Applied Machine Learning

Assignment 1 Report
COMP 551

Authors: *Ling Fei Zhang, Brandon Ma, Giordano Di Marzio*

Instructor:	*Yue Li*
Department:	Computer Science
Date:	October 6, 2022

Contents

1 Abstract

In this assignment, we investigated the performance of two machine learning models on two benchmark datasets consisting of multiple features and a binary classification. Feature reduction techniques, such as Random Forest and Principal Component Analysis, were applied to help fit K - Nearest Neighbor and Decision tree models. We observed that the model performed with higher accuracy while using Random Forest for the second dataset, but was more accurate with PCA for the first dataset. We also compared different cost functions for both learning models such as Euclidean distance and Manhattan distance for KNN along with Entropy and Gini Index for Decision Tree. Additionally, we found that the Decision Tree approach achieved worse accuracy than K - Nearest Neighbor for both dataset 1 and 2. Precision recall and Receiving Operating Characteristic curves were finally used to evaluate each of the models on the two principal component features (drawn from PCA) and the two selected features from both datasets (drawn from Random Forest). Analysis of these results would further encourage that the optimal dimensionality reduction was in fact Random Forest . A large difference in accuracy would also be brought to light between the two datasets following this analysis. Weighted KNN was also implemented and shown to have greater accuracy than KNN for the first dataset, while having lower accuracy than KNN for the second.

2 Introduction

The main task of the assignment was to implement K - Nearest Neighbor (KNN) and Decision tree models on two datasets. The first one involved the survivability of patients who had hepatitis, which contained 19 features for 155 patients. The second dataset contained 19 features for 1151 instances of images that characterized whether the patient showed signs of diabetic retinopathy. In order to plot decision boundary graphs for KNN and Decision Tree, we needed to reduce the dimensionality of the features . Random forest and Principal Component Analysis (PCA) were used to obtain the two most relevant features and the principal components respectively. These were generally observed to be Protime and Albumin or Bilirubin for the first dataset and the number of MA1 and exudates or distance for the second dataset (they could vary depending on how you split the data, because some features had very similar selection importance from the random forest) . These features coincide with related work, where for example, macular exudates are the primary cause for blindness (dataset 2) [1] and albumin is a biomarker for chronic liver disease (dataset 1) [2].

The most effective dimensionality reducer varied depending on the classifiers: PCA gave a higher for KNN while Random Forest was most effective for Decision Tree. Additionally, there was a significant difference in the performance of both models between the two datasets, but consistently, KNN held greater accuracy . Within these classifiers, multiple cost functions were considered such as Euclidean and Manhattan distance for the hepatitis data, along with Entropy and Gini index for the other. Our observations showed that the Euclidean and Manhattan distance both achieved relatively similar accuracy whereas the Entropy cost function seemed to obtain slightly higher accuracy compared to the Gini index. Then, weighted KNN's accuracy was shown to hold greater accuracy for the first dataset compared to both KNN and Decision Tree, but holds the weakest results for the 3 classifiers considering the second dataset.

3 Methods

The two main methods used to model the data were K - Nearest Neighbor (KNN) and Decision tree (DT). KNN is a supervised learning classifier which uses proximity to predict the grouping of an individual data point. Different amounts of neighbors K as well as different cost functions, such as Euclidean and Manhattan distances were tested, and best results were kept. Decision tree is also a supervised learning classifier which uses thresholds to split the data points. Different depths were explored to obtain the best splits. We also evaluated two different cost functions to measure information gain during the splitting of the data: Entropy and Gini index.

4 Datasets

The datasets were first filtered by missing values. In other words, any row with a missing value is deleted from the dataset. Next, we converted the type of the dataset to float, in order to do calculations among the values. Furthermore, we computed basic statistics for both datasets. This includes listing the minimum, maximum, mean, standard deviation and class correlation for each feature. Next, we also applied Random Forest to evaluate the importance of each feature. We ran the algorithm multiple times to then choose the two most important features on average. This is later put against the Principal Component Analysis and the method with the highest accuracy would be chosen. Finally, we decided to split the datasets as follows: dataset 1 was split into 60% training, 15% validation and 25% test while dataset 2 was split into 70% training, 10% validation and 20% test.

5 Results

Overall, it was observed that KNN held the highest accuracy.

In dataset 1, we saw high accuracy for both classifiers. KNN yielded slightly better results when using the Euclidean distance function to calculate the distance between points. KNN was also found to be working best when K had a value of around 9. Given these two tweaks, the accuracy of the prediction ranged from 80% to 85%. On the other hand, Decision Tree worked best at very low depths (1 or 2) using entropy as our cost function. The accuracy observed was around 85% to 90%. For KNN, it was determined that Principal Component Analysis had an edge over choosing the two best features. However, this was not the case for DT, where choosing the two best features proved to give better accuracy.

