Moreover, while the 'small' dataset contained less data, it retained moderately correlated features which were dropped in the larger dataset in order to have more examples. Retaining these features seems to have played an important effect on predictive ability, as few features in the hepatitis dataset were strongly correlated to the label. The best K value from validation set testing was determined to be 5. While the decision boundary in figure 6 suggests that the model may have been slightly over-fitting the data, the histograms in the data processing Notebook show that there are numerous positive examples with a normalized 'ascites' score above 0.5. The decision tree algorithm performed best on the 'large', non scaled dataset with poorly correlated features removed. The ideal tree depth for this dataset was found to be 1, which seemed rather shallow. We suppose that as few examples were available for training, the tree was over-fitting at a shallow depth. Notably, the decision tree algorithm performed slightly worse than KNN. This seems to indicate that the hepatitis dataset was

not linearly separable – positive and negative examples could be found in the same regions of the feature space. The histograms in the Data Processing Notebook highlight this fact. This suggests that one is better of having few, but better features when using decision trees, which aligns with our results. We suppose that the decision tree struggled to accurately classify hepatitis, while KNN was able to look at the combination of many slightly important features in aggregate to get an accurate classification.

In conclusion, while in our results show that K-Nearest Neighbours did outperform Decision Trees, there is still a large use for Decision Trees because of their quick computation and they perform fairly well, which might be preferable in some medical situations, whereas K-Nearest Neighbours may be used for slightly more accurate results by giving up more computation time.

5. Statement of Contribution

The distribution of work is as follows: Theo Janson worked on the data preprocessing and K-Nearest Neighbours Implementation. David Schrier worked on the beginning of the Decision Tree implementation and wrote majority of the report in Overleaf. Jacob McConnell finished the Decision Tree implementation and worked on other miscellaneous techniques and debugging.

References

[1] U.S. breast cancer statistics. (2021, February 04). Retrieved February 06, 2021, from https://www.breastcancer.org/symptoms/understand_bc/statistics
[2] Viral hepatitis: A hidden killer gains visibility. (2017, May 23). Retrieved February 06, 2021, from https://www.who.int/publications/10-year-review/hepatitis/en/
[3] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.