COMP 551 Assignment 1

Shelley Xia, Mingze Li, Yiwei Cao

February 8, 2022

Abstract

In this project, we investigated the performance of two machine learning models, K-Nearest Neighbors and Decision Tree, across two benchmark datasets. In particular, we examined the impact of different cost functions, distance functions, and stopping criteria on accuracy. We also experimented with cross-validation, hyperparameter tuning, and different methods of feature selection in an attempt to achieve a better model fit.

For both datasets, we found that the KNN model achieved higher accuracy than the decision tree model. Specifically, the highest accuracy KNN gives on the first dataset is 91.67% while the decision tree model yields 87.5%. With regards to the diabetes dataset, the two models did not differ to such an extent: the KNN model has an accuracy of 70.43%, while the decision tree model yields 68.12%. Overall, we concluded that the KNN model performed better than the decision tree model on both datasets.

Function-wise, we learnt that using the cosine distance function yields slightly better performance than the manhattan and euclidean distance functions for our KNN model, and the misclassification cost function produces a higher accuracy than the Gini index and entropy cost functions for our decision tree model. In addition, the 'max_depth' stopping criterion outperformed other stopping criteria in our tree. Moreover, we discovered that performance varies significantly with different k values for the KNN model. On the other hand, tuning the L value in L-Fold cross-validation has a lesser impact on the test accuracy rate. Implementing feature selection does not significantly improve performances.

1 Introduction

Different machine learning models have long been used for medical diagnosis. In this project, we examined the performance of two machine learning models, K-Nearest Neighbors and Decision Tree, on diagnosing patients with Hepatitis and Diabetic Retinopathy. To train and test our models, we used two benchmark datasets provided by the UCI Machine Learning Repository [1][2]. The datasets have been used effectively by previous researchers to investigate different applications of machine learning, including data mining [3], Support Vector Machines [4], and Neural Networks [5].

In our experiments, we examined the effects of hyperparameter tuning of key hyperparameters, cross-validation, feature selections, and different cost and distance functions on both the Decision Tree and KNN models. We found that the performance of our KNN model peaks when we tuned the number of nearest neighbor restraints in KNN (referred to as k below) as $k = 3$ for the hepatitis dataset and as $k = 12$ for the diabetes dataset. Remarkably, the model performance of the hepatitis dataset drops significantly with k greater than 6, suggesting underfitting when the number of nearest neighbor restraints gets ineffectively large. However, for the diabetes dataset, the correlation of k's value with the performance varies greatly as we tune the value of L in L-fold cross-validations. We also investigated the effects of hyperparameter tuning of the maximal depth restraint, referred to as d below, for our decision tree model. We found that the model performance peaks at $d = 3$ for the hepatitis dataset. Similar to the KNN performance on the same dataset, the performance drops significantly when we set d to higher values. For the decision model, our results show that $d = 8$ yields the best model performance for the diabetes dataset.

We also learnt that the selection of distance functions and cost functions for the KNN model and decision tree model, respectively, has an impact on the model performance, although not as significant as the selection of key hyperparameter values. In general, the cosine function yields slightly better model performance than the Manhattan and euclidean function for our KNN model, and the Gini-index cost function produces slightly better model performance than other cost functions. We found no significant impact on model performance with the addition of feature selection.

2 Datasets

2.1 Data Cleansing and Basic Statistical Analysis

The first dataset contains 19 features regarding the physiological characteristics of hepatitis patients, as well as information on whether each patient lives or dies. In the initial data processing phase, we found that 75 out of 155 instances in the dataset contained at least one missing feature. To deal with this problem, we tried two different data cleansing strategies. The first is simply removing instances with missing features. The second strategy involves data imputation, where missing values are filled with either the median (in the case of categorical features) or the mean (in the case of numerical features) of the feature. In the end, we decided to adopt the first method. Upon examining the class distribution, we found that 67 instances were classified as class 2 (patient lives) while 13 instances were classified as class 1 (patient dies), after data cleansing.

The second dataset contains 18 features relevant to the diagnosis of diabetic retinopathy (DR). We did not find any missing data in this set. Upon examining the class distribution, we found that out of 1151 instances, 611 were classified as class 1 (containing signs of DR) while 540 were classified as class 0 (no signs of DR).

2.2 Data Normalization

After properly cleaning the datasets, we then proceeded to data pre-processing, a crucial step in machine learning. We noticed that the values of the categorical features in the first dataset are 1's and 2's, while those of the continuous features range from 0 to 100. Without normalization, intuitively, the continuous features would be given much higher weights than the categorical ones, resulting in undesired behaviors. As suggested by Singh, D., & Singh, B. (2020) [6], data normalization could ensure that each feature has a numerically equal contribution to the classifiers. Thus, we hypothesized that data normalization would be beneficial in our case. Specifically, we chose Z-score Normalization as our primary method for its simplicity and consistency with original data [7]. We applied Z-score normalization to all continuous features of both datasets. Moreover, we adjusted the values of categorical features down by 1 such that most of the values in the datasets are around the interval [0,1]. With proper testing, results suggest normalization does result in better model performance.