In dataset 2, we observed much lower accuracies for both models. In this case, KNN using the Euclidean distance function and a high K value of around 20 to 30 gave best results. Decision Tree yielded best results using the same parameters as the previous dataset, i.e. using the Gini index cost function and low depth of 1 or 2. As in dataset 1, KNN yielded better results with PCAs and DT had better outcomes with the top two features.

To see how well the model fit the data, we plotted four Precision Recall curves for dataset 1 and four Receiver operating characteristic curves for dataset 2.

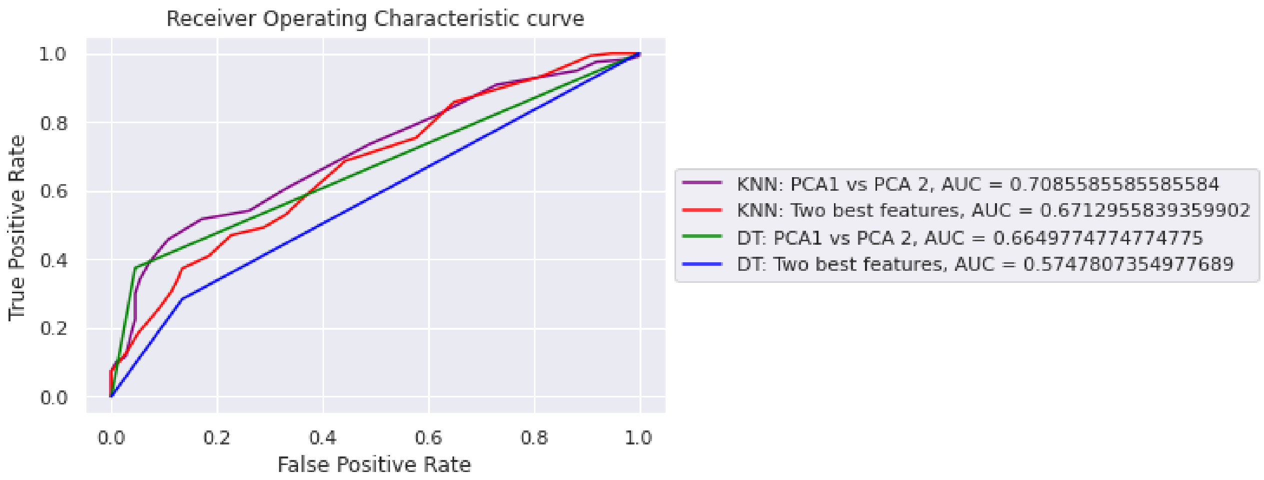

Figure 1: Receiver Operating Characteristic Curve

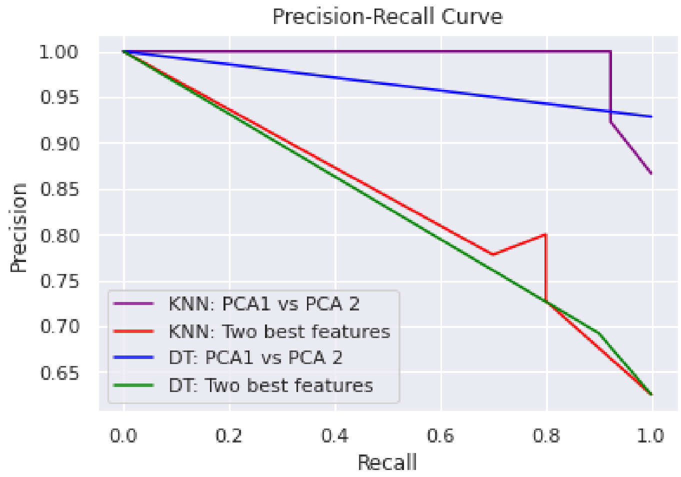

Figure 2: Precision-Recall Curve

In the plot, we compared KNN and DT with the top two features as well as with PCAs. An area under the curve (AUC) is also shown. As we can see, KNN using PCA as it's features has the highest AUC, indicating a better fit.

In order to improve the accuracy in our models for the second dataset, we decided to implement weighted KNN for K-Nearest Neighbors. For dataset 1, it was observed that the accuracy for KNN slightly increased by around 5% overall while for dataset 2, there was a slight decrease of about 5%.

6 Discussion and Conclusion

In dataset 1, KNN and decision tree both yielded high accuracy when classifying the test points. This was likely due to the fact that the data was more clearly separated unlike the data points in dataset 2, where the points were all clumped together.

