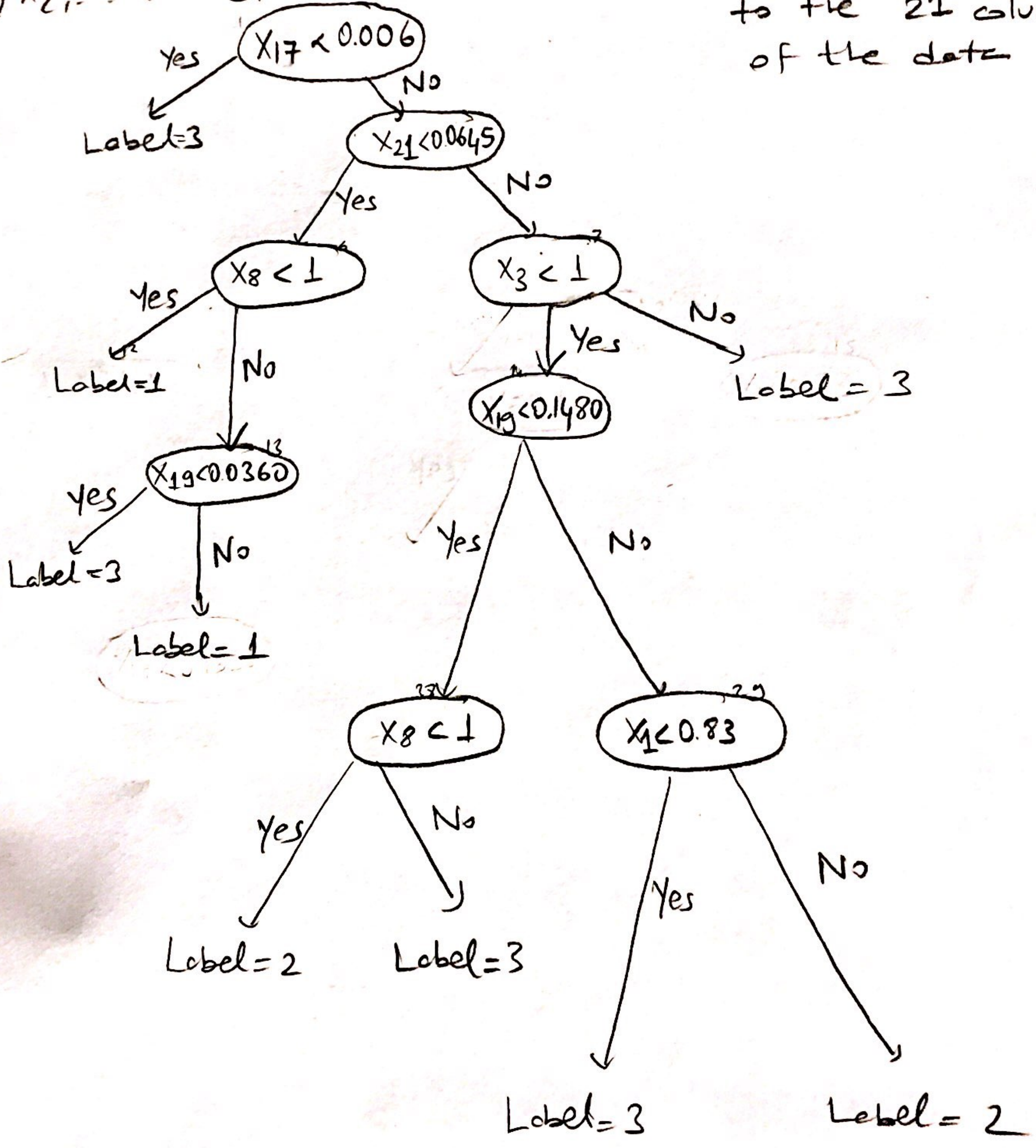
Oytun GÜNEŞ

ID: 20901170

CS 550 HOMEWORK 1

Part 1)

a) $X_1, X_2, \ldots X_{21}$ are the features from the 1st column to the 21st column of the data



In this homework our aim is to perform classification on the data sets: training set and test set. Decision Tree is a common classification method which constructs a tree from the instances, and from the tree it predicts a label for the new instance. For the given Thyroid data set our aim is to classify 3 labels from 21 features (15 binary and 6 continuous features). In the lecture we have learnt to construct a classification tree as follows:

