Mini Project 1: Getting Started with Machine Learning

Group 60: Alvin Chen, Ziqi Li, and Emily Tam

February 5, 2021

1 Abstract

In this project we compared two models: K-Nearest Neighbors and Decision Trees and their accuracy with regards hyperparameters. For the KNN model we tested the accuracy using different values of K and distance functions, and for the Decision Tree model we tried different maximum tree depths and different cost functions. We found that our KNN model had an accuracy of up to 98% using Euclidean distance function and $k = 20$ for the Breast Cancer dataset and the Decision Tree model had an accuracy of 95% for the Hepatitis dataset using an entropy cost function with maximum depth of 5. In our analysis, we found that increasing K in the KNN model resulted in clear underfitting, and increasing depth for the decision trees resulted in overfitting. We also investigated selecting only important features to simplify the dataset, and used PCA to visualize the model predictions.

2 Introduction

The Wisconsin Breast Cancer dataset and the Hepatitis dataset are two well-known data-sets that are frequently used in testing machine learning models, having a simple 2-class label and a plethora of features to be fitted. In this project, we investigated the performance of the K-Nearest Neighbour and Decision Tree models on these two benchmark datasets from the UC Irvine Machine Learning Repository. Academic models are able to consistently achieve an accuracy of 97% on both the hepatitis and breast cancer classification [2], and our goal was to find a combination of hyperparameters to achieve such accuracy. Our experiments include: comparing the accuracy of these two model on the two datasets, observing the impact of changing hyperparameters and different distance/cost functions on the accuracy of both models, observing the decision boundary plots for each model and observing the change of accuracy when only using correlated features as input features for the KNN model. After running these experiments, we discovered that as we increase K for the KNN model, the training data accuracy decreases from underfitting, and as we increase the maximum depth for the Decision Tree model, the training data accuracy converges to 1 due to overfitting. On average, the Euclidean distance function gave a better accuracy for the KNN model, and the entropy cost function gave a better accuracy for the Decision Tree model. The decision boundaries show the potential overfitting in the form of “islands” in the KNN plots and long thin rectangles in the decision tree plots. While exploring ways to improve this experiment, we hypothesized that you could keep features with high correlations to the classification and eliminate relatively irrelevant features to the classification during the training and test process to construct a model with better accuracy performance. Since the Breast Cancer dataset only contained continuous features [1], we demonstrated the use of PCA reduction to help visualize the model predictions. We decided against reducing dimensions with PCA on the hepatitis set since the data contained both categorical (binary) data and continuous data.

3 Datasets

The two datasets we are using for testing our models are: the Breast Cancer dataset (699 instances) and the Hepatitis dataset (155 instances). After dropping samples with missing instances (e.g. ‘?’), we were left with 683 clean instances for the Breast Cancer dataset and 80 clean instances for the Hepatitis dataset. For the Breast Cancer dataset, we split 100 instances as testing data and the rest of the instances were used as training data. We found that one of the features, ’Bare Nuclei’, was parsed as a string datatype, so we manually casted it into an int, under the assumption that all of the features for the breast cancer set are continuous. As visualized in the histograms, some of the continuous features were biased towards a certain value (unimodal peak near one end of the range).

For the Hepatitis dataset, we split 20 instances as testing data and the rest of the instances were used as training data. It is worth mentioning that the continuous features in the hepatitis dataset do not have a same scale. Thus, we also normalized these continuous features to make them have a same scale.

Moreover, for a better understanding on these two datasets, we have shown the distributions of each feature using histograms and the correlations between features using a correlation matrix. We also explored pairwise relationships between features, and have included these figures in the code file.

There are many possible ethical concerns that can arise when working with these kinds of datasets. Most notably, there is the issue of imparting bias and discrimination during the training of models using these datasets

[3]. It is important to examine whether the dataset is representative of the general population, especially if it is applied in medical diagnostics [4]. Another issue to consider is the role of human judgement, which could impart additional biases to the trained model [3].

