Implementation & Analysis of KNN and DT

COMP 551 - Assignment 1

Group 102
Sevag Baghdassarian (260980928)
Mike Zhu (261069547)
Rosa Wu (260901912)

Abstract

The assignment consisted of implementing the K-Nearest Neighbours (KNN) and Decision Tree (DT) machine learning algorithms from scratch, applying them on two fixed benchmark data sets acquired from the web, and analyzing their performance.

To ensure maximal accuracy, we tested on many different values for the hyper parameters, and also implemented a weighted version of KNN to compare performance. We found that KNN performed better than DT on both data sets in terms of loss, AUROC, and AUPRC, even though DT proved to have better accuracy on the Hepatitis data sets which is imbalanced.

Introduction

The KNN and DT algorithms were written and tuned according to two benchmark data sets concerning Hepatitis and Diabetic Retinopathy Debrecen. The data sets contained missing values, which had to be discarded for the KNN implementation. Upon analyzing the data, we found that the Hepatitis data set is overly imbalanced, which led us to do over-sampling (1). Since the data contains more categorical features than numerical ones, all the numerical features were converted to categorical ones, and as such, the magnitudes of the binary features should not bear any meaning. In the end, the DT model obtained better accuracy on the Hepatitis testing data set, and the KNN model obtained better accuracy on the Diabetes testing data set, but KNN had better precision-recall overall.

Methods

The KNN algorithm predicts the label of a new data point by finding the most common label among the K points whose features are most identical to the point. These points are determined by a distance function defining what it means for two points to be “close,” such as Euclidean distance (for numerical features), or Hamming distance (for categorical features).

On the other hand, the DT algorithm classifies a new data point by running the point through a tree, each level containing a decision node which is a conditional classifier on one of the features of the point, leading it through a certain branch. The label of the new data point is determined by the leaf node that the point reaches.

Data Sets

The first data set consists of features regarding characteristics of people with Hepatitis, such as age and sex, and the labels correspond to whether they lived or died. Malformed data (points with missing features) were cleared from the data set. As apparent in the Figure 1 provided in the report, the Hepatitis data is heavily imbalanced in terms of labels. Further, the feature importance for the Hepatitis data set is demonstrated in Figure 5.

The other data set consists of predicting whether an image contains signs of diabetic retinopathy. The labels for this data set are much more balanced. **Feature scaling** was also used to normalize the range of the features. Figure 6 shows feature importance before and after scaling.

Results

KNN achieves an accuracy of 85% upon prediction on the Hepatitis testing set, whereas DT achieves an accuracy of 87.5%. However, since the Hepatitis data set is heavily imbalanced, the accuracy is not a great measure of performance, and it's better to compare the AUPRC. A comparison of the AUPRC can be seen in Table 1. KNN has better AUPRC than DT. We found that indeed, in all cases, KNN obtained higher AUPRC than DT.

On the other hand, to deal with the Diabetes data set, since most of its features are numerical, we also implemented a **weighted KNN** model (2), discussed in detail further in the report. A comparison of the performances of all three models following different distance/cost functions can be found in Table 2. As apparent from the table, DT exhibited significantly greater loss than both KNN models on this data set.

The Hepatitis labels were heavily imbalanced (see Figure 1), which pushed us to do **over sampling** on the training set. We also analyzed feature importance using random forest classifier. Feature importance is discussed in more detail further in the report.

For both data sets, validation loss was used to determine the hyperparameter (e.g. k for KNN and max depth for DT). We used validation loss instead of validation accuracy because validation accuracy is not a perfect evaluation metrics on imbalanced data sets. Loss was measured using binary cross-entropy, since we have binary labels. Figure 2 shows the validation loss for the Hepatitis data set for different values of k. For the Diabetes set, we decided to also implement a weighted KNN version for increased accuracy, by giving more weight to the nearest neighbours. Table 2 presents a comparison of the performance of regular KNN and weighted KNN per distance function.

Generally, with all cost functions used for the DT model, an increase in max depth resulted in an increase in validation loss. Validation loss per max depth for DT on the Hepatitis data set can be found in Figure 2.

For KNN, we tried using two different distance functions on the Diabetes set (since its features are numerical): Euclidean and Manhattan. Figures 7 and 8 compare the confusion matrices for Euclidean and Manhattan distance functions on regular KNN and weighted KNN, respectively. For both regular and weighted KNN, the case for Euclidean distance has slightly more elements along the diagonals than Manhattan distance (more points which are predicted correctly).

In the case of DT, we tested three different cost functions: Missclassification cost, Entropy cost, and Gini index cost. For the Hepatitis set, all three cost functions resulted in a similar best loss value of 3.2378. As for the Diabetes set, Entropy cost turned out to yield the lowest best loss at 2.7995. The AUROC comparisons for these three cost functions on the Diabetes set can be found in Table 2.