1. List all possible splits
2. Calculate the entropy for every possible split
3. Select the one with the minimum entropy

For the **binary** features split value is either 0 or 1. However for the **continuous** features we need to list all possible splits. For the splitting criterion I have used the entropy at node m, as we did in the class:

$$I(m) = -\sum_{i=1}^{C} P_m(C_i)\log\left(P_m(C_i)\right)$$

$where\ P_m(C_i)\ is\ the\ probability\ of\ having\ i-th\ class\ at\ node\ m.$

Entropy of a binary split is:

$$I(s) = P_{left}I(left) + P_{right}I(right)$$

I have used entropy threshold in order to stop growing. The stopping criterion is needed in order to prevent overfitting.

b)

Training set:

Predicted Class

True Class

	Class 1	Class 2	Class 3
Class 1	93	0	0
Class 2	0	191	0
Class 3	3	8	3477

Accuracy is calculated as follows:

$$Accuracy = \frac{Number\ of\ correctly\ classified\ labels}{Total\ number\ of\ labels}$$

Overall Accuracy = 0.9971, Class Based Accuracies (C1, C2, C3) = [1.0000 1.0000 0.9968]

Test set:

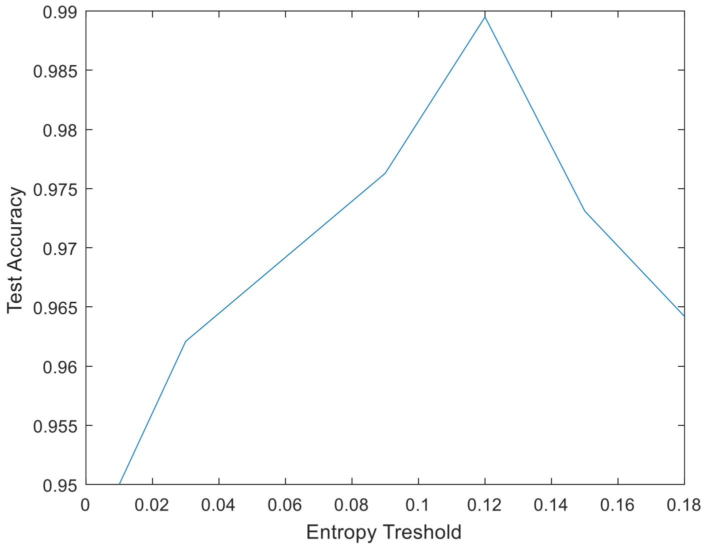
		Predicted Class		
		Class 1	**Class 2**	**Class 3**
True Class	**Class 1**	71	0	2
	Class 2	0	176	1
	Class 3	11	20	3147

c) In the class we have stated that unbalanced data may have difficulties in the minority class. In our example I have performed oversampling for the minority classes in order to make the data balanced, but I did not find out much difference in the accuracy comparing with the previous unbalanced data. We do not need normalization for this dataset since the feature magnitudes are not that different.

overall_accuracy = 0.9895 , class_based_accuracies = [0.9825 0.9756 0.9920]

d) **3-fold cross validation** : I have randomly splitted the data into three, and I have used 2 of them for training and the other one is testing the data. I have done the testing 3 times and calculated the average accuracy.

e) My parameter was entropy threshold, I have changed with respect to the test accuracy. I have selected the parameter which maximizes the test accuracy.



Part2)

I have used MATLAB Statistics and Machine Learning Toolbox, which is really powerful since there are many algorithms implemented for both classification and regression problems. I have used a “fitctree” function to build a classification tree model. `tree = fitctree(X,Y)` where X is the input matrix and Y is the label vector. For the model one can define cost of misclassification, cross-validated decision tree, leave one out cross validation, maximum number of splits and so on.

Another function to be used is “predict” where `Ynew = predict(tree,Xnew)` . Again, Ynew is the predicted labels “ Ynew” from “Xnew” data using the model “tree”. There are many parameters one can change, such as stopping criterion, optimization criterion.