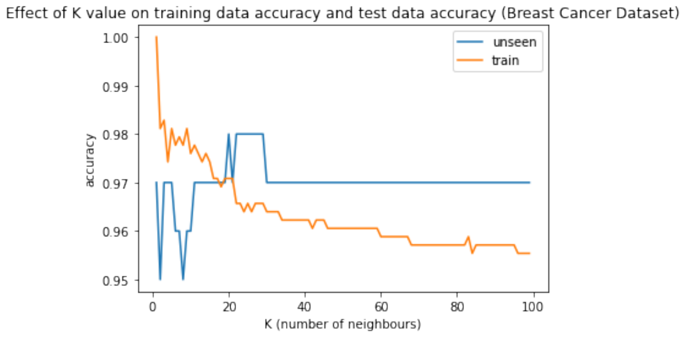
4 Results

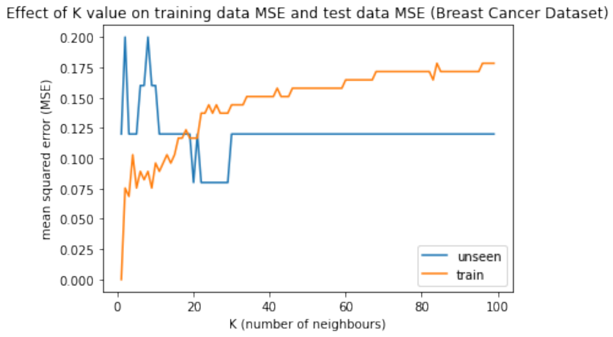
Task 3.1

For Task 3.1, the best hyperparameters for each model and dataset pair were chosen using the results we obtained from the task3.2 3.4. The KNN model used on the Breast Cancer dataset was most accurate (accuracy = 98.0) when K = 20 and the euclidean distance function was used. The KNN model used on the Hepatitis dataset was most accurate (accuracy = 90.0) when K = 8 and the euclidean distance function was used. The Decision Tree model used on the Breast Cancer dataset was most accurate (accuracy = 97.0) when maximum tree depth = 5 and the entropy cost function was used. The Decision Tree model used on the Hepatitis dataset was most accurate (accuracy = 95.0) when maximum tree depth = 3 and the entropy cost function was used. Overall, for the Breast Cancer dataset, the highest accuracy was achieved with the KNN model. For the Hepatitis dataset, the highest accuracy was achieved with the Decision Tree model. (see code file)

Task 3.2

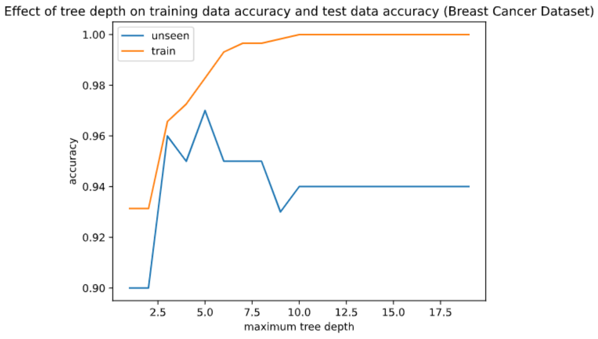
Task 3.2 involved testing different K values to see how it affects the training data accuracy and test data accuracy. The highest accuracy was achieved with K = 20 for the KNN model used on the Breast Cancer dataset (accuracy = 98.0) and K = 8 for the KNN model used on the Hepatitis dataset (accuracy = 98.0). 10-fold cross validation supported these choices for K-value (refer to Appendix).

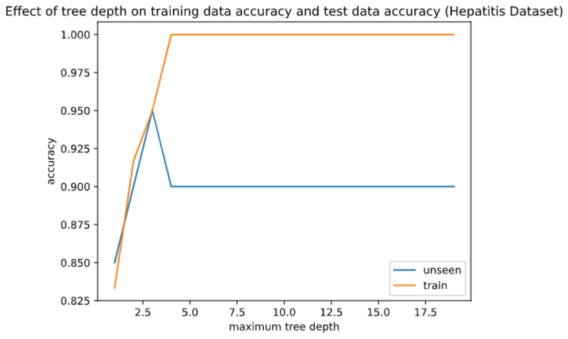




The general trend observed was that as the K-value increased, the test data accuracy increased up to a certain value, before decreasing and stabilizing at a slightly lower value. On the other hand, the training data accuracy starts at 100.0 and decreases as the value of K increased.

Task 3.3

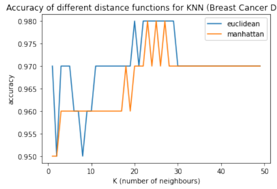


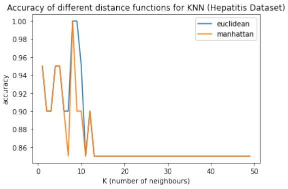


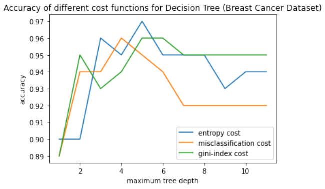
For both the Breast Cancer dataset and the Hepatitis dataset, as we increase the maximum tree depth, the training data accuracy keeps increasing and eventually converges to 1. This is because the model is overfitting to the data as the maximum tree depth increases. When the model overfits the training data, the testing data accuracy decreases because the model cannot generalize to this unseen data.

