DS 8002 Machine Learning Project 1 – Data set classification analysis (October 2016)

Najlis, Bernardo J. (Student Number: 500744793)

Abstract—**In this project, you will apply three algorithms to two data sets. The data and the algorithms are provided by R. Please answer each question in the order they appear. Do not skip to later steps to answer earlier questions that ask you to predict outcomes based on your analysis of the data and understanding of the algorithms. Submit your report in D2L by midnight on the due date.**

I. Introduction

IN this project we perform analysis on two data sets (iris and contact lens) over three different types of classification machine learning algorithms. The analysis was done using R and libraries for each type of algorithm used. As the original assignment was meant to be done using Weka, most of the R code is just a shell interface to Weka classes and libraries.

This report is accompanied by an R markup document with all the code and its output to support all answers and reasoning presented here.

II. DATASETS

The two datasets are different in at least four distinctive ways:

1. **Data format**: Contact Lens is text separated by spaces, Iris is comma separated values
2. **Number of rows**: Contact Lens has only 24 rows, Iris has 150 rows
3. **Attribute data types**: Contact Lens has all numeric integer and discrete values for the attributes, Iris has numeric decimal continuous values for the attributes.
4. **Class attribute**: Contact Lens has a numeric discrete value for class, Iris has a string text for class.

Which algorithm do you expect to perform best on the Contact Lens data data? Why?

I expect the **Multilayer Perceptron** to perform better on the contact lens data, as it tends to perform better over smaller datasets if done with larger structures and high number of epochs.

Which algorithm do you expect to perform best on the Iris data? Why?

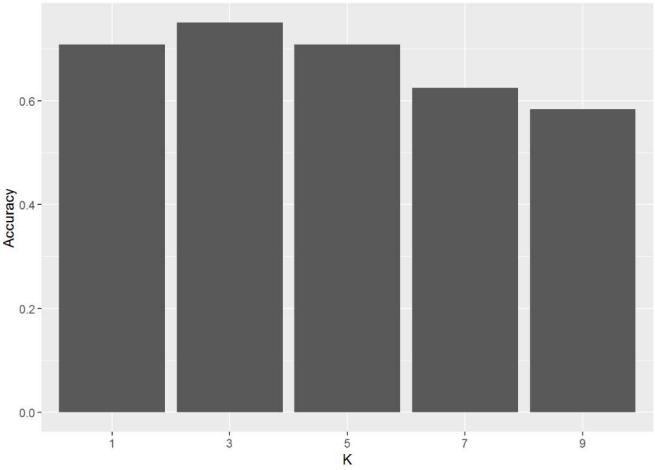
I expect decision trees to perform best on the iris data as it is a larger dataset where the entropy calculations can be applied reasonably, and will not tend to over fit the structure.

III. KNN on the Contact Lens Data

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

Both K = 1 and K = 3 give the best values, as expected. K = 1 should always give 100% accuracy = 1 and K = 3 also has 100% accuracy as the underlying data has 3 classes.

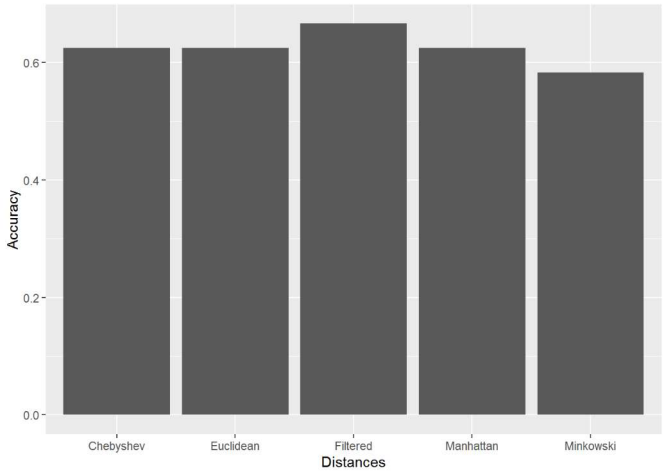


Holding K constant, try different distance functions on each data set.

