Mathematics and Natural Sciences faculty

Machine Learning Methods

P160B124

Report no. 2

Report author:
Stud. Marius Arlauskas
MGDMI-0

Supervisor:
Doc. Tomas Iešmantas

Kaunas, 2020

Task 1 LDA QDA

1. Plot 16 images of signs representing that particular letter (use 4x4 grid). Discuss data variability and its impact on the classifier training:

Asnwer:

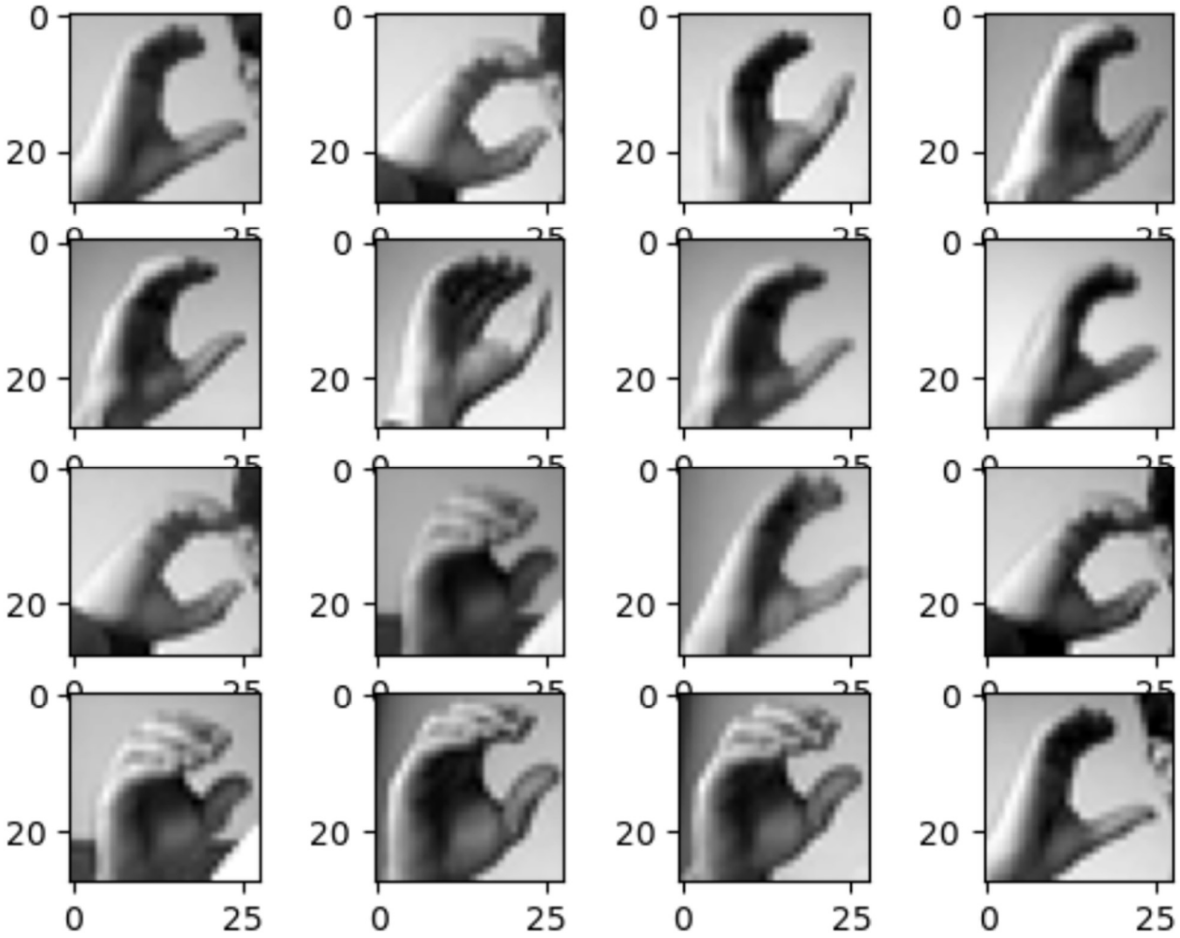

Figure 1 training photos plotted, label=2

The hand is not always perfectly centered, there are even some identical duplicates in the dataset(bottom row), if there are more duplicates, it will be overweighted in the training, some photos include a part of the head in the photo, so that will introduce more noise into the dataset, we should consider cropping the photo a little bit more.

2. Fit LDA to train dataset and report accuracy for test dataset (perhaps a better way to do this is to visualize, because of large number of classes).

Answer:

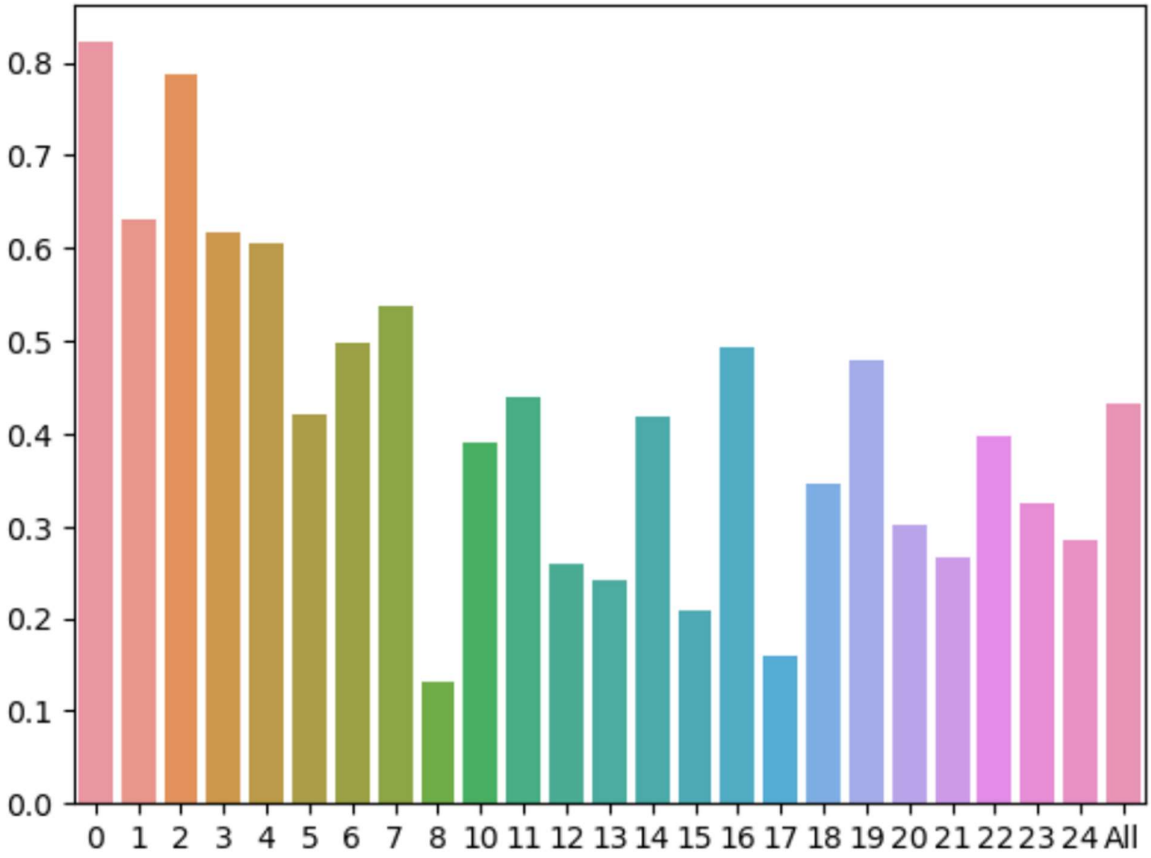

Figure 2 overall accuracies

Overall accuracy is ~43% which is not great, also I noticed that there are no class with label 9, which is weird, was it removed from dataset on purpose?

3. Reduce number of features and report overall accuracies.

Answer:

- Taking every second pixel returns overall accuracy of around ~58%, which is quite a bit better.
- Reducing image quality with interpolation to 14x14 returns overall accuracy of around ~61%, even better.
- After reducing image quality with interpolation, then cutting off top and bottom 2 rows of pixels returns also ~61% overall accuracy, although we removed 56 columns, we could speculate we removed noise and reduced chances of overfitting.
- Reducing image quality with interpolation + antialiasing returns overall accuracy of ~62%, which is a slight improvement, but most likely just random, not significant improvement. Antialiasing probably would have bigger impact on higher resolution images.

Reducing features could have improved perfomance of the model in this case because the hands became into more general shapes, and there were less variability in the data because the position of the hand in the photo became less important.

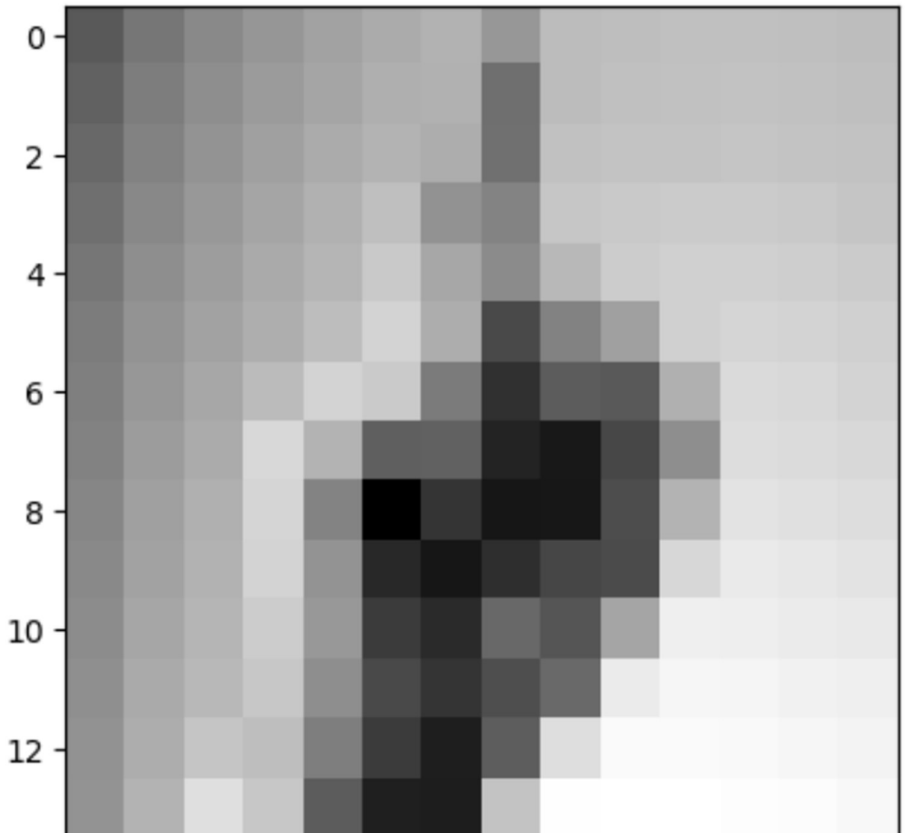

Figure 3 Image with reduced quality by interpolation+antialiasing

4. Fit QDA and report accuracies for each class in test set

Answer:

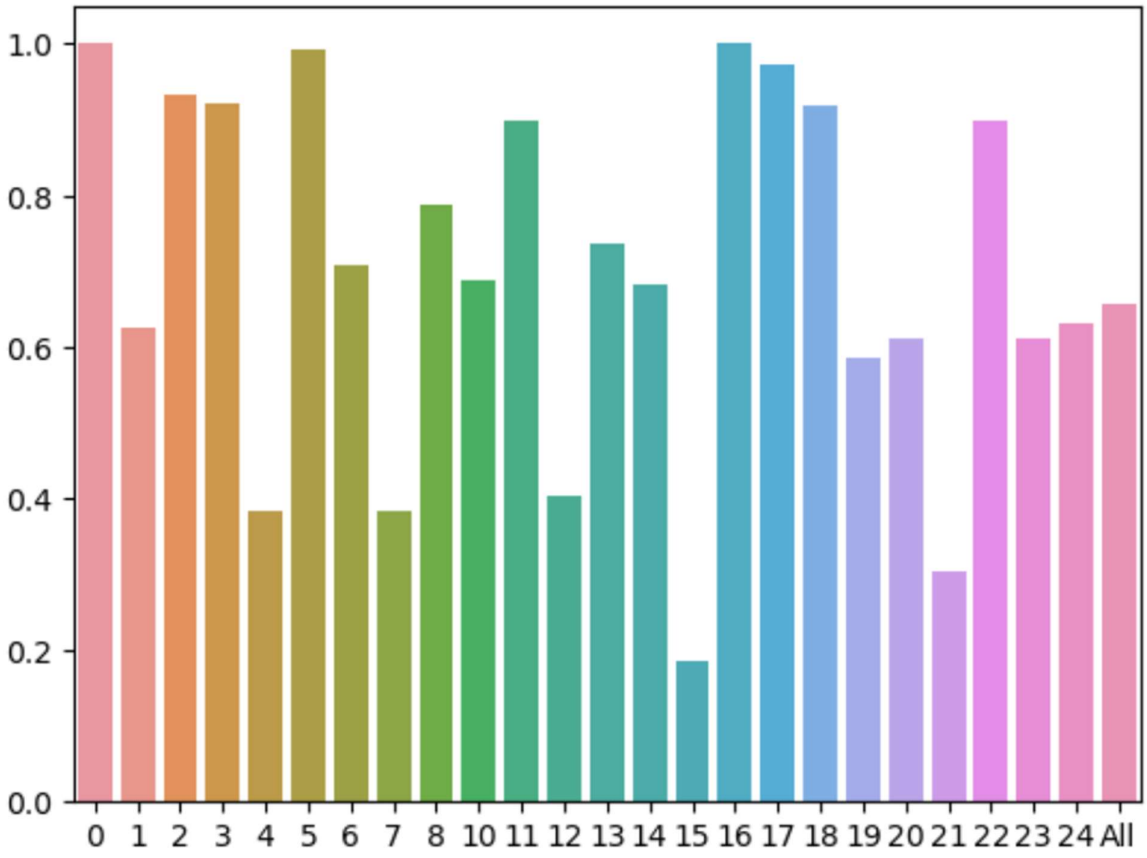

Figure 4 overall accuracies of QDA, images unaltered

Overall accuracy ~66%, already a great improvement over LDA.

5. Repeat the similar analysis as in part (3) and discuss the results.

Answer:

- Taking every second pixel returns overall accuracy of around ~70%, a noticeable improvement.
- Reducing image quality with interpolation to 14x14 returns overall accuracy of around ~73%, geat improvement.
- After reducing image quality with interpolation, then cutting off top and bottom 2 rows of pixels returns also ~72% overall accuracy, the training time for QDA was somewhat long, so reducing the image to 14x14 and then cropping it reduced training time significantly, and the accuracy dropped only by 1% which is highly likely just by random chance, this method would be preffered for final model by me.
- Reducing image quality with interpolation + antialiasing returns overall accuracy of ~73%, no significant improvement over interpolation alone.

Task 2 Decision trees

1a). Split Spotify dataset into training and testing subset (use createDataPartition function from Caret library with p=0.8). Fit a decision tree to a training dataset (use option control = rpart.control(cp=0) inside rpart function). Report CP table plot and discuss which cp value should be chosen and why.

Answer:

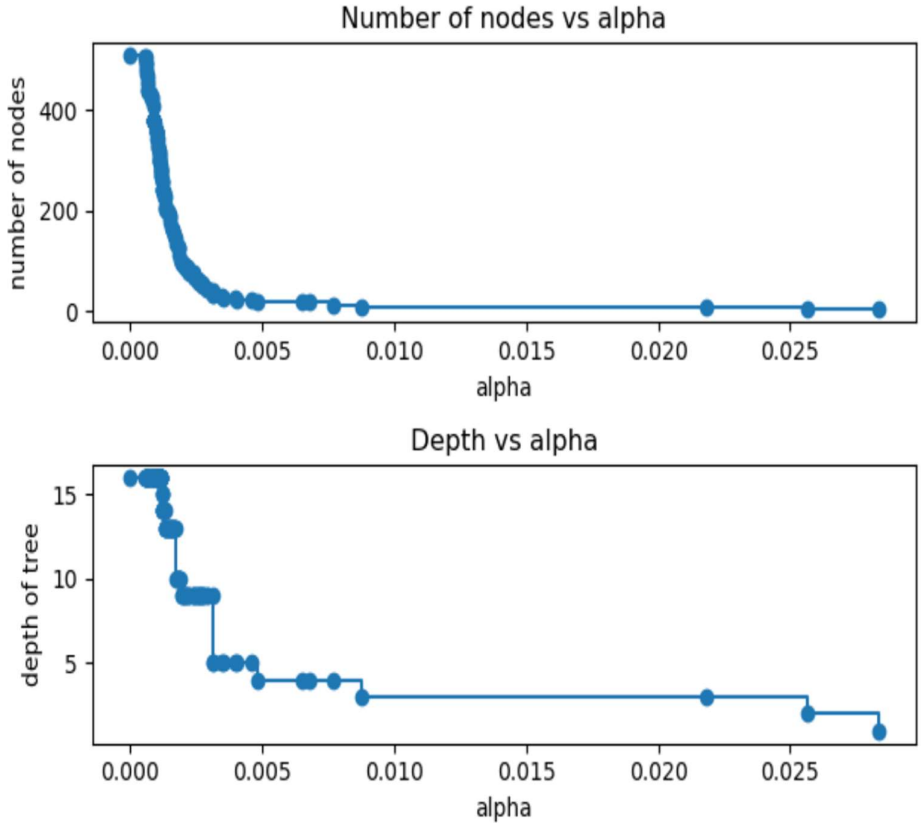

Figure 5 alpha effects on decision tree

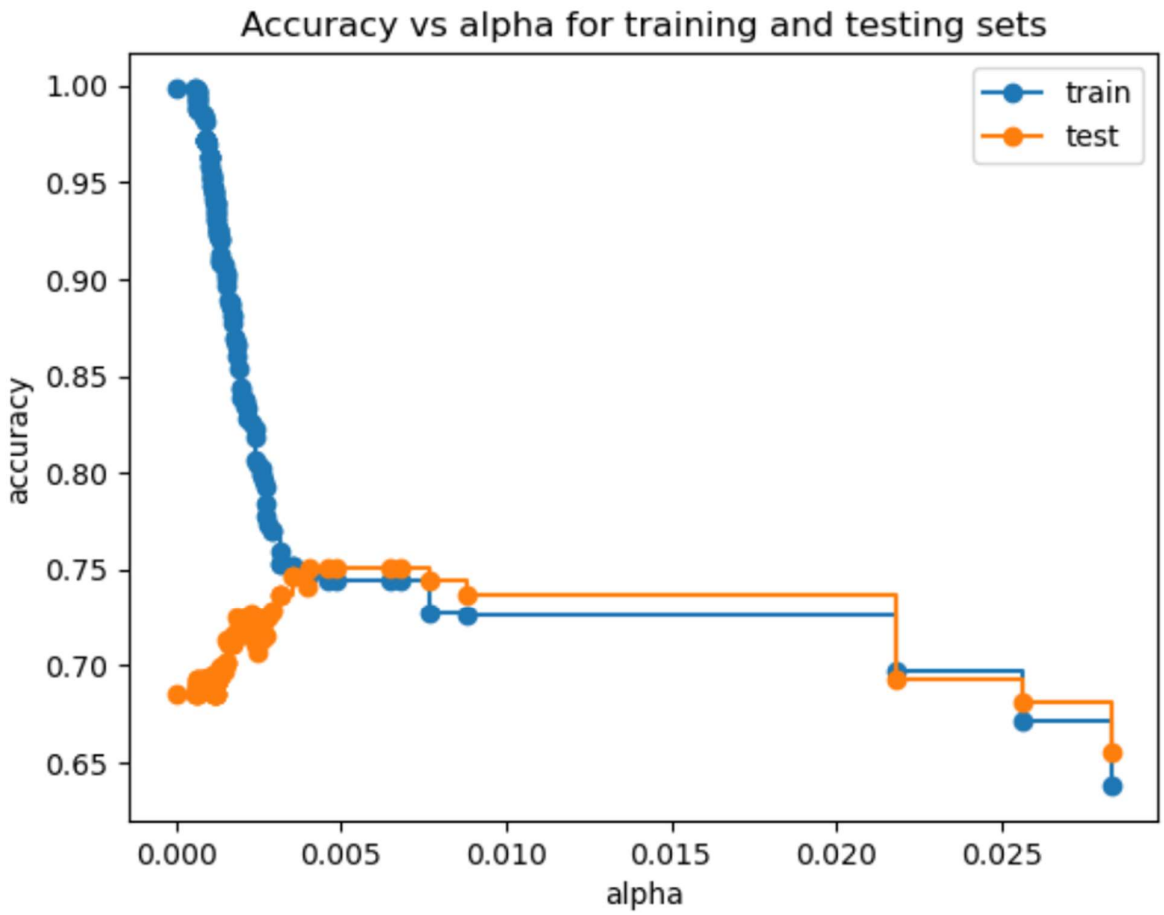

Figure 6 alpha effects accuracy

I would chose cp value to be in the range [0.004;0.008], since the train and test accuracies seem to converge on these values, so we would be less likely to overfit. Also decision trees at these values have considerally smaller number of nodes and depth. Specific value chosen is 0.007.

1b). Using cp value from the above task, prune the tree and plot it.

