Comp551 Project1 Writeup

Sibo Yang(260906371), Rui Song(260890776), Jeff Zhang(260840033)

January 2021

Abstract

In this project we first implement two machine learning models, which are the K-nearest-neighbor algorithm and the decision tree algorithm. Next, we compare the performance of these two algorithms on two benchmark datasets: the Wisconsin breast cancer dataset and the Hepatitis dataset. Based on our results, we conclude that the KNN approach overall achieves better performance than decision tree. Further, in order to obtain a set of parameters for optimizing the models, we examine how different hyper-parameters, such as the number of neighbors and the maximum depth of decision tree, could affect test accuracy by plotting accuracy against the hyper-parameter values. To additionally reduce generalization error, we implement cross-validation in our hyper-parameter testings.

Intruduction

Guided by the conceptual knowledge of KNN and decision tree given in class, we launch an investigation into the detailed implementations of these algorithms as well as how the performance of these algorithms could depend on varying hyper-parameter values. Our investigation is based on two open-source datasets. The first one, the Wisconsin Breast Cancer dataset, contains 699 instances of data vectors that each represents the characteristics of an image of breast mass. The second dataset, which is the Hepatitis dataset, consists of 155 instances of data vectors that represent hepatitis patient's physical conditions. Both the breast cancer dataset and the hepatitis dataset are drawn from the University of California website. [3] [2] In the end, we also compare the performances of KNN against that of the decision tree to identify the more accurate approach, given the two real-world datasets.

Datasets

We import the dataset with Pandas and for both datasets, and removed instances with missing or malformed features. Since we are going to use the KNN which is a distance based algorithm in this assignment, we rescaled the distribution of features to [0, 1] to make sure our model will not be affected by the magnitude of features.

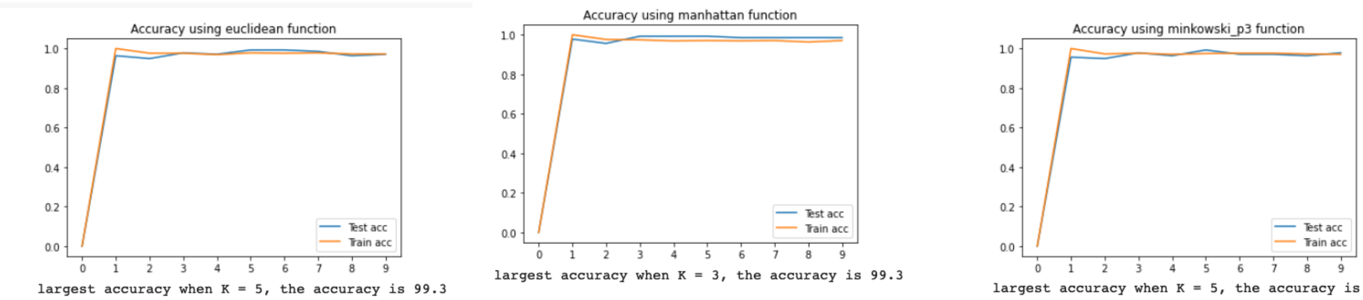
Breast Cancer dataset

This dataset contains 699 instances with 9 attributes, each representing characteristics of the cell nuclei in a image of FNA of a breast mass, and they are classified as either malignant diagnosis or benign diagnosis. After cleaning the dataset, 444 instances are classified as benign diagnosis (class=2), 239 instances are classified as malignant diagnosis (class=4). Since the 'color' option can only accept array of len=3 or 4, we also change the 4 and 2 to 2 and 1 respectively, for the convenience of drawing decision boundaries.