Since the data sets have high-dimensional features, it would be infeasible to plot them in a 2D graph. As such, we took two important features from each data set using a feature importance graph, and plotted one against the other. Given that most of the features of the Hepatitis data set are categorical, we transformed the numerical features to categorical ones and used one-hot encoding to encode them. As such, we have 62 features for the Hepatitis set in total. Figure 5 shows the resulting feature importance graph for the KNN model. As apparent from the graph, feature 53 turns out to be the most important, which is the same as the feature importance analyzed using random forest classifier. The decision boundary plots on both the KNN model and the DT model for the Hepatitis set using two important features can then be found in Figure 8. For the case of DT, feature 2 turns out to be the most important on the Diabetes set, which is also the same result using random forest classifier.

Conclusion

All in all, we learned that imbalances in the data sets may lead to very broad results; accuracy being high doesn't mean the model has good performance, especially on such data sets. Strategies must be taken into account these effects, such as over-sampling and feature scaling. Furthermore, we learned that different learning models may perform very differently on different data sets, and depending on the cost/distance functions chosen, there are different ways of interpreting results. It is up to the designer to determine which functions to choose and why by analyzing performance on the validation set, and that is an interesting topic to investigate further.

Statement of Contributions

Sevag Baghdassarian: report write-up and code review
Mike Zhu: Task 1, 2 & 3, report review
Rosa Wu: Task 2 & 3, report review

References

[1] J. Brownlee, *Random oversampling and undersampling for imbalanced classification*, Machine Learning Mastery, 04-Jan-2021. [Online]. Available: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. [Accessed: 05-Oct-2022].

[2] J. McCaffrey, *Weighted K-NN classification using Python*, Visual Studio Magazine. [Online]. Available: https://visualstudiomagazine.com/articles/2019/04/01/weighted-k-nn-classification.aspx. [Accessed: 06-Oct-2022].

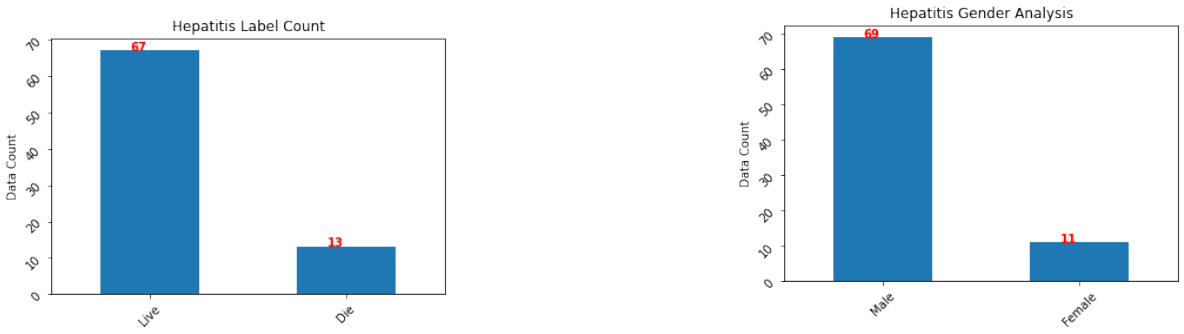
Appendix

	KNN	DT
Loss	0.3273	0.9067
Accuracy	0.85	0.875
AUROC	0.9264	0.6991
AUPRC	0.984	0.9418

Table 1: Performance of KNN and DT models on the Hepatitis data set

	KNN Euclidean	KNN Manhattan	Weighted KNN Euclidean	Weighted KNN Manhattan	DT Misclassification	DT Entropy	DT Gini
Loss	0.6368	0.7145	0.6300	0.6275	2.6683	2.0801	4.1415
Accuracy	0.6319	0.6215	0.6475	0.6562	0.625	0.6180	0.6059
AUROC	0.6927	0.6976	0.7013	0.7148	0.6379	0.6658	0.6108
AUPRC	0.7306	0.7264	0.7393	0.7453	0.6532	0.7255	0.7073

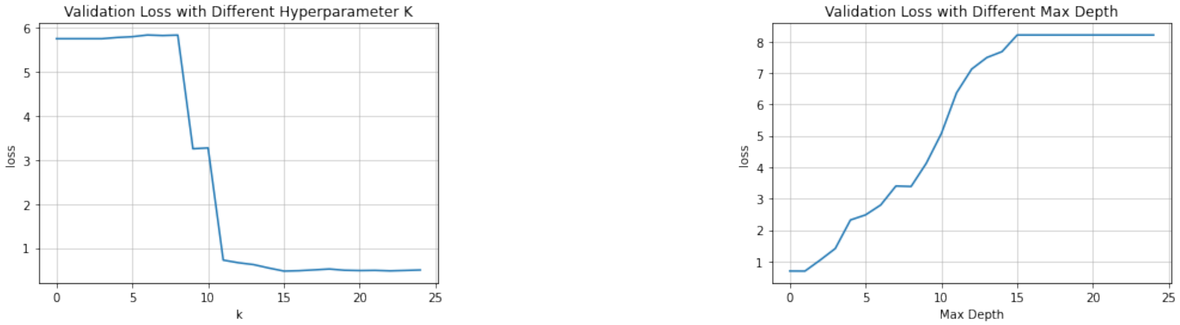
Table 2: Performance of KNN, weighted KNN and DT models on the Diabetes data set per distance/cost function