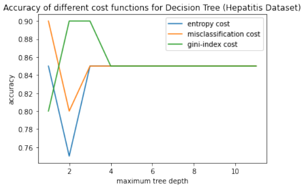
Task 3.4

For the KNN model, the graphs show that in term of accuracy the Euclidean distance function performs better on the Breast Cancer dataset and Hepatitis dataset as we increases K. (It is also worth mentioning that we have used Hamming distance function for the categorical features in the Hepatitis dataset, and the Euclidean and Manhattan functions are only applied to continuous features.) For the Decision Tree model, the graphs show that the entropy cost function gives the highest accuracy for the Breast Cancer dataset and the Hepatitis dataset. (Refer to ipynb file for larger images.)





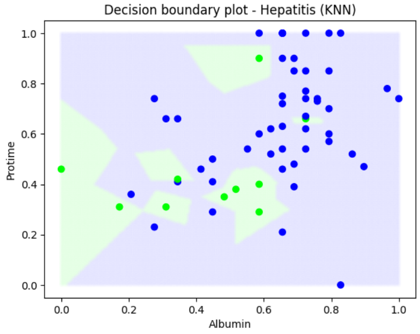


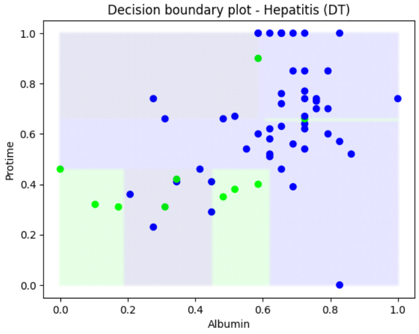


Task 3.5

In trying to plot decision boundaries for the graphs, we realized that there were too many dimensions to represent in a comprehensive decision boundary plot. We thus decided upon two ways to visualize the data: map the boundaries based off the PCA reduced training sets, or to choose two highly correlated features, selected from the correlation heatmap to train and plot a decision boundary on. In our research, we found that for categorical data, PCA was not as valuable/accurate in reducing dimensions since PCA works with variance, a concept in continuous values only[4], therefore we only included the breast cancer PCA reduced set.

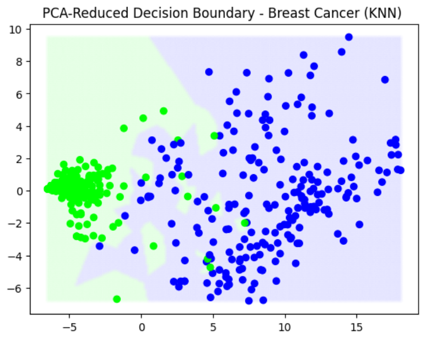
In terms of just using two highly correlated features, in the breast cancer set we chose to plot uniformity of cell shape vs. uniformity of cell size, and for the hepatitis set, we chose albumin and protime since they were highly correlated and continuous features. The following plots show the decision boundaries on the hepatitis set ($k = 5$):

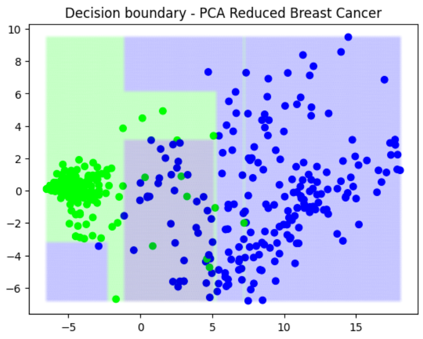




For the decision boundary plots on the KNN models, we see the regions bounded are irregular in shape, as expected due to the Euclidean distance used to predict, whereas in the following decision boundary on a Decision Tree model, we observe that the boundaries are orthogonal to the axes, clearly indicating and defining the split at each level. The same observations apply to the decision boundaries for the PCA-reduced breast cancer datasets ($k = 5$).

Observe that in the decision tree plots, to account for certain data points that exist in a mislabeled location, the tree model will ”slice” the space with a very thin rectangle. This can be seen in both the hepatitis dataset and breast cancer dataset and appears to be overfitting. The same overfitting can be observed in the KNN boundaries, where the overfit points simply appear as “islands”.