Hepatitis dataset

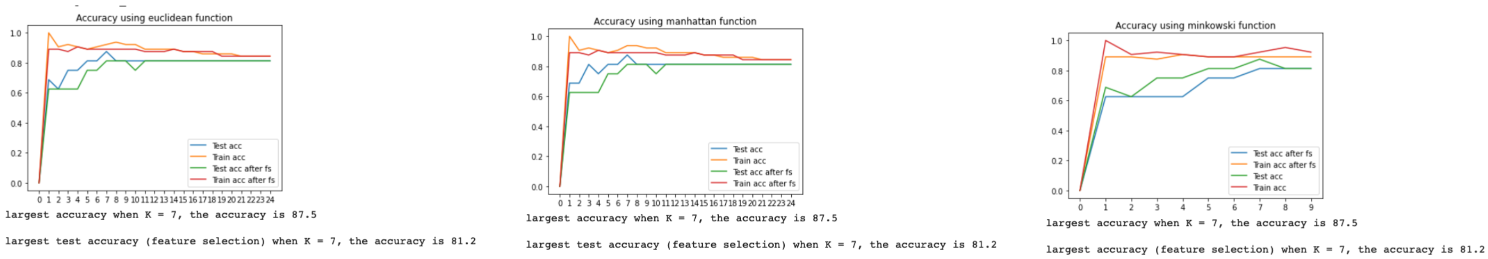
This dataset contains 155 instances of hepatitis patient's physical condition with 19 attributes, and they are classified as either DIE or LIVE. After cleaning the dataset, 13 instances are classified as DIE (class=1), 67 instances are classified as LIVE (class=2). We can see that this dataset is highly imbalanced in terms of class distribution. But it will not affect the accuracy, since even if we classify every test instance as LIVE, we still get a 84% accuracy. This assignment asks us to evaluate the performance using accuracy, so we didn't make some changes to our experiment to deal with the imbalance, we will discuss about it further in the discussion section.

Results



```
largest accuracy when K = 5, the accuracy is 99.3
largest accuracy when K = 3, the accuracy is 99.3
largest accuracy when K = 5, the accuracy is 99.3
```

Figure 1: KNN performance on breast cancer dataset using difference K values and distance functions



```
largest accuracy when K = 7, the accuracy is 87.5

largest test accuracy (feature selection) when K = 7, the accuracy is 81.2
```

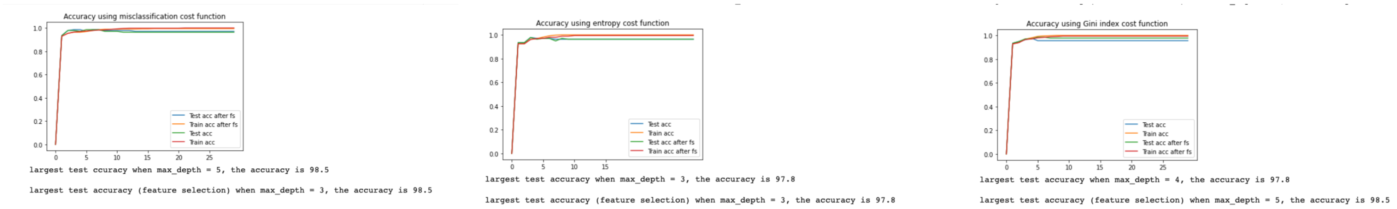
```
largest accuracy when K = 7, the accuracy is 87.5

largest test accuracy (feature selection) when K = 7, the accuracy is 81.2
```

```
largest accuracy when K = 7, the accuracy is 87.5

largest accuracy (feature selection) when K = 7, the accuracy is 81.2
```

Figure 2: KNN performance on hepatitis dataset using difference K values and distance functions

As shown in Figure 1 and Figure 2. We first test how different K values and cost functions can affect the performance of KNN of on both datasets. We found that as K increases, the model underfits the data and the training accuracy becomes lower. We also found that by applying different distance functions, the accuracy of model on both datasets didn't have significant changes. Then by experimenting with different K values, we found that in the breast cancer dataset,K = 5 will give the best performance which is 99.3% and in the hepatitis dataset, K = 7 will give the best performance which is 87.5%. In the hepatitis dataset which consists of 80 valid instances with 19 attributes, we believe that the high dimension and the small size of dataset will result in low accuracy of the model, so we decided to select 4 features from the 19 features using the Chi-square test with the help of SciKit learn library. However, the interesting observation is that using the dataset after feature selection for training, the accuracy is lower than before which can be seen in Figure 2.