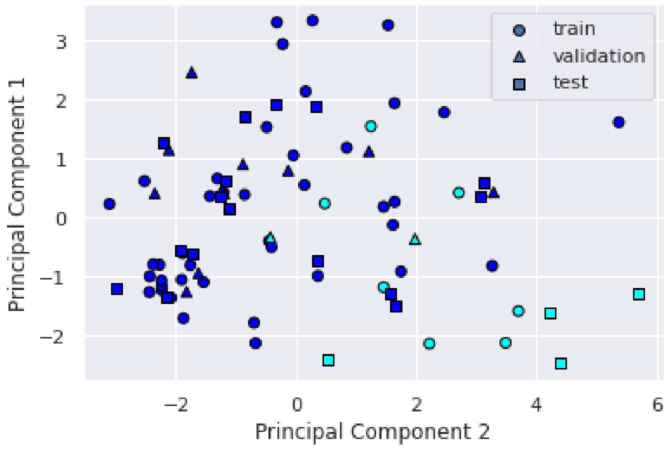

Figure 3: Data1: Principal Component Analysis

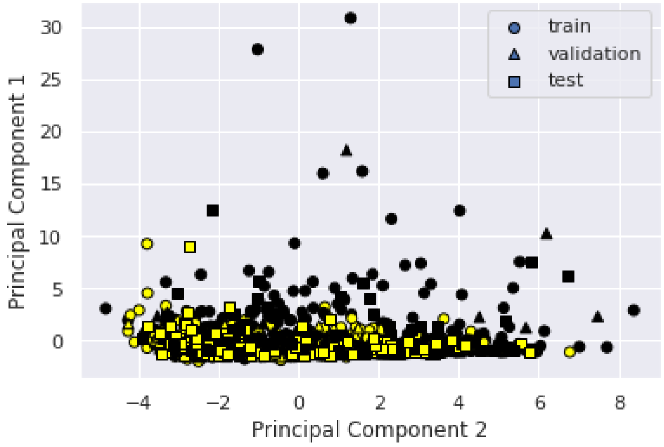

Figure 4: Data2: Principal Component Analysis

This makes it very hard for KNN since the points are sitting on top of each other, and makes it even harder for Decision Tree since adding thresholds would rarely create pure data in a clump. This is probably the reason why DT was so hesitant to choose deeper depths because clumped data doesn't offer any information gain when splitting it. Hence, KNN and DT gave very low accuracies when attempting to predict the data. This is further shown by the ROC graph shown above. The lines barely deviating from the diagonal is an indication that the models do not fit the data very well.

In summary, KNN was the slightly better model for both datasets, but it was made fairly obvious that if the data is not properly dispersed, it becomes very difficult to correctly create an accurate model. This was shown explicitly in dataset 2. However, because of this, weighted KNN performed better than KNN on dataset1 , but performed much worse on dataset 2. For future improvements and for when we get there in the course, one could have tried to implement K-fold Cross-Validation, Naive Bayes and/or Support Vector Machines and compared those classifiers to the ones we implemented .

7 Subject Extension/Originality

Diabetic retinopathy [1], a complication in diabetes, is a disease that can often be diagnosed using fluorescein angiography, a way to record blood flow in the retina . Post-imaging, the features vital to the diagnosis include microaneurysms appearing as small red dots in the superficial retina layers, flame-shaped hemorrhages as splinter hemorrhages in the superficial nerve fiber layer, macula exudates as shaded regions, etc. Needless to say, the identification of features through medical imaging and processing of biomarkers are an amazing tool to diagnose an upcoming disease. Similarly to fluorescein angiography, where we try to identify features through imaging of the retina, brain imaging presents a universe of potential with regards to disease prevention. An example of this is where machine learning was used to distinguish pediatric brain tumor types using multi-parametric magnetic resonance imaging . The study demonstrated that multivariate analysis, multiparametric MRI and machine learning techniques such as KNN and PCA for dimensionality reduction can be employed to distinguish between high and low grade with high accuracy (85% BAR) [3].

8 Statement of Contributions

Work was equally distributed among the three team members.

References

[1] T. Taveira-Gomes, “Machine learning on the diabetic retinopathy debrecen data set data set,” 2016. [Online]. Available: https://rpubs.com/tiagotaveira/debrecen#:~:text=Debrecen%20Diabetic%20Retinopathy%20Dataset,or%20an%20image%2Dlevel%20descriptor

[2] R. J. Rosaria Spinella, Rohit Sawhney, 2015 Sept. 29. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26420218/

[3] L. M. James T. Grist, Stephanie Withey, “Distinguishing between paediatric brain tumour types using multi-parametric magnetic resonance imaging and machine learning: A multi-site study,” *ELSEVIER*, p. 6, 2020.