COMP 551 Mini-Project 1

Yingjie Xu YJ.XU@MAIL.MCGILL.CA (260862481)
Hengxian Jiang HENGXIAN.JIANG@MAIL.MCGILL.CA (260830557)
Tianhao You TIANHAO.YOU@MAIL.MCGILL.CA (260830663)

Abstract

In this project, we investigated the performance of two machine learning methods — K-Nearest-Neighbors and Decision Tree. Each method is implemented and tested by two datasets separately to evaluate the accuracy of the outcomes. Normalization is adopted to improve the predicting model. After comparing the outcome accuracy, we found that KNN predicts more accurately when the dataset contains more numerical features and the Decision Tree can be easily overfitted when dealing with small datasets. We found that the Decision Tree approach had almost the same accuracy as K - Nearest-Neighbor in the Breast-Cancer dataset and the Decision Tree approach achieved better performance than K-Nearest-Neighbor in the hepatitis dataset. By analyzing these two datasets, we conclude that malignant breast cancer would generally have larger values in features that describe the breast cancer than the benign ones, and patients who died from hepatitis would generally suffer from more symptoms than people who live.

1. Introduction

Machine learning is a method used to make predictions or decisions by building mathematical models based on training data. To generate good predictions of data based on decision boundaries, K-nearest-neighbors(KNN) and Decision Tree(DT) are investigated to implement the model.

Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. The task given by the project is to analyze the dataset Breast-Cancer-Wisconsin created by Dr. William H. Wolberg and the hepatitis dataset. The goal of this project is to predict whether the diagnosis result of the breast cancer is malignant or benign and to predict the living chances(divided into two classes: live or die) by analyzing the features given in these two datasets.

By analyzing the features of the Breast cancer dataset, we found that the higher the value of those features describing the tumour in the dataset, the higher the probability that the tumour is malignant. As for the hepatitis dataset, we found that patients with the symptoms described had a higher probability of dying from hepatitis.

Some noticeable discoveries when implementing these two model to the datasets:

Model 1. KNN predict more accurately when the dataset contains more numerical features. 2. Decision tree can be easily overfit when dealing with small datasets.

Dataset 1. Malignant breast cancer would generally have larger values in features that describe the breast cancer than the benign ones. 2. Patients who died from hepatitis would generally suffer from more symptoms than people who live.

The key take away in this project was that by adjusting and manipulating the parameters of each learning model, we have a better understanding of the impact of changing those parameters would have on the model and the interactive relationship between different parameters.

2. Datasets

2.1. Summary

For this mini-project, we are given two datasets. The first one is Wisconsin diagnostic breast cancer data. Ten real-valued features are computed for each cell nucleus from a digitized image of a fine needle aspirate of a breast mass. Those values are used to predict the diagnosis result (M = malignant, B = benign). The second dataset is the hepatitis dataset. The information about the patient is given including age, sex and a lot of other factors. We use that information for predicting the class (live or die).

2.2. Data processing

For both datasets, we loaded all the data into pandas and then perform the data cleaning. We removed all the instances with missing features and removed all the duplicated data.

2.3. Training, validation and testing sets

After initial data processing, we have 675 samples for the breast cancer dataset which is a relatively larger dataset, so we decided to use 70% for training, 15% for validation and 15% for testing. By contrast, we got 80 samples remaining for the hepatitis dataset, which is a relatively small dataset. Hence, we decide to give more percentage of the dataset for testing and validation. 60% of samples are used for training, 20% for validation and 20% for testing for our machine learning models.

2.4. Basic statistics

For the breast cancer Wisconsin dataset, the first thing we did is taking the mean over all the columns except class, since it is the reference that we divide the data into different groups. Then we divide the whole data set into two groups by class 2 and class 4. We can easily find the difference of the mean value in each column between two groups in the bar plot (Figure 1). It is obvious that the mean of class 2 is much lower than the mean of class 4 in each column, which indicates most class 4 would have larger values. (In Figure 1, the x legends correspond to the actual column names in the dataset)

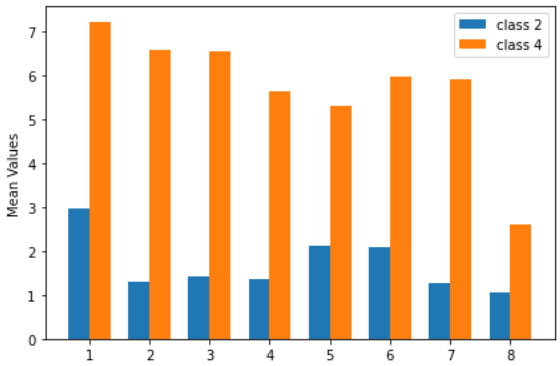

Figure 1: breast cancer dataset

For the hepatitis dataset, since most of the features are categorical data with yes or no values. We decided to take a look at the mode value for each of those categorical features. In the plot, class 1 means die and class 2 means live. It seems that patients who die generally suffer from steroids, malaise, liver firm and some other health problems. By

looking at this plot (Figure 2), we could see that people who died from hepatitis would generally suffer from more symptoms than people who live. (In Figure 2, the x legends correspond to the actual column names in the dataset)