Figure 3 and Figure 4 shows the result of how different maximum tree depths and cost functions can affect the performance of decision tree on both datasets. We observed that in both datasets, test accuracies are consistently lower than training accuracies because the model has the tendency to overfit the training data as maximum tree depth grows. In the breast cancer dataset, models that used classification function achieved a slightly higher accuracy than models that used other cost functions. however in the hepatitis

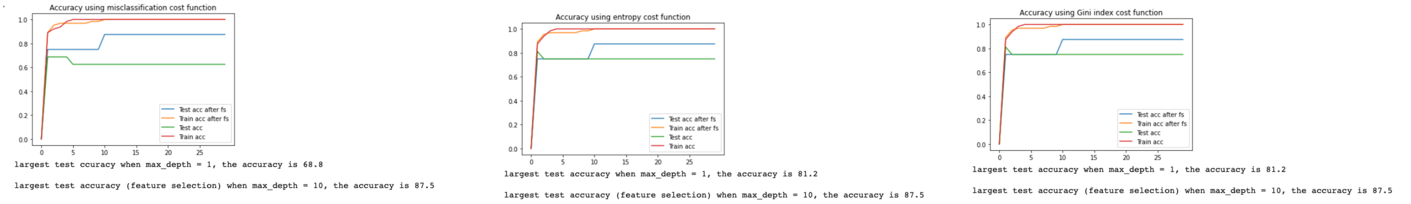


```
largest test ccuracy when max_depth = 5, the accuracy is 98.5

largest test accuracy (feature selection) when max_depth = 3, the accuracy is 98.5
```

```
largest test accuracy when max_depth = 3, the accuracy is 97.8

largest test accuracy (feature selection) when max_depth = 3, the accuracy is 97.8
```

```
largest test accuracy when max_depth = 4, the accuracy is 97.8

largest test accuracy (feature selection) when max_depth = 5, the accuracy is 98.5
```

Figure 3: Decision tree performance on breast cancer dataset using difference maximum tree depth and cost functions



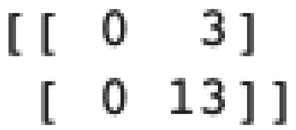
```
largest test ccuracy when max_depth = 1, the accuracy is 68.8

largest test accuracy (feature selection) when max_depth = 10, the accuracy is 87.5
```

```
largest test accuracy when max_depth = 1, the accuracy is 81.2

largest test accuracy (feature selection) when max_depth = 10, the accuracy is 87.5
```

```
largest test accuracy when max_depth = 1, the accuracy is 81.2

largest test accuracy (feature selection) when max_depth = 10, the accuracy is 87.5
```

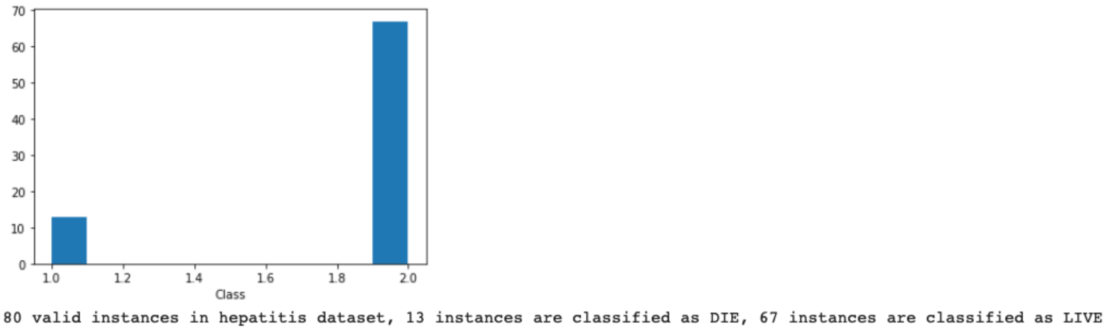
Figure 4: Decision tree performance on hepatitis dataset using difference maximum tree depth and cost functions

dataset, the models that used classification cost function have significantly lower test accuracies. Following the same idea mentioned in analysing the KNN model, we reduced the number of features from 19 to 4 to train the model. Then we observed that the performance of models had a increase comparing to model that used 19 features for training.



```
[[ 0  3]
 [ 0 13]]
```