Which distance function(s) work best for each data set? Why?

Using Weka we set K fixed at 9 (which is the least accurate model using Euclidean distance in the first analysis) and recreated models using Chebyshev, Filtered, Manhattan and Minkowski distance calculations.

Out of all the distance calculations (Chebyshev, Manhattan, Minkowski) the best for the lenses data set is the **Filtered** distance.

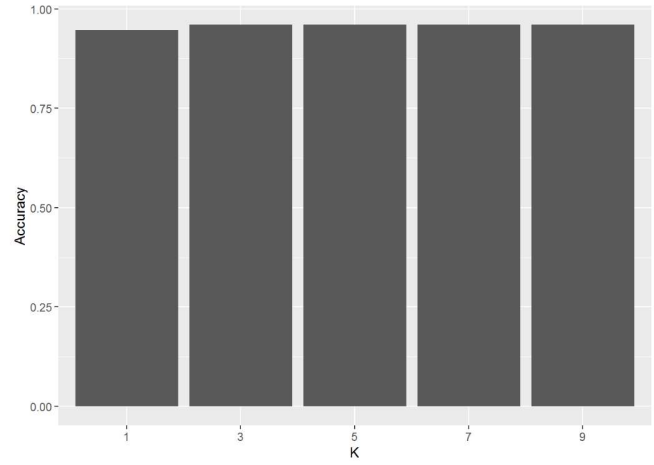


IV. KNN ON THE IRIS DATA

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

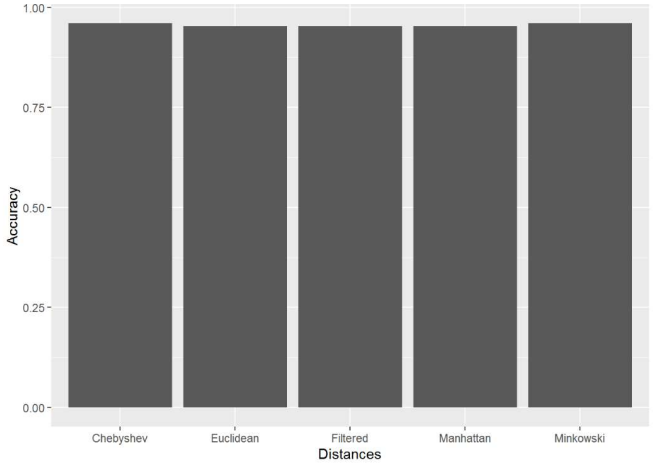
Same as with the contact lens data, the best value is K = 3, as the iris dataset is also arranged in three classes. Accuracy goes down when increasing K over 3. The difference in accuracy is marginal.



Holding K constant, try different distance functions on each data set.

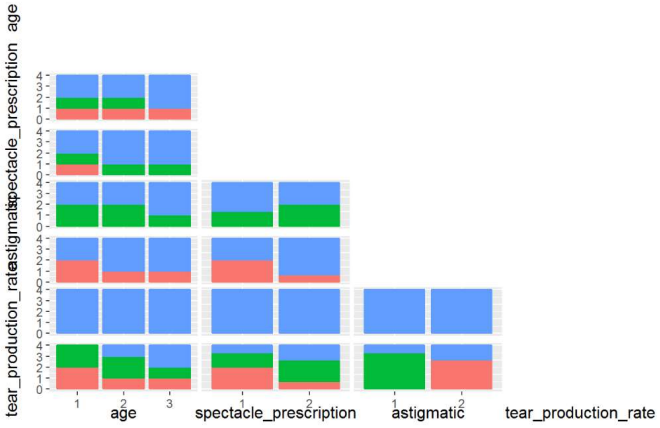
Which distance function(s) work best for each data set? Why?