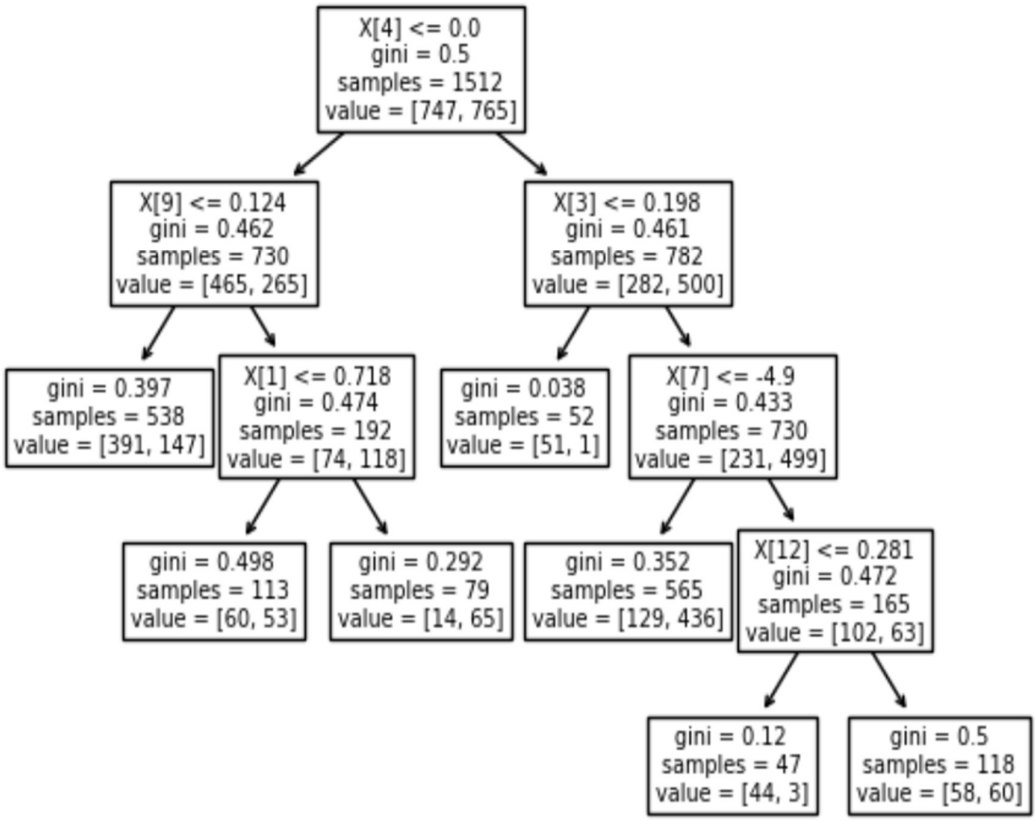

Figure 7 pruned tree

1c). How many leaves that pruned tree has? What is the 10-fold cross-validation error of that tree?

Answer:

The decision tree has 6 leaves.

10-fold cross-validation error is ~35%

1d). Write down one classification rule, learned by the tree.

Answer:

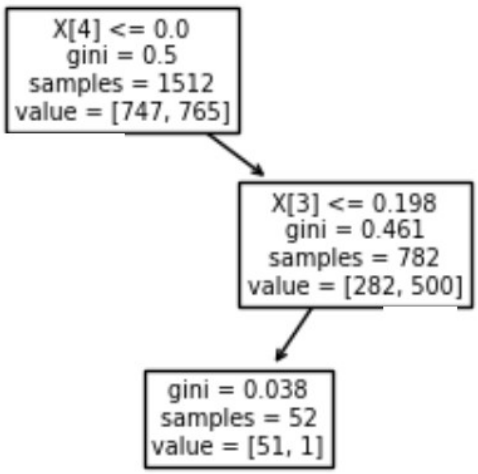

Figure 7 classification rule

If X[4] <=0 and X[3] <= 0.198 then object is classifiend as having class 0. The rule is very accurate on the training dataset, since in only missclasifies 1 object, but whether it is accurate on testing dataset, we do not know.

1e). Use the unpruned (from part (a)) and pruned (from part (b)) decision trees to predict target values from the test dataset and report overall accuracies as well as accuracies for each class. Would you say that the overfitting occurred in this case or it did not?

Answer:

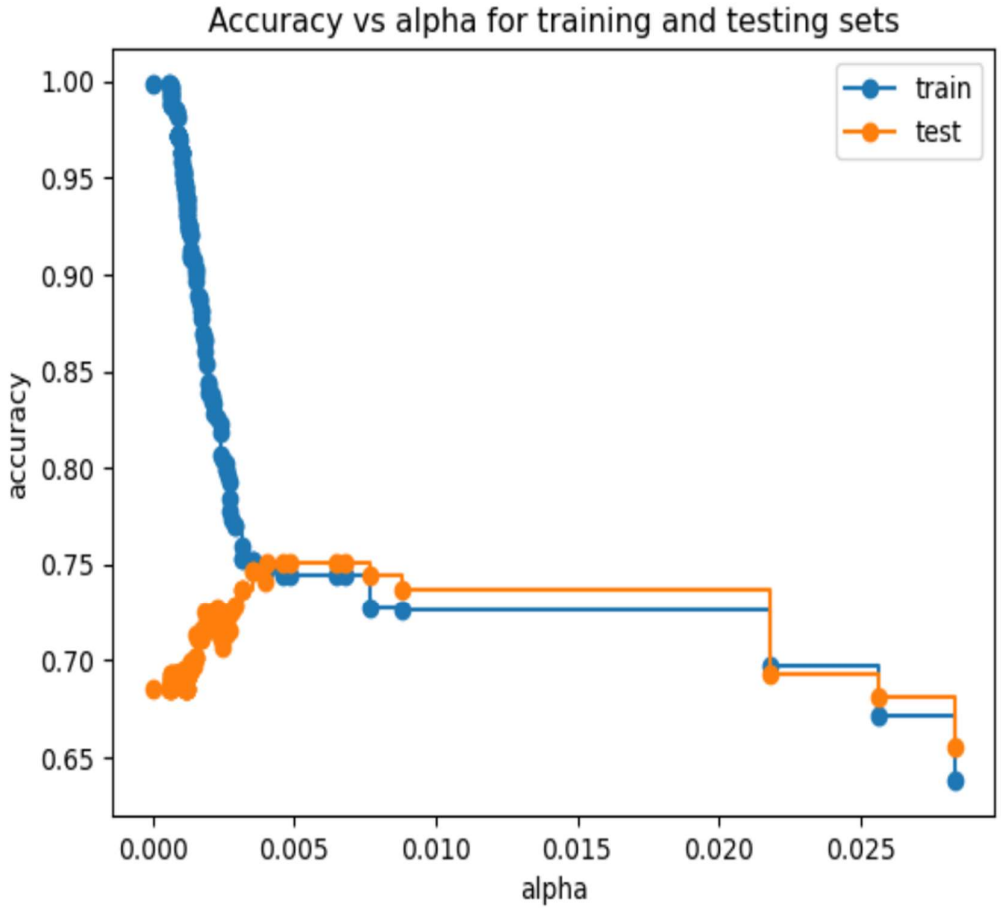

Figure 6 alpha effects accuracy

Note:Since there are only 2 classes, I will only report overall accuracy.

Unpuned was majorly overfitted, since training and test dataset accuracies were vastly different, training dataset had 100% while test dataset only around 68%, I would say the model was the least overfitted when the alpha is in the earlier suggested range of [0.004;0.008].

1f). Was you decision tree a better classifier than logistic regression classifier (use your results from the previous lab report)?

Answer:

Yes, logistic regression model had ~68% accuracy, while decision tree ~73%.

1g). Discuss variable importance. You can access variable importance values by this line: fitted tree$variable.importance. Also you can use plot function to visualize for better understanding.

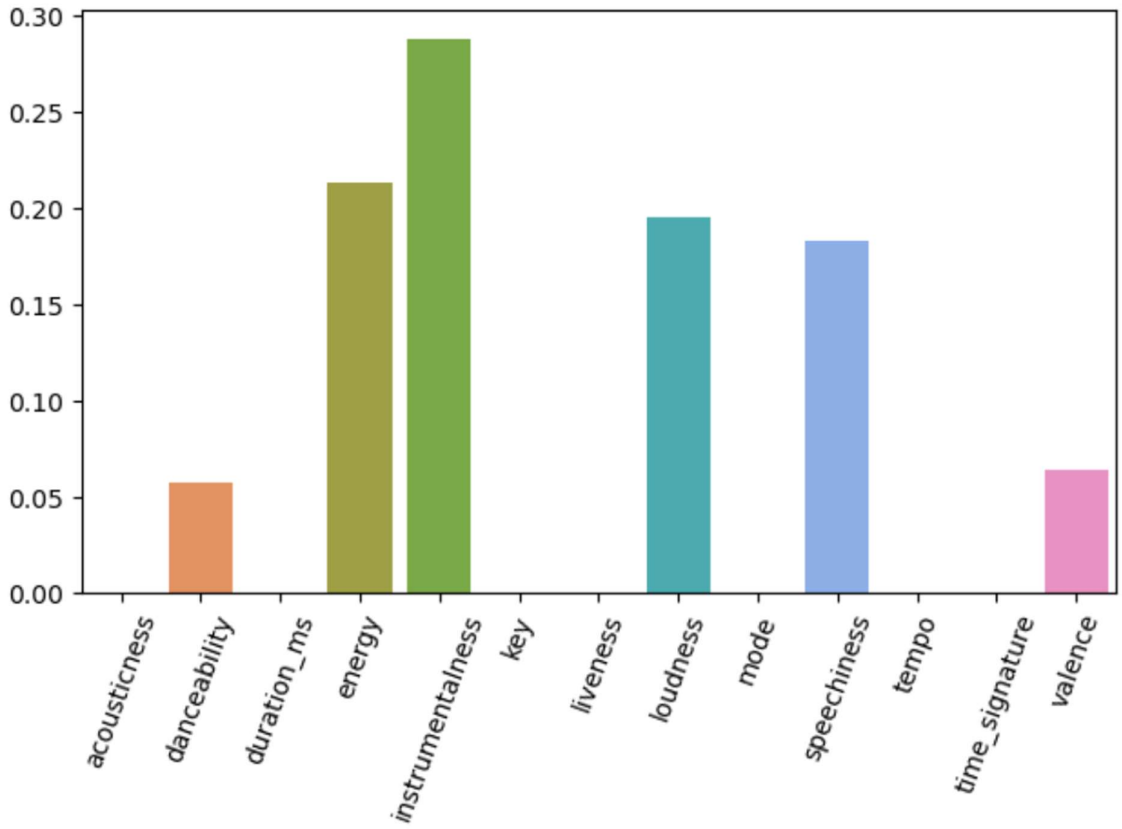

Figure 8 feature importance

The model basically ignored alot of features, ignored features seems quite logical to ignore, very interesting that decision tree seems to have some sort of natural feature selection feature, but in the further lab assignment task it seems to fail on the MNIST dataset.

2.a) Fit a decision tree to a training dataset (use option control = rpart.control(cp=0) inside rpart function). Report plot and discuss which cp value should be chosen and why.