Figure 5: KNN confusion matrix of hepatitis dataset



```
80 valid instances in hepatitis dataset, 13 instances are classified as DIE, 67 instances are classified as LIVE
```

Figure 6: KNN confusion matrix of hepatitis dataset

When ploting the decision boundaries, we use PCA to reduce the dimension of each dataset to component0 and component1, before PCA, we standardize the dataset's features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance[1].
Figure 7 and 8 are the variation of decision boundaries for breast cancer dataset for two distance function of KNN. Although the shape of boundary is a little bit different, we can still see that the boundary is

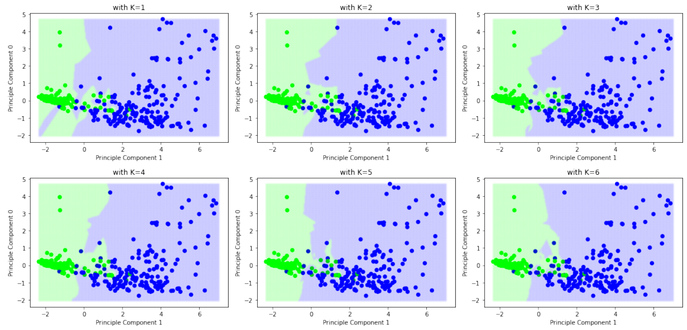


Figure 7: decision boundaries for breast cancer dataset by eucildean for KNN

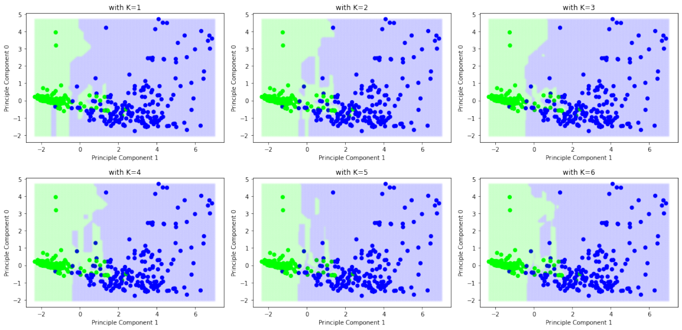


Figure 8: decision boundaries for breast cancer dataset by manhattan for KNN

$Component1 = 0$ in general. Manhattan causes more branches than Euclidean and the shape is more upright, we think it is caused by the complexity of the distance formula.

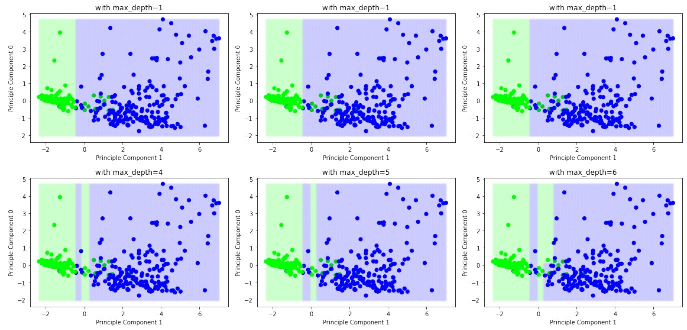


Figure 9: decision boundaries for breast cancer dataset by misclassification for Desicion Tree

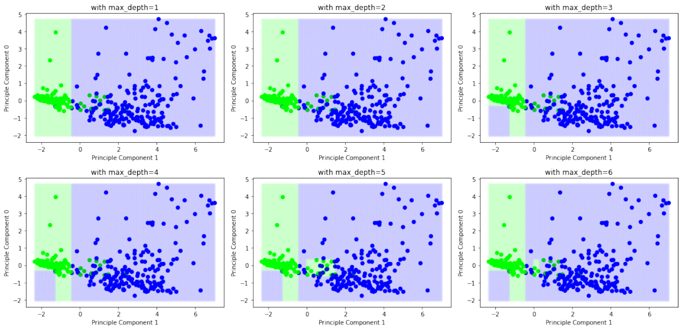


Figure 10: decision boundaries for breast cancer dataset by gini and entropy for Decision Tree

Figure 9 and 10 are the variation of decision boundaries for breast cancer dataset for different cost function of Decision Tree (since the decision boundary for entropy and gini is so similar,we only use one set of picture). Although the shape of boundary is a little bit different, we can still see that the boundary is $Component1 = -0.5$ in general and for misclassification, another boundary is $Component1 = 0$ and $Component1 = 0.5$.