2.3 Ethical Concerns

Through analyzing the datasets, we acknowledged that there were potential ethnic concerns such as biases by gender and age groups. In the first dataset after cleaning, we found that there are 11 patients with 'sex' belonging to class 1 while there are 69 patients whose 'sex' belongs to class 2. The highly asymmetrical distribution in gender may result in failures in predicting class 1 patients' situations. Similarly, most patients in the dataset were in their 30s and 40s, which may result in biases in our models.

3 Results

3.1 Feature Selection

For the first dataset, we performed chi-square tests between each categorical feature and the class. Only the features that significantly correlate with class (those with $p<0.05$) were kept. We observed that doing so did not have a significant impact on the performance of the the KNN model or decision tree model. We further tried to examine the relationship between the numerical features and the class through performing two-sample t-tests, and only kept the features for which there is a significant difference between the mean of the instances classified as class 1 and the mean the instances classified as class 2. However, eventually, we decided to keep all numerical features since this feature selection strategy led to a decrease in prediction accuracy.

For the second dataset, we noticed that some features seem to contain repetitive information, based on the dataset description [2]. Therefore, we initially computed the correlation between each feature and discarded the ones for which the correlation is higher than 0.9. Similar to previous attempts, this decreased

the accuracy of the model prediction rather than increasing it. Thus, we did not include the feature selection step in our final version of the code.

3.2 Hyperparameter Tuning & Cross-validation

We tested the impact of different K values in the KNN model and different L values in L-fold cross-validation on model performance (L=2,3,4,5,6,10,20, K = 1, ..., 20). For the hepatitis dataset, the accuracy peaks at approximately K=3 (depending on the distance function used). At K > 6, the accuracy drops significantly, indicating underfitting. For the diabetes dataset, however, the optimal k value varies with different L values in L-fold cross-validation. Based on our limited observations from the diabetes dataset, we hypothesize that the optimal K value tends to be larger when L value is smaller, whereas when L value gets bigger, the optimal K value tends to get smaller. Across all tested L values and all K values, the highest accuracy is generated with K = 12, with 2-fold cross-validation and cosine as the distance function.

We further investigated how tuning the maximal depth restraint in the decision tree model can influence the model performance (d = 1, ..., 20). The optimal max depth value d = 3 leads to the highest accuracy rate of 87.5% in the hepatitis dataset. When the cost function is the misclassification cost, the performance drops when we tuned the maximal depth value d to higher values. Interestingly, when the used cost function is Gini-index, the performance maintains the high accuracy of 87.5% despite when value we tuned d for. For the diabetes dataset, the patterns of how the accuracy changes with the maximal depth value d vary greatly, depending on the different cost function employed. When the cost function is Gini-index, the model performance peaks at d = 8. However, when misclassification is used as the cost function, performance peaks at d = 5.

For the KNN model, we investigated the effects of different distance functions on model performance. In particular, we implemented and tested 3 distance functions: Manhattan distance, Euclidean distance, and Cosine similarity. Their formulas are as follows:

$$D_{euclidean}(x, y) = \sqrt{\sum_{d=1}^{D}(x_d - y_d)^2} \tag{1}$$

$$D_{manhattan}(x, y) = \sum_{d=1}^{D}|x_d - y_d| \tag{2}$$

$$D_{cosine}(x, y) = \frac{x^T * y}{\|x\|\|y\|} \tag{3}$$

While Manhattan distance only provides horizontal and vertical comparisons [8], Euclidean distance and Cosine similarity can account for multidimensional vector space. We surmised that the cosine similarity and the euclidean distance would excel, but this was not supported by empirical data. Results upon testing suggested that the three distance functions have similar effects. As for the diabetes dataset, we observed that the cosine similarity renders a slightly better model performance, compared to the manhattan and euclidean distances. We also noticed that the manhattan distance slightly won over the euclidean distance. For a fixed k value determined by hyperparameter tuning, model accuracies for both datasets based on different distance functions are summarized in Table 1.

	Manhattan	Euclidean	Cosine Similarity
Hepatitis (k=3)	91.67%	91.67%	91.67%
Diabetes (k=12)*	69.86%	68.99%	70.43%

Table 1: Model accuracies based on distance functions.
*Note: value corresponds to the cost function with the highest accuracy.