(a) Label imbalance

(b) "Gender" feature imbalance

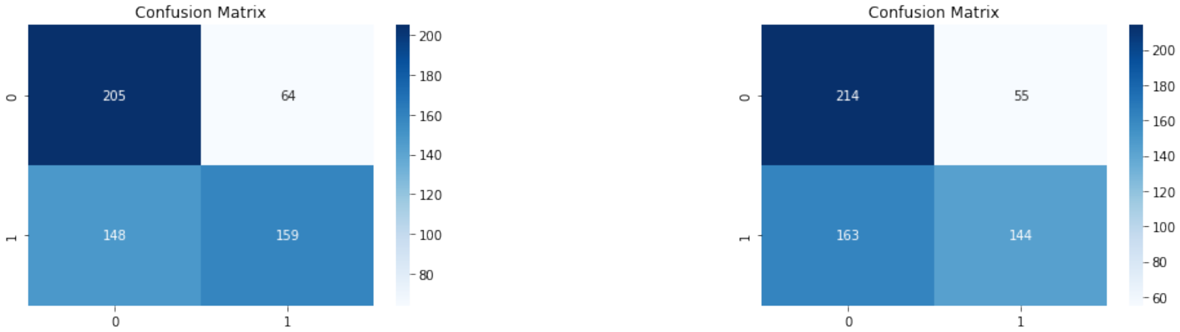
Figure 1: Label and feature imbalance in the Hepatitis data set



(a) Validation loss for KNN

(b) Validation loss for DT

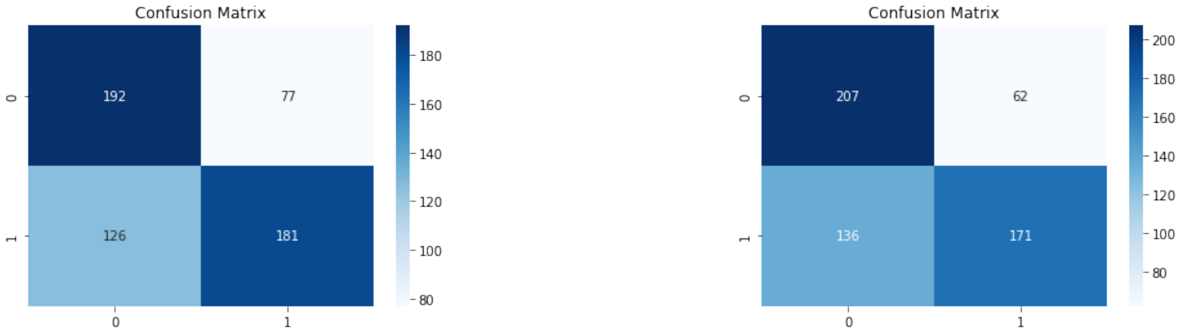
Figure 2: Validation loss graphs for KNN and DT on the Hepatitis data set



(a) Confusion matrix for the Diabetes set using Euclidean distance

(b) Confusion matrix for the Diabetes set using Manhattan distance

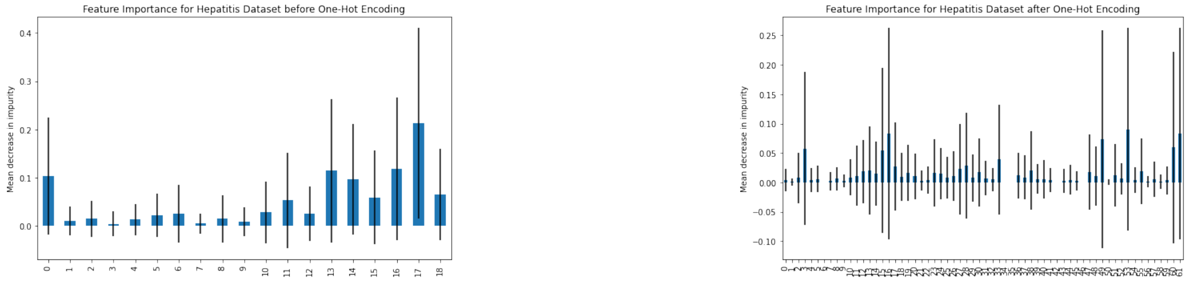
Figure 3: Comparison of confusion matrices for the two distance functions (regular KNN)