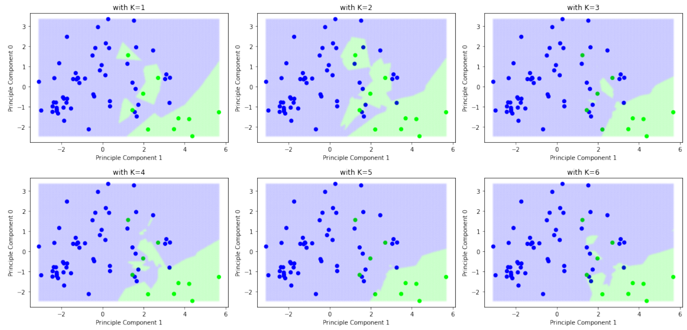


Figure 11: decision boundaries for hepatitis dataset by eucildean for KNN

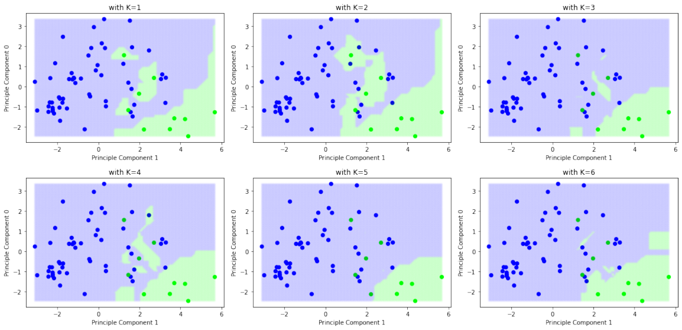


Figure 12: decision boundaries for hepatits dataset by manhattan for KNN

Figure 11 and 12 are the variation of decision boundaries for hepatitis dataset for two distance function of KNN. The shape of boundaries of these two function are similar in the end, and we can calculate that the boundary approximately is ($Component0 = \frac{1}{3}$ $Component1 - 2$)in general.

Figure 13, 14 and 15 are the variation of decision boundaries for hepatitis dataset for different cost function of Decision Tree. As we can see, the boundary for these three are different.For misclassification:

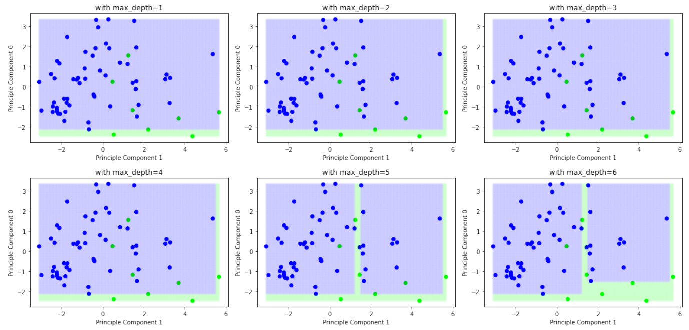


Figure 13: decision boundaries for hepatitis dataset by misclassification for Desicion Tree