Figure 2: hepatitis dataset

2.5. Ethical concerns

Both datasets are healthcare-related data. If we want to use that information, there would have to consent the patient associated with that data in order to proceed. Moreover, even if we got consent, we still need to use those data carefully and without disclosing any personal information to researchers. In the consent, we should let the patient know what kinds of information and data could be used and be as detailed as possible.

3. Result

3.1. Dataset 1: Breast Cancer Wisconsin Data Set

3.1.1. K-NN

The task is to apply the K-NN model for the given dataset and try to find the highest accuracy with a particular K-value which is a hyperparameter. In the beginning, we checked for the data type in each column. Since there is no categorical data in this data set, we can directly apply euclidean distance or manhattan distance to the K-NN model. As mentioned before, we chose to slice the data set with proportion 70%, 15%, 15% as our training set, validation set and testing set. We constructed an evaluation program. In order to get an ideal K value, we use the validation set to tune the hyperparameter. We drew the accuracy result for different K as a graph (Figure 3). Through the graph, we can easily find the approximate accuracy over different K values. By looking at the graph, when K = 3, the validation set accuracy reaches the highest point and the value is small so that the model won't be too complex. Hence, we chose K = 3 as our ideal K-value. The resulting accuracy obtained from the testing set is 96.1%.

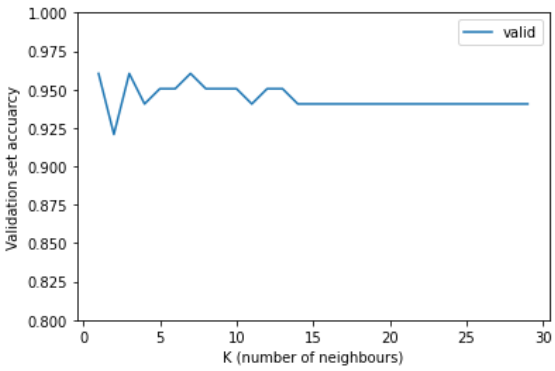

Figure 3: KNN validation set accuracy

3.1.2. Decision Tree

During the construction of the decision tree, we first sliced data with 70%, 15%, 15% as our training set, validation set and testing set. After this, we constructed an evaluation program with the validation set. To find the highest validation set accuracy, we plot the accuracy at each max depth (Figure 4). It is obvious that the accuracy reached the highest point at max depth equals to 6. The

resulting accuracy obtained from the testing set is 96.1%.

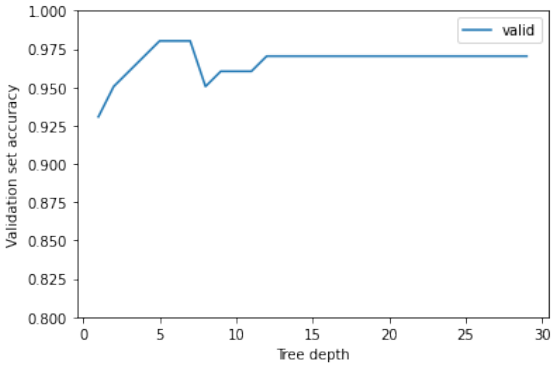

Figure 4: DT validation set accuracy

Decision boundary: We chose uniformity of cell size and clump thickness to plot the decision boundary (Figure 5). From the figure, we could see that when the uniformity of cell size is smaller than 4 and the clump thickness is smaller than 6, most of the patients would have benign cancer.

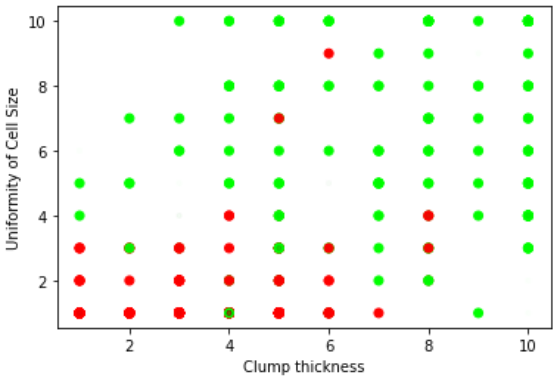

Figure 5: Decision boundary

3.2. Dataset 2: Hepatitis Dataset

3.2.1. K-NN

Since it is a small dataset, we chose to use proportion 60%, 20%, 20% for dividing the set. In the second dataset, different features have different ranges. To avoid the scaling problem of K-NN, We standardized the data by using the min-max normalization. Since the dataset is combined with both categorical and numerical data. We divided two types of data and converted them into int type and float type respectively.