Concerning the decision tree model, we examined the behaviors of our model predictions corresponding to both different cost functions and distinct stopping criteria. Here we present 3 common cost functions: the

misclassification cost, the entropy cost, and the Gini-index. Their formulas are as follows:

$$\text{Misclassification cost: } cost(\mathbb{R}_k, \mathcal{D}) = \frac{1}{N_k} \sum_{x^n \in \mathbb{R}_k} \mathbb{I}(y^n \neq w_k) = 1 - p_k(w_k) \tag{4}$$

$$\text{Entropy cost: } \frac{N_{left}}{N_{node}} cost(\mathbb{R}_{left}, \mathcal{D}) + \frac{N_{left}}{N_{node}} cost(\mathbb{R}_{right}, \mathcal{D}) \tag{5}$$

$$\text{Gini-index: } cost(\mathbb{R}_k, \mathcal{D}) = \sum_{c=1}^{C} p(c)(1 - p(c)) \tag{6}$$

We found that the entropy cost and Gini-index functions give better model performance than the misclassification cost. We postulate this is due to lack of homogeneity in the subtrees. Overall, we conclude that the Gini-index works the best in terms of model prediction in the decision tree model. For a fixed max_depth determined by hyperparameter tunning, model accuracies for both datasets based on different cost functions are summarized in Table 2.

	Misclassification	Entropy	Gini-index
Hepatitis (max_depth=3)	79.17%	83.33%	87.5%
Diabetes (max_depth=8)	58.55%	66.09%	68.12%

Table 2: Model accuracies based on cost functions.

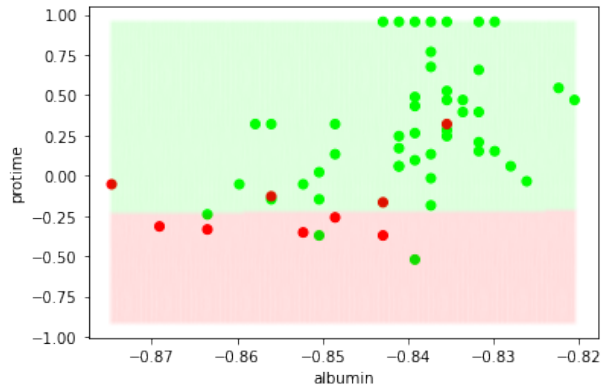
Similarly, we implemented 3 conventional stopping criteria: 'max_depth', 'small_cost', and 'min_leaves'. 'Max_depth' criterion (M.D.) forces a decision tree to stop splitting after it reaches a certain max_depth. 'Small_cost' (S.C.) criterion allows the tree to stop when the current cost of the current node is small enough. For our purpose, we set the threshold cost to be 0.05. 'Min_leaves' (M.L.) criterion instructs the tree to stop splitting when the number of instances in potential subtrees are too small. Our implementation allows users to specify 3 stopping criteria as arguments to the fit function. The decision tree model will stop splitting when either of these 3 stopping criteria is met. If no criteria is specified, then the default stopping criterion is 'max_depth' (M.D.). Note that there can be a minimum of 1, and a maximum of 3, stopping criteria in play. We found that 'max_depth' (M.D.) gave the best performance among 3 criteria. When we fix the max_depth given by hyperparameter tuning and the cost function to be the Gini-index, the results of different combinations are demonstrated in Table 3.

	M.D.	S.C.	M.L.	M.D.+S.C.	M.D.+M.L.	S.C.+M.L.	M.D.+S.C.+M.L.
Hepatitis (max_depth=3)	87.5%	83.33%	83.33%	87.5%	87.5%	83.33%	87.5%
Diabetes (max_depth=8)	86.12%	64.93%	64.93%	68.12%	68.12%	64.93%	68.12%

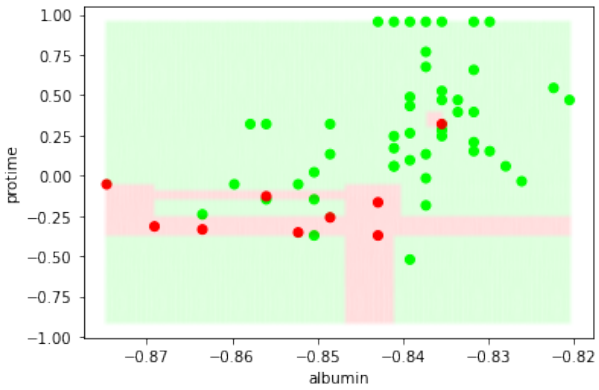
Table 3: Model accuracies based on stopping criteria.