Answer:

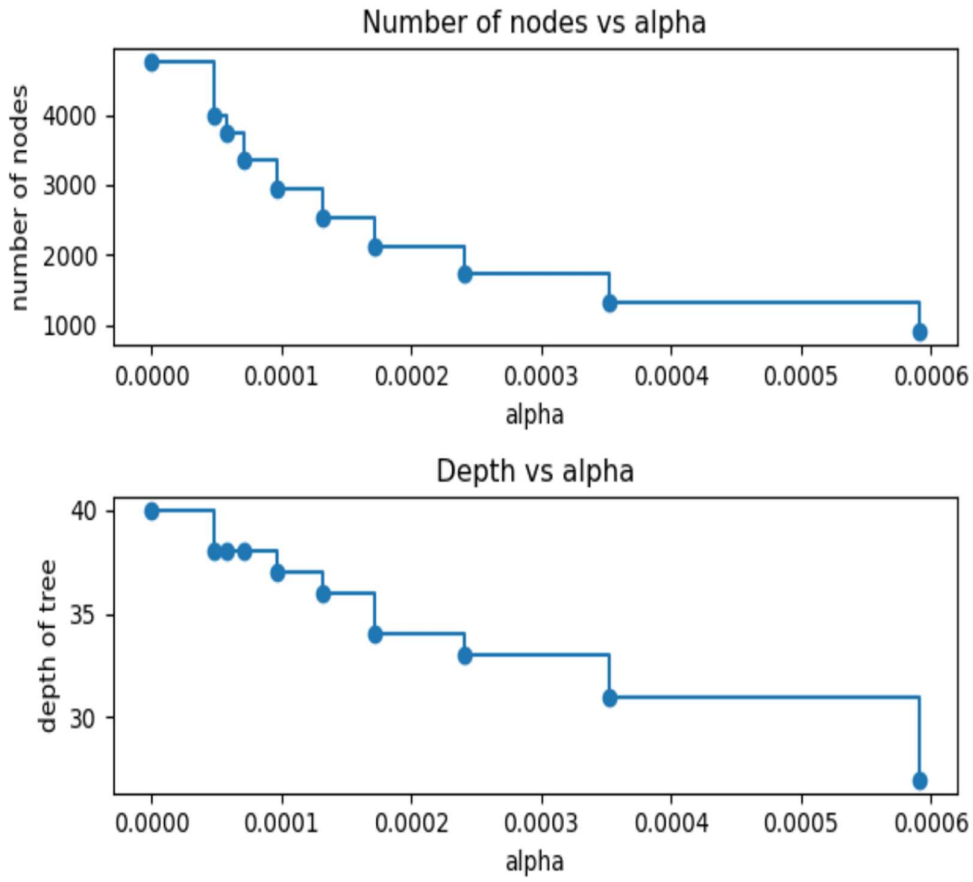

Figure 9 cp effect on decision tree

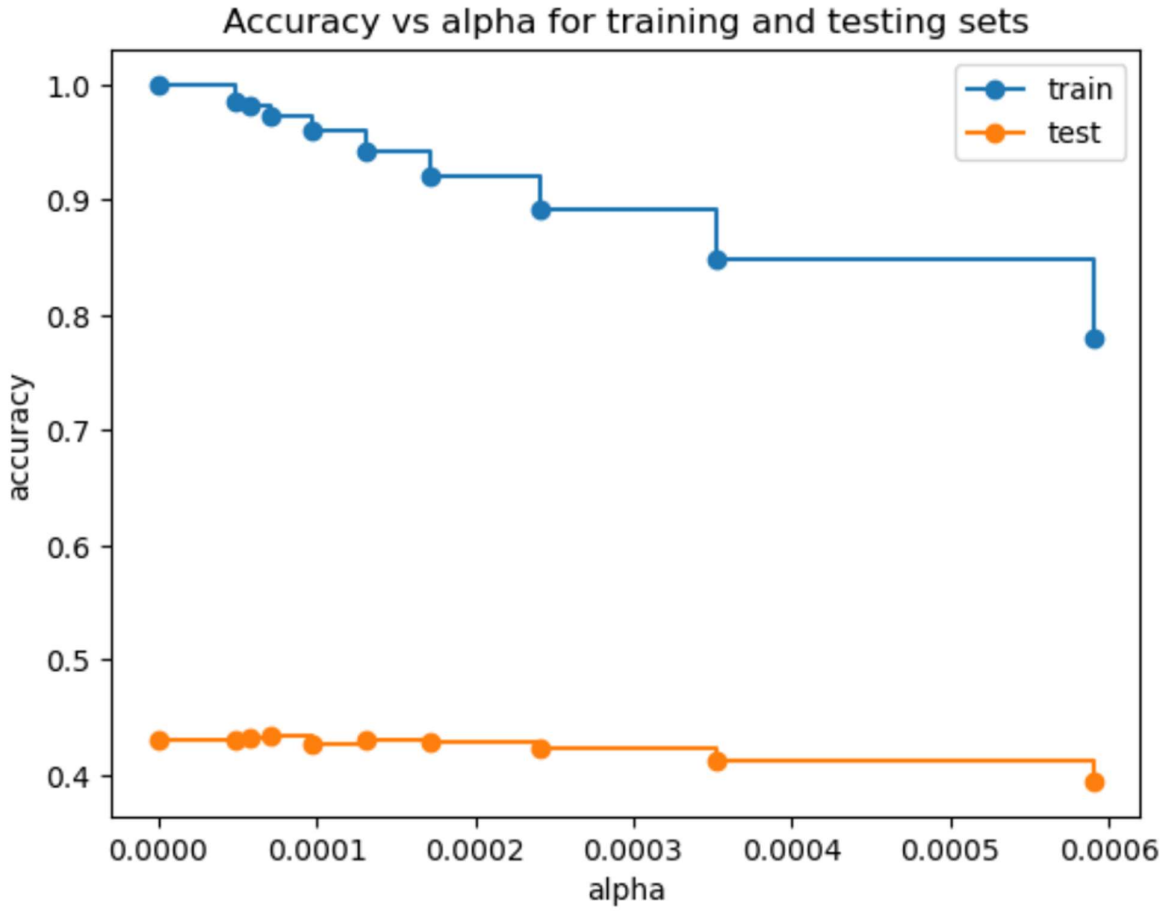

Figure 9 cp effect on decision tree‘s accuracies

These plots surpised me, not sure what to do, I guess we could just pick alpha with the best accuracy on the testing dataset, but that would be cherry picking results, I guess we could take the largest ccp value of 0.006 too, but in general If I would see this kind of perfomance on the dataset, I would just abandon the model and try out different models.
Picked cp=0.0003 ...

2b). How many leaves that pruned tree has? What is the 10-fold cross-validation error of that tree? Could not find a method that returns leave count, but the node count is around 2000, so the leaves are probably in the range of [500; 1300].

10-fold cross-validation error is ~41%, but taking the consideration that it is on the training dataset.

2.c) Use the unpruned (from part (a)) and pruned (from part (b)) decision trees to predict label values from the test dataset and report overall accuracies as well as accuracies for each class. Would you say that the overfitting occurred in this case or not?

Answer:

The overall accuracies are reported in the above figure.

I would say that extreme overfitting occured, but also not sure even if it is possible to generalize the model with this dataset, might be a better idea to look to try out different models.

2.d) Investigate, whether deleting some features (pixels) improves the performance?

Answer:

Interpolation without antialiasing returned overall accuracy of ~25%, definetely not a great strategy, abandoned other strategies, since the training times were extremely long.

Because of the training times, tried out PCA algorithm:

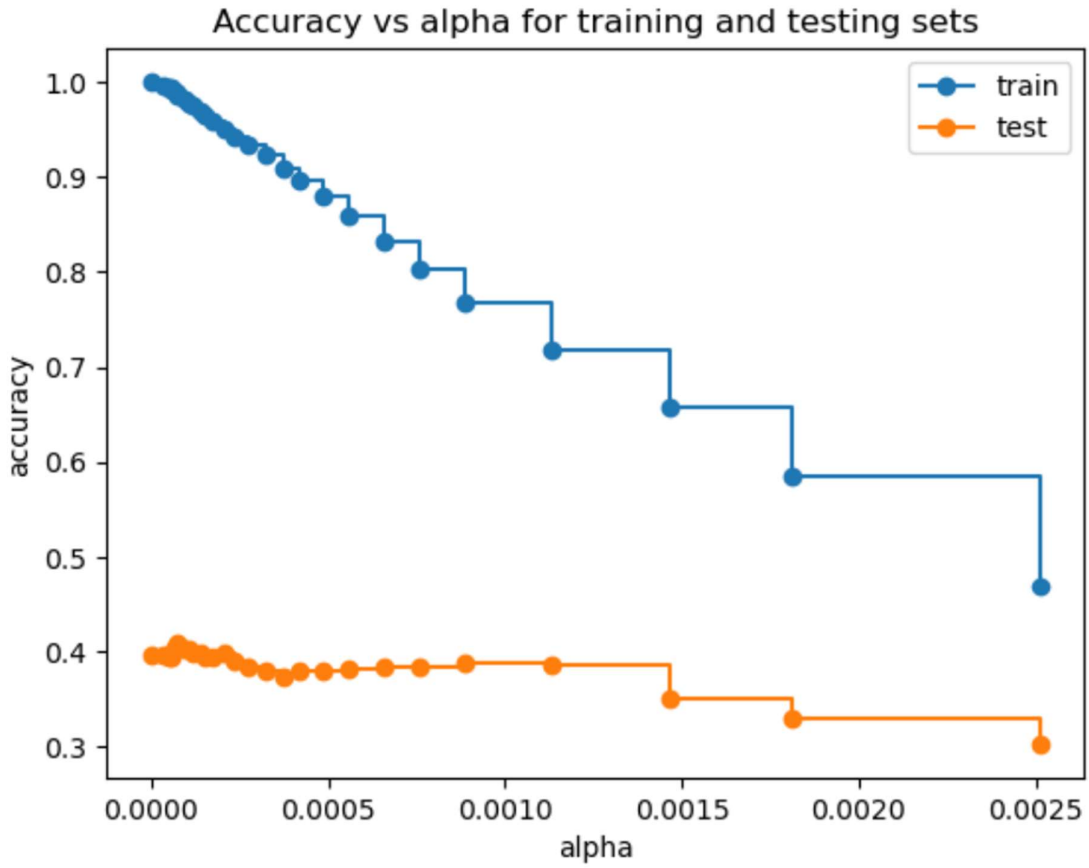

Figure 10 cp effect on decision tree‘s accuracies after PCA

Firstly, I standardized the features, then set the PCA algorithm to retain 95% of the variance,
It dropped 671 features and left only 113 features, and overall accuracy was ~40% which is great, since we greatly reduced time spent on training roughly 5 times, and lost only 1% of overall accuracy.

2.e) Was you decision tree a better classifier than LDA or QDA classifiers? (use your results from the previous assignment)?
Nope, final QDA overall accuracy was ~72% and LDA ~62%, the decision tree overall accuracy was ~41%.