Technische Hochschule Ingolstadt, M-APE, Prof. A. Schiendorfer Lukas Lodes M.Sc.

Intro to AI and Neural Networks (Summer 2023)

Assignment 04

Exercise 1 *(Questions on Decision Trees)*

a) If a decision tree is overfitting the training set, is it a good idea to try decreasing its depth?

b) If a decision tree is underfitting the training set, is it a good idea to try scaling the input features?

c) If you hadn't learned about entropy, gini index, or misclassification rate, what metric would you design to quantify "pureness" in datasets? Show its evaluations on some small example distributions like $(0.8, 0.2)$ or $(0.25, 0.25, 0.25, 0.25)$.

Exercise 2 *(Building a Decision Tree)*

Imagine you are a newspaper deliverer that encounters a number of dogs on your tour. Some of them (try to) bite, some of then only bark. The dogs are described by the following binary features: Heavy, Smelly, Big and Growling. Consider the following set of examples:

Heavy	Smelly	Big	Growling	Action
No	No	No	No	Bark
No	No	Yes	No	Bark
Yes	Yes	No	Yes	Bark
Yes	No	No	Yes	Bite
No	Yes	Yes	No	Bite
No	No	Yes	Yes	Bite
No	No	No	Yes	Bite
Yes	Yes	No	No	Bite

Round all values within this exercise back to 4 decimal places. Note that we've prepared the dataset as a NumPy array in `dogtree.ipynb`. You may use this notebook for your calculations or work out the answers manually – depending on your liking.

a) What is the entropy of the target value 'Action' in the whole dataset?

b) Which attribute would the ID3 algorithm choose to use for the root of the tree? What is its information gain?

c) Draw the full decision tree that would be learned for this data using ID3 without pruning.

d) Suppose three new dogs appear in your round as listed in the table below. Classify them using the decision tree from the previous question. What is the accuracy of your tree on this test set?

Name	Heavy	Smelly	Big	Growling	Action
Buster	Yes	No	Yes	Yes	Bark
Pluto	No	No	No	Yes	Bark
Zeus	Yes	Yes	No	No	Bite

e) Repeat the steps of the previous four subtasks using the **Gini coefficient** as splitting criteria. Are there differences between the resulting trees and their performance on the above test set? Think about the reasons.

Based on Machine Learning at VU University Amsterdam https://mlvu.github.io/