Same as with the comparison on Euclidean distance with fixed K, the difference in accuracy over different distance calculation methods is marginal.



V. DECISION TREES ON THE CONTACT LENS DATA

Based on R's visualizations, which attribute do you expect to be chosen as the split attribute at the root node?



Based on the R visualization, I expect the age to be the root of the decision tree.

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.

How do ID3 and J48 compare in terms of performance?

As the data set is so small, there is no noticeable performance difference between the two methods.

How does pruning affect test performance and generalization performance?

In general, pruning reduces test performance as it removes branches or leaves that do not improve performance over the

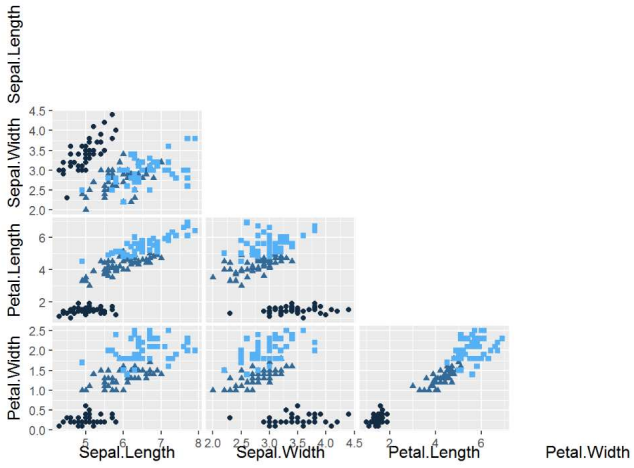
training data.

What does that suggest about overfitting?

This suggests that the tree now is more overfitted in the training data and will not generalize as good as comparable non-pruned tree.

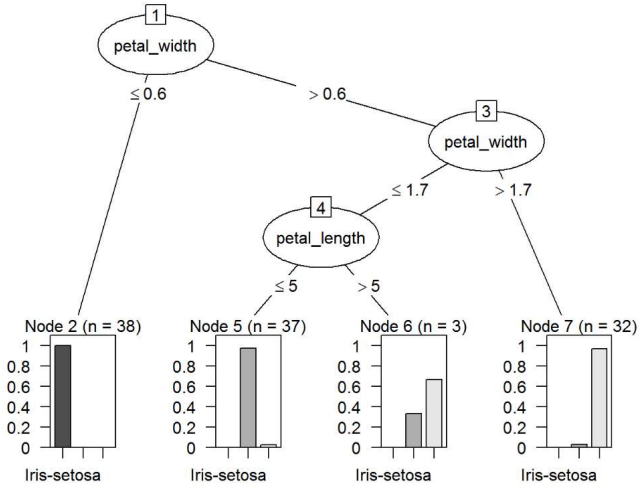
VI. Decision Trees on the Iris Data

Based on R's visualizations, which attribute do you expect to be chosen as the split attribute at the root node?

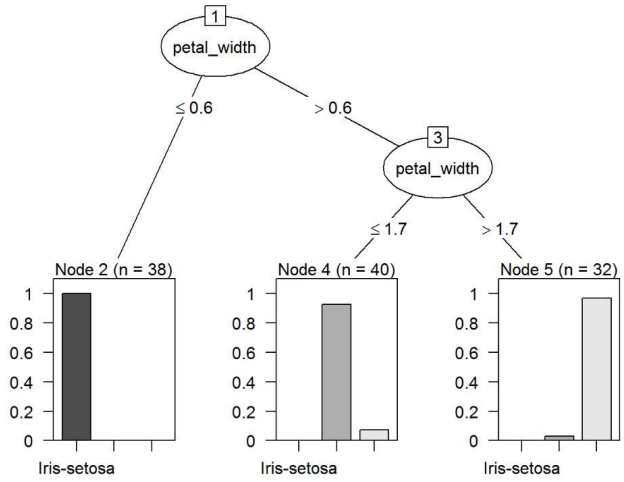


Based on the R visualization, I expect the root of the tree to be Petal.Width.

