Asam Mahmood – z23050331

CAP 5615 Introduction to Neural Networks

Assignment 2

Question 1 [2 points: 0.25/each]: Please use your own language to briefly explain the following concepts:

•Training data vs. test data: Training data is used to help create a model and the test data is used for helping to validate the model.

• Decision Trees: Decision trees in machine learning are a type of method/tool used for creating predictions and used for classifications. It has nodes that represent a type of validation on a certain feature and the branch represents the results of that validation the final node of a branch is the leaf node and usually holds a class for labeling a result. When you draw out the whole flow for each node it ends up looking like a type of flow chart or a type of tree. This when you look at the final output graphically can remind you of a giant if/else statement.

• Overfitting: Overfitting is a result of training a model overly on the training data. It does not recognize patterns very well and will separate the data completely without considering other features. It is considered a poorly performing model when overfitting happens.

• Prepruning and Post pruning for decision tree learning:
Pre pruning is a process that is stops the growth of a tree and helps prevent overfitting. It checks the cross-validation error on each splitting if the error does not get small enough then the growth of tree stops. Post Pruning is a process when the decision tree is built out completely then its cut back down. This may sometimes result in an overfitted tree.

• Entropy: Entropy is the measure of a randomness in the data. The way to look at it is if the higher entropy value the greater the difficulty to get a solid conclusion from results. It is used in creating decision trees and helps the process of where to split the data. A formula representation is below

$$Entropy\ (p) \ = \ - \sum_{i=1}^{N} p_i \ log_2 \, p_i$$

• Information Gain: Information gain is how much a feature/attribute gives us relevance/information about a classification. It is also valued calculated by comparing entropy results before and after data splitting. Formula can be viewed below:

Information Gain = Entropy after splitting data - Entropy before splitting data

• Information Gain Ratio: Gain ratio is used to normalize the information gain of a feature/attribute against the entropy of that feature.

Gain Ratio = Information Gain / Entropy

• Gini-Index: Gini Index is used to calculate the probability of a specific feature that is wrongly classified when chosen randomly. The Gini Index goes from 0 to 1. Closer to 1 depending on use case means higher chances of incorrectly/correctly classifying feature.

$$Gini\,(P) = \sum_{i=1}^{n} p_i\,(1 - p_i\ ) = 1 - \sum_{i=1}^{n} (p_i\,)^2$$

Question 2 [2 pts]: The following figure shows a toy dataset with two numeric attributes/features (x1 and x2) and nine types of instances (color coded using different shapes). The sub-figure on the right panel shows a constructed decision tree from the toy dataset. Please explain

• Roles of interior nodes vs leaf node of the decision trees [0.5 pt].

Interior nodes have branches and are part of inner workings of a decision tree. They split and have paths to sub-node or to root. But they always end to a final node which we call the leaf node.

A leaf node is the final node that is reached when traversing the tree it can also be viewed as the result of the decision tree's classification on certain feature. There is no splitting of node on leaf nodes

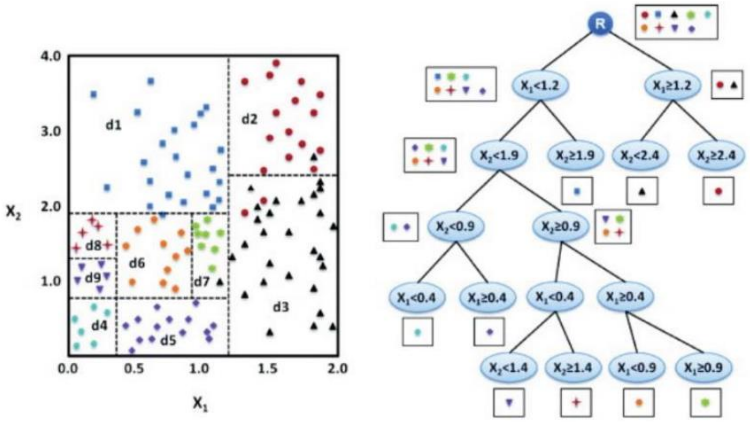
• Explain the process of building a decision tree though recursive partitioning of the feature space [1.0 pt].

When creating a decision tree through recursive partitioning of the feature space, a selected feature is used to segment instances in to at least two or more branches by using certain features. The process continues with one element being chose each time in every iteration. This continues until it keeps creating sub trees and smaller sub trees until a final leaf is created.

• What are the meaning of the dashed lines on the left panel, and how does each dashed line correspond to the node on the right panel [0.5 pt].

The dashed lines is used to determine which branches left or right the dot instances for every single feature belong to depending on the conditional value at that instance. Each Dashed line represents an inner node. The value inside the node determines which features and threshold used for splitting.

Asam Mahmood – z23050331



(a)

(b)

ID	Days	Outlook	Temperature	Humidity	Wind	Class
1	Mon	Sunny	Hot	High	Weak	No
2	Tue	Sunny	Hot	High	Strong	No
3	Wed	Overcast	Hot	High	Weak	Yes
4	Thu	Rain	Mild	High	Weak	Yes
5	Fri	Rain	Cool	Normal	Weak	Yes
6	Sat	Rain	Cool	Normal	Strong	No
7	Sun	Overcast	Cool	Normal	Strong	Yes
8	Mon	Sunny	Mild	High	Weak	No
9	Tue	Sunny	Cool	Normal	Weak	Yes
10	Wed	Rain	Mild	Normal	Weak	Yes
11	Thu	Sunny	Mild	Normal	Strong	Yes
12	Fri	Overcast	Mild	High	Strong	Yes
13	Sat	Overcast	Mild	Normal	Weak	No
14	Sun	Rain	Hot	High	Strong	Yes
15	Mon	Rain	Mild	High	Strong	No

Table 1

Question 3 please view html for solution

Question 4 please view html for solution

Question 5 please view html for solution

Question 6 please view html for solution

Question 7 please view html for solution

Question 8 please view html for solution