After conversion, we applied hamming distance to the categorical data and Euclidean distance to the numerical data. However, it didn't give a better performance. Hence, we decided to continue with the Euclidean distance approach. Similarly, we use the validation set to find the ideal K value. The graph (Figure 6) shows that the accuracy reaches the highest at K = 3, and the final accuracy we got from the testing set is 81.2%.

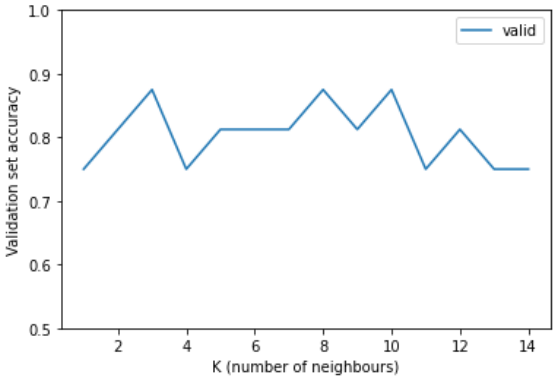

Figure 6: KNN validation set accuracy

3.2.2. Decision Tree

First of all, the proportion of each set is different from the previous dataset. The training set, validation set and testing set occupied 60%, 20%, 20% respectively. By applying different max depth to the validation set, we got the validation set accuracy plot (Figure 7). It is interesting to notice that the accuracy seems to be constant when K > 7, this might be related to the size of the dataset.

When we chose a quite large number as our max depth, the program ends with overfitting problems. Hence, it is safe to choose 4 as our max depth because it has the highest accuracy. The final accuracy from our decision tree by using the testing set is 93.8%.

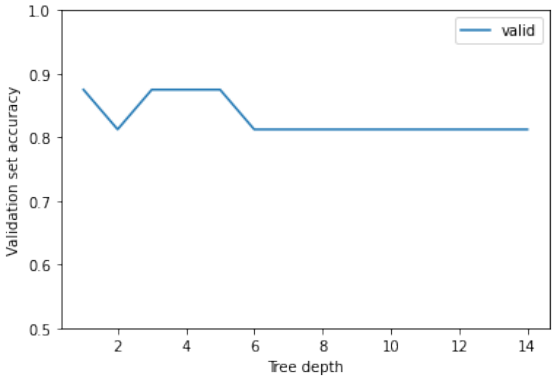

Figure 7: DT validation set accuracy

Decision boundary: We chose steroids and antivirals to plot the decision boundary (Figure 8). There are only 4 points on the plot because both steroid and antivirals are categorical data. By looking at the graph, we could conclude that if a patient has antivirals, then he is more likely to live.

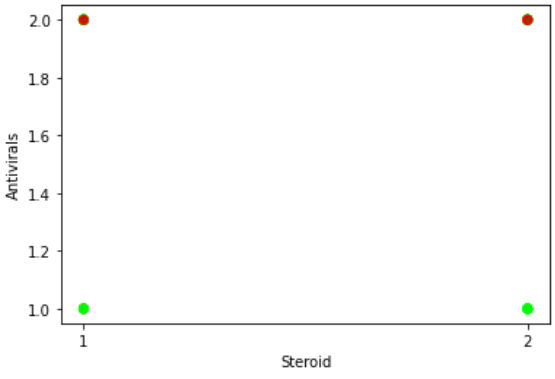

Figure 8: Decision boundary

4. Discussion and conclusion

4.1. Key takeaways

The key take away in this project was that by adjusting and manipulating the hyperparameters of each learning model, we have a better understanding of the impact of changing those parameters would have on the model and the interactive relationship between different parameters. And after evaluating the performance and accuracy for both machine learning model, we found that the decision tree gets a great performance for predicting the datasets with many categorical features.

4.2. Future investigation

Due to the limited size of the second dataset, the performances of both models are not as well as the first dataset. We may need to find more data in the future investigation. Selecting a model to predict a dataset is not easy, we may need to analyze the advantages and disadvantages of the models before applying them. We could possibly draw more dicision boundary plots on different combination of features and get a better understanding of the dataset.

5. Statement of Contributions

Hengxian was responsible for implementing the Decision Tree for the first dataset, and experimented with KNN and Decision Tree model and wrote the abstract, introduction and discussion and conclusion section in the report. Yingjie was responsible for data cleaning and processing, implemented KNN and Decision Tree for the first and the second dataset respectively, and wrote the dataset section in the report. Tianhao was responsible for data cleaning and processing, implemented KNN for both datasets and wrote the result section in the report.