5 Creativity

Part 1: Accuracy performance when using only correlated features

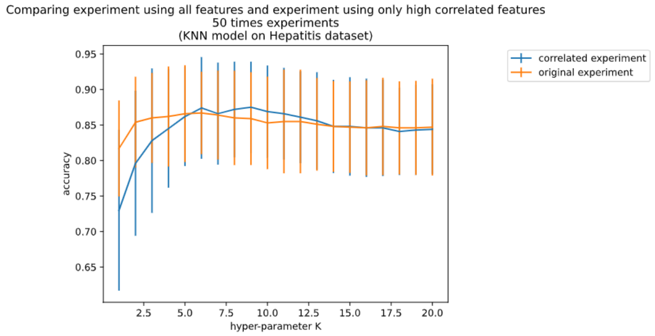


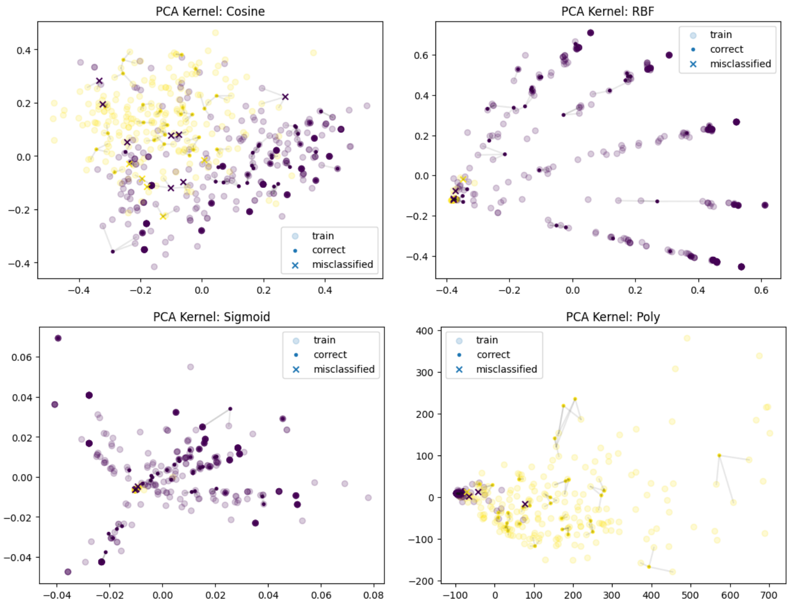
Figure 6: The average accuracy of the original model and the simplified model (error bars show variance)

Idea: Running experiments using KNN while keeping features with high correlations to the classification and eliminating relatively irrelevant features to the classification during the training and testing processes (using the correlation matrix shown in Task 1). Observed the average accuracy of the original model (taking all features as input) and this new model (taking only ASCITES, VARICES, ALBUMIN, PROTIME and HISTOLOGY as input features).

Results: From the following graph (average accuracy graph), we can see that after running 50 experiments (re-sampling training set and test set for each experiment) on the original KNN model (taking all features as input) and the simplified KNN model (taking only ASCITES, VARICES, ALBUMIN, PROTIME and HISTOLOGY as input features), the average performance of the simplified model is better than the average performance of the original model (especially in the range of K=5 to K=13). Also, the highest accuracy is given by the simplified model. This result shows that for the hepatitis dataset, we could eliminate relatively irrelevant features to the classification to improve the average accuracy of our prediction for some range of K's.

Part 2: Investigating Kernels for PCA

While we were considering options for visualizing the high dimensional data, the use of Principal Component Analysis to reduce dimensions was suggested. We discovered that PCA's could be extended with kernels for some interesting pattern recognition. We tested out a few kernels included in the sklearn.Decomposition library in hopes of the data conforming better to one of the kernel patterns (with regards to the default linear). We noted that only the cosine kernel PCA reduction had a well-defined boundary between the two classes, while the rest of the kernels were not useful in visualizing the higher dimensions. In the end, the linear PCA still had the clearest representation of the data. The following use $k = 3$:



Other creativity parts:

- Correlation matrix for both datasets (shown in Task 1 in the code file) to find most correlated features
- Normalization of the continuous features in the hepatitis dataset
- Using weighted sum method for predict in the hepatitis dataset, which contains both categorical and continuous features

6 Discussion and Conclusion

In this project, we learned how hyperparameters, and different distance and cost functions affect the accuracy performance of the KNN and Decision Tree models by training them on two distinct health datasets. The results of the experiments found that the KNN model trained on the Breast Cancer dataset achieved the highest accuracy out of the four different training models. We improved the accuracy performance by normalizing the continuous features in the hepatitis dataset and using a weighted-sum method when predicting using the KNN model for the Hepatitis dataset. The Breast Cancer dataset had a greater sample size than the Hepatitis dataset, and thus was able to generalize better during test phases. When plotting decision boundary plots, we gained insight about potential overfitting in the form of “islands” within the KNN plots and long thin rectangles in the decision tree plots. Additionally, as part of the creativity requirement, we investigated the impact of features and dimensionality on the models. We found that one could keep features with high correlations to the classification and eliminate relatively irrelevant features to the classification during the training and testing process to construct a model with better accuracy performance [1]. A possible direction for future investigation could be using the analysis of correlations between features and the classes to reduce the dimensionality of our datasets and improve the accuracy performance of all four models. This could be accomplished semi-automatically via PCA dimension reduction, which we experimented with to reduce the breast cancer dimensions for visualization purposes, since the features were all continuous. Another potential experiment could investigate the accuracy of higher dimension fitted models vs. PCA reduced-fitted models.