3.3 Decision Boundary

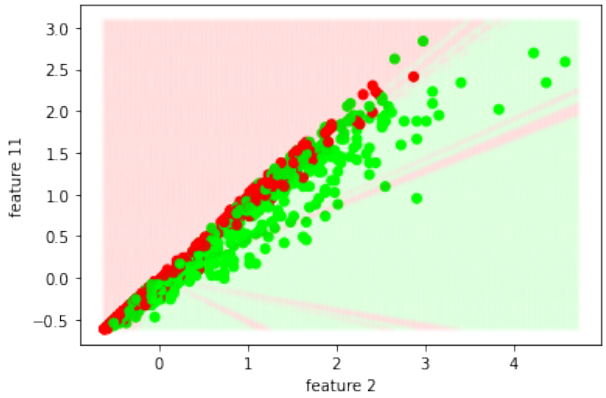
The decision boundaries for each model are shown in Figure 1. The features being plotted are selected based on their degree of correlation with the class. Since both datasets contain multiple features but only 2D graphs can be plotted, these graphs may not be the best representation of the actual decision boundaries. Nonetheless, some potentially useful observations could be made. For example, in Figure 1(b), a single data point has a decision boundary around it, indicating the possibility of overfitting. On the other hand, the decision boundary shown in Figure 1(a) may suggest the possibility of underfitting.



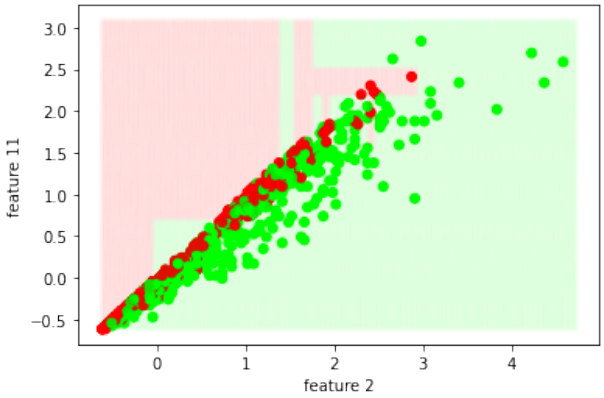
(a) Hepatitis Dataset, KNN



(b) Hepatitis Dataset, Decision Tree



(c) Diabetes Dataset, KNN



(d) Diabetes Dataset, Decision Tree

Figure 1: Decision boundaries for each dataset and each model.

4 Discussion and Conclusion

In the medical realm, machine learning models have been widely used for diagnosis given gathered information of patients [9]. In our work, we trained both the KNN model and the decision tree model on two datasets from the UCI machine learning repository [1]: the hepatitis dataset and the diabetes dataset. We then investigated the effects of different parameterized functions and various values of hyperparameters on the resulting model accuracies. We found that the KNN model beats the decision model performance-wise. We also observed that the cosine similarity as the distance function in the KNN model gives the best model prediction out of all tested distance functions. It is also worth pointing out that changes in the hyperparameter k, the number of nearest neighbors, has a great impact on the results. As k increases, the accuracy initially increases and then gradually drops due to underfitting. With regards to the decision tree model, it was evident that the 'max_depth' stopping criteria produces the best model performance. Cost-function-wise, we saw the highest accuracy was given by the Gini-index. Furthermore, our attempt of feature selection did not improve the model fit and instead decreased it. We acknowledge that due to lack of domain knowledge, we are unable to perform further feature analysis to generate a more representative selection of useful features. Our future work would involve improvements of future selection and data preprocessing through inquiring domain experts. We hope to increase our model fit with a better selection and normalization of features. Lastly, we plan to seek reasons why the KNN model scores much higher than the decision tree model in these datasets.

Statement of Contributions

All three authors contributed to writing the manuscript and writing the code. In particular, Shelley focused on data normalization, distance functions, cost functions and stopping criteria; Mingze focused on cross validation and hyperparameter tuning; Yiwei focused on data cleansing, feature selection, and plotting

decision boundaries.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] Bálint Antal and András Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, 60:20–27, 2014.

[3] Heitor S Lopes. An ant colony based system for data mining: Applications to medical data, Jan 2001.

[4] E. Smirnov, I. Sprinkhuizen-Kuyper, and G. Nalbantov. unanimous voting using support vector machines: Semantic scholar, Jan 1970.

[5] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.

[6] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, 2020.

[7] Zixuan Zhang. Understand data normalization in machine learning, Aug 2019.

[8] Test your skills on k-means clustering algorithm, May 2021.

[9] Thomas Bocklitz, Melanie Putsche, Carsten Stüber, Josef Käs, Axel Niendorf, Petra Rösch, and Jürgen Popp. A comprehensive study of classification methods for medical diagnosis. *Journal of Raman Spectroscopy*, 40(12):1759–1765, 2009.