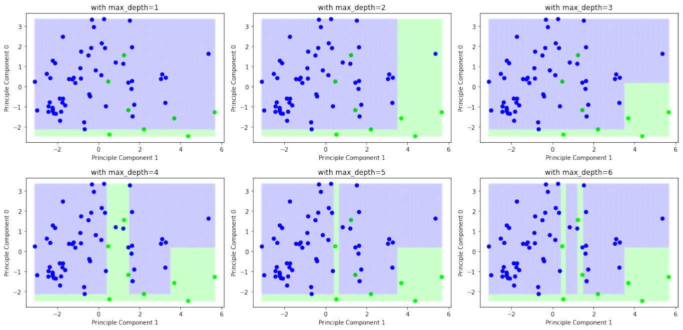


Figure 14: decision boundaries for hepatitis dataset by gini for Decision Tree

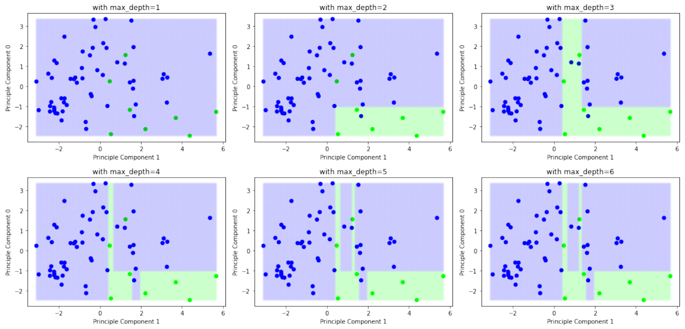


Figure 15: decision boundaries for hepatitis dataset by entropy for Decision Tree

$Component0 = -2$ when $Component1<1$, $Component0 = -1.2$ when $Component1>1$ and $Component1 = -1$ and $Component1 = -1.5$ and $Component1 = 5.5$; For gini: $Component0 = -2$ when $Component1<-3.5$, $Component0 = 0$ when $Component1> - 3.5$, $Component1 = -0.2$ and $Component1 = -0.5$ and $Component1 = 1$ and $Component1 = 1.5$; For entropy: $Component0<-1$ when $0.2<Component1<1.5$ and $Component1>2$, $Component1 = -0.2$ and $Component1 = -0.5$ and $Component1 = 1$ and $Component1 = 1.5$.

It seems that for hepatitis dataset, the performence of Decision Tree is clearly worse than KNN. What's more, we can find that KNN decision boundaries is similar to a linear boundary. So maybe, the classification method of Decision tree is not a good choice for this dataset. Maybe, we can find a way similar to linear regression to make a boundary.

Comparing two ways for both dataset, we can find that the decision boundaries are similar but KNN has more curves and Decision Tree has more upright line. We think this is because the classification methods for them – KNN is calculating the distance while Decision tree only use Boolean determination.

Discussion and Conclusion

We can observe from the results that hyper-parameters on graphs usually have the shape of an inverted parabola, such that there always exists a value where the performance would be optimal. This follows from the idea that large K would likely result in under-fitting, and low K values would likely result in over-fitting. The same applies for the decision tree where the larger tree depths cause over-fitting, since each partition holds very few data points. And similarly smaller tree depths result in under-fitting and bad precision. After experimenting how different models perform on these two datasets, we can conclude that KNN model will achieve a higher accuracy on breast cancer dataset, and decision tree model is able to reach a better performance on hepatitis dataset. As mentioned in the Dataset section, we observed that the class distribution of hepatitis dataset is imbalanced as shown in Figure 6, the accuracy is probably not a good way to evaluate the model performance on this dataset. So we evaluated the confusion matrix of test result as shown in Figure 5, and found out the precision is actually 0, which is very dangerous for disease prediction. So we can apply techniques to address the data imbalance issue to reach a better precision in the future. For Figure 7 to 15, we compare different boundaries for among different functions also between different models. We can find that both KNN and Decision Tree do well in breast cancer dataset, which implies that there should be specific independent features influence the label. However, for hepatitis dataset, KNN and decision tree get totally different models and KNN does better and the boundary is like some linear line. Maybe this implies that the features have relevance and effect the labels together.Therefore, for the future investigation, we may can try combine some features together and use some relation to set a model and make predictions well.

Statement of contributions

Everyone contributed to model implementing, experiment implementing, data processing and writing the writeup.

References

[1] Michael Galarnyk. PCA using python (scikit-learn). *Towards Data Science*, 2017.

[2] G Gong. *hepatitis dataset.* University of California, 1988.

[3] William Wolberg. *breast cancer dataset.* University of California, 1995.