One major downside of the KNN model is that it is sensitive to feature scaling and noise. In future experiments, we could compare the KNN model to other models in these domains. Ideally, more important features should maximally affect the classification, and should have larger scale, while noisy and irrelevant features should have a smaller scale. We could also experiment with using a weighted KNN model, where the weights are inversely proportional to the distance. This could improve the accuracy and generalize the model to testing data. Although we experimented with cross-validation for the KNN model trained on Breast Cancer data, we did not use cross-validation on the other models. Future experiments should implement this, as it would allow for a better estimate of the generalization error and uncertainty measure.

7 Statement of Contributions

Task 1: Alvin, Ziqi
Task 2: Alvin, Ziqi
Task 3 and Creativity: Emily, Alvin, Ziqi
Report: Emily, Alvin, Ziqi

References

[1] Taha, I., & Ghosh, J. (1996). Characterization of the Wisconsin Breast cancer Database Using a Hybrid Symbolic-Connectionist System. Proceedings of the Intelligent Engineering Systems through Artificial Neural Networks Conference. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.9360&rep=rep1&type=pdf

[2] Nahato, K. B., Harichandran, K. N., & Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine, 2015, 460189. https://doi.org/10.1155/2015/460189

[3] Pazzanese, C. (2020, October 26). Ethical concerns mount as AI takes bigger decision-making role in more industries. The Harvard Gazette. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

[4] Ledford H. (2019). Millions of black people affected by racial bias in health-care algorithms. Nature, 574(7780), 608–609. https://doi.org/10.1038/d41586-019-03228-6

[5] Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLoS computational biology, 15(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

[6] Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Nature, 7(9). https://doi.org/10.1057/s41599-020-0501-9

Appendix

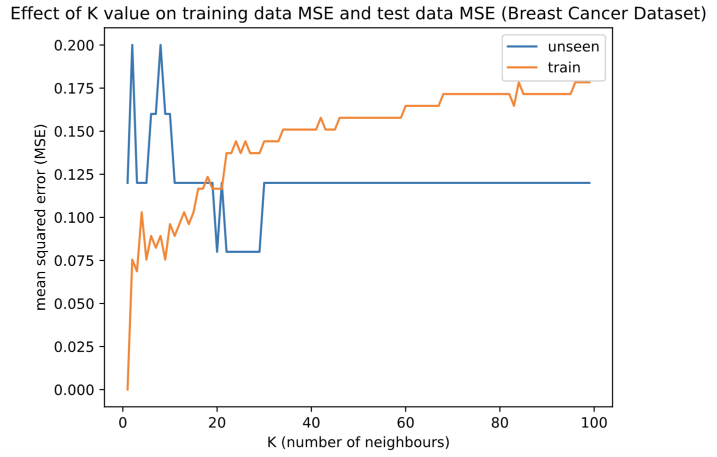


Figure 1: K-value vs. Mean-Squared Error

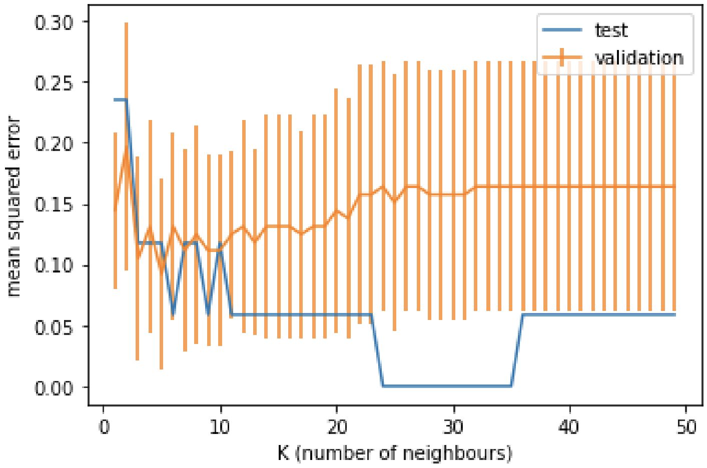


Figure 2: Cross Validation