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.



The unpruned version of the tree has 2 levels deep and accuracy of 97.27%.



The pruned version has 1 level deep and accuracy of 96.36%.

The difference in accuracy most likely does not justify the extra computation required if the unpruned version is used.

Why can't you run ID3 on the Iris data?

It doesn't make proper sense to run ID3 on the iris data because the attributes are continuous numeric values and this implementation of the algorithm doesn't account for this types of attributes. In another type of decision trees, these attributes can be handled by taking decision ranges for the continuous variables.

How does pruning affect test performance and generalization performance?

In general, pruning reduces test performance as it removes branches or leaves that do not improve performance over the training data.

What does that suggest about overfitting?

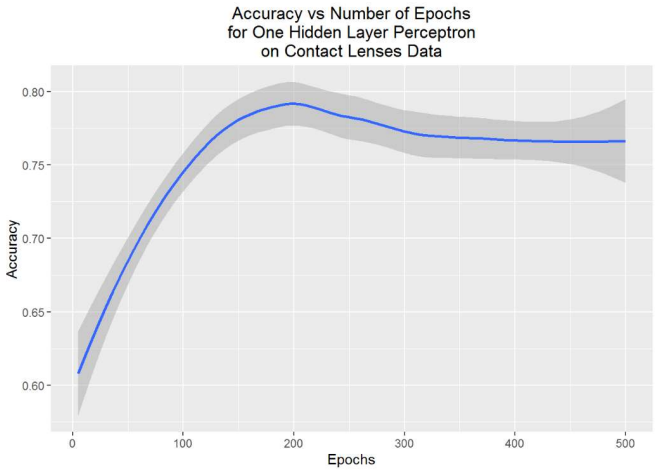
This suggests that the tree now is more overfitted in the training data and will not generalize as good as comparable non-pruned tree.

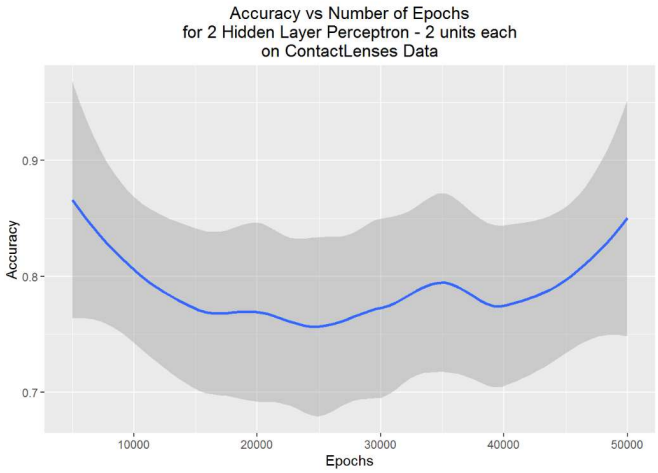
VII. Multilayer Perceptron on the Contact Lens Data and Iris Data

Experiment with different network structures (e.g. extra hidden layers, extra units). Report the results in graphs that show training time (epochs) versus error rate or accuracy.

Which network structures result in the most overfitting?

When comparing One hidden layer vs a two hidden layer – two units multilayer perceptron network, the latter is more overfitting than the former, yielding to the conclusion than the more complex the network the more it will overfit.


Accuracy vs Number of Epochs
for One Hidden Layer Perceptron
on Contact Lenses Data
Accuracy
0.80
0.75
0.70
0.65
0.60
0
100
200
300
400
500
Epochs


Accuracy vs Number of Epochs
for 2 Hidden Layer Perceptron - 2 units each
on ContactLenses Data
Accuracy
0.9
0.8
0.7
10000
20000
30000
40000
50000
Epochs