(a) Confusion matrix for the Diabetes set using Euclidean distance

(b) Confusion matrix for the Diabetes set using Manhattan distance

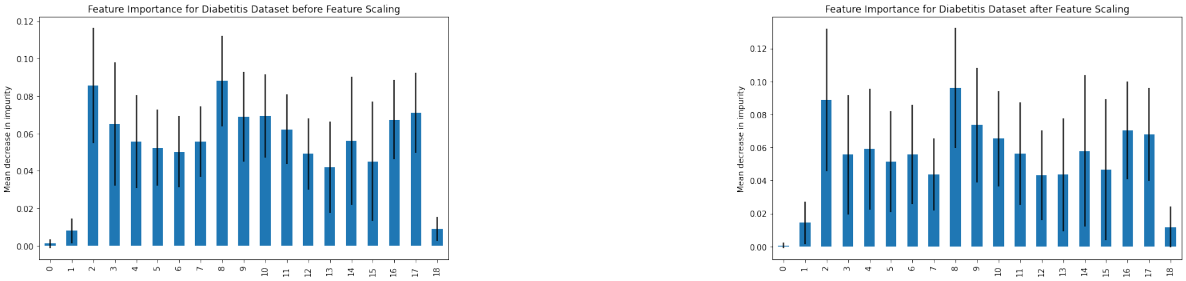
Figure 4: Comparison of confusion matrices for the two distance functions (weighted KNN)



(a) Before one-hot encoding

(b) After one-hot encoding

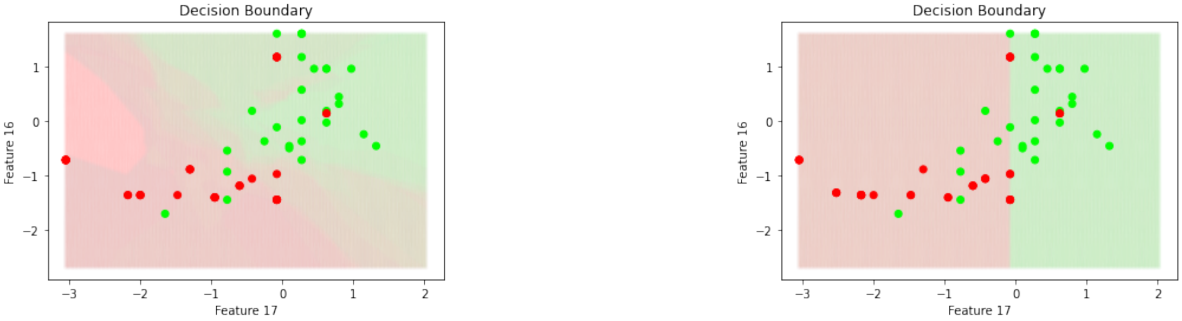
Figure 5: Feature importance for the Hepatitis data set before and after one-hot encoding



(a) Before feature scaling

(b) After feature scaling

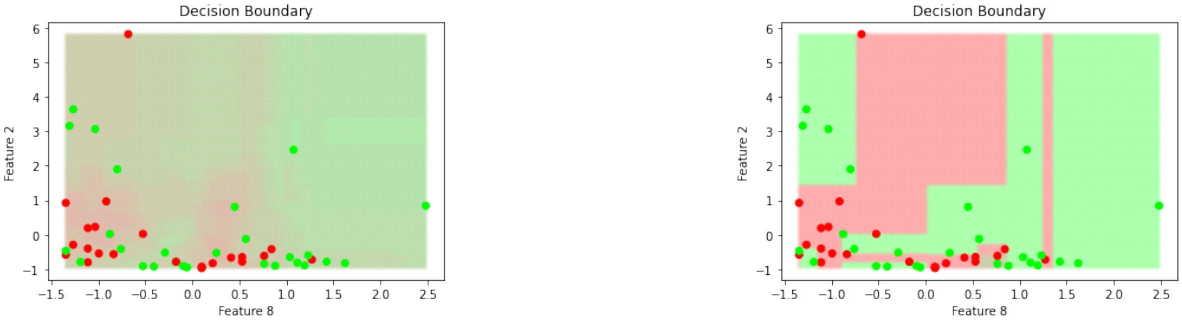
Figure 6: Feature importance for the Diabetes data set before and after feature scaling



(a) KNN decision boundary plot using Hamming distance

(b) DT decision boundary plot using Gini index cost

Figure 7: Decision boundary plots for KNN and DT on the Hepatitis data set



(a) KNN decision boundary plot using Manhattan distance

(b) DT decision boundary plot using Gini index cost

Figure 8: Decision boundary plots for KNN and DT on the Diabetis data set