I wanted to experiment pruning level versus the training data accuracy to understand the behaviour of pruning. For that, I have used different levels of pruning from Level 1 to Level 8, and calculated training accuracy for all pruning levels. I have obtained the following by plotting:

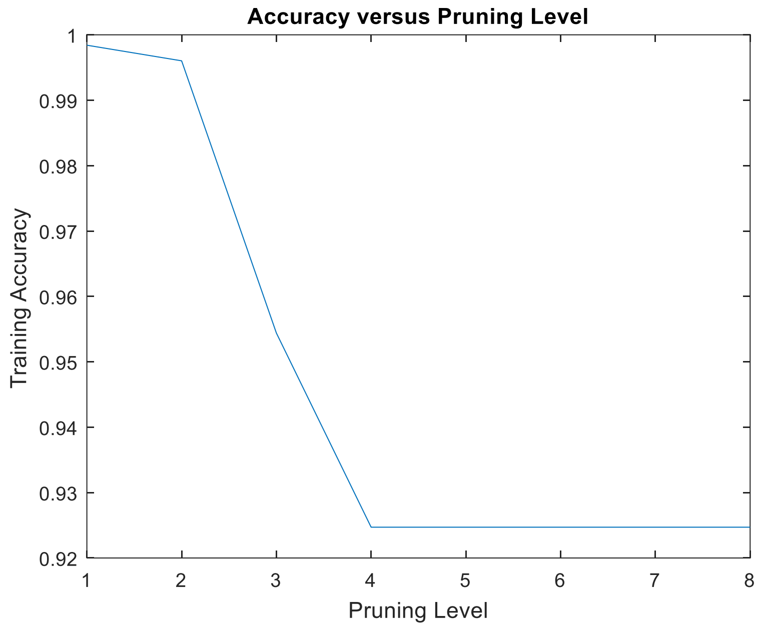


Figure 1: Accuracy versus Pruning Level

From this graph, it can be said that as we increase the pruning Level 4 the training accuracy decreases and after Level 4 it is constant. I think this level depends on mostly data and class distribution on the data. By using the optimum value of Level1, I have calculated overall accuracy, class-based accuracies and confusion matrices for both training and test data.

For that I have obtained:

Training:

		Predicted Class		
		Class 1	**Class 2**	**Class 3**
True Class	**Class 1**	91	0	2
	Class 2	0	191	0
	Class 3	0	1	3487

train_accuracy = 0.9992, class_based_accuracies_train = 0.9785 1.0000 0.9997

Testing:

		Predicted Class		
		Class 1	**Class 2**	**Class 3**
True Class	**Class 1**	73	0	0
	Class 2	0	175	2
	Class 3	3	0	3175

test_accuracy = 0.9985, class_based_accuracies_test = 1.0000 0.9887 0.9991

Part 3)

1) Diagnosis of ovarian cancer using decision tree classification of mass spectral data

In [1], the aim is is to perform discriminate ovarian cancer from benign disease and healty using classification and regression tree (CART) which is commercially available algorithm biomarker patterns software (BPS). 139 patients with either ovarian cancer, benign pelvic diseases, or healthy women are examined using the BPS software. 122 protein clusters or features are used. For the stopping criterion they have used gini criterion and perform 10-fold cross validation analysis. The paper calculates sensitivity in order to determine the performance of the decision tree for the learning set, cross-validation, and test set. They have found % 94.9 for the learning set, %84.6 for cross-validation and %80 for the test set. They found the decision tree effective comparing to the other classification methods.

2) Predicting corporate financial distress based on integration of decision tree classification and logistic regression

In the paper [2] , the aim is to improve the financial distress model. According to the paper, there is a need for the executives of the firms to estimate financial distress possibility in the short or long run. I t has collected 100 listed companies as the initial samples. A total of 37 features from the samples are available, they have used principle component analysis (PCA) to find the suitable feautures. The decision tree (DT) classification methods (C5.0, CART, and CHAID) were used.

- C5.0 model gives training accuracy of %99.25 and test accuracy of %97.01
- CART model gives training accuracy of %96.24 and test accuracy of %95.83
- CHAID model gives training accuracy of %97.74 and test accuracy of %92.29

References:

[1] Vlahou, Antonia, et al. "Diagnosis of ovarian cancer using decision tree classification of mass spectral data." *BioMed Research International* 2003.5 (2003): 308-314.

[2] Chen, Mu-Yen. "Predicting corporate financial distress based on integration of decision tree classification and logistic regression." *Expert Systems with Applications* 38.9 (2011): 11261-11272.
http://www.sciencedirect.com/science/article/pii/S0957417411003976