Machine Learning - Assignment 1 - S Deepak Narayanan, 16110142

Q1.

a. No. Choosing date is not a good choice at all. It doesn't actually cover any necessary and needed information for a decision tree. It will be chosen as the root because of its extremely high information gain. We would actually have a tree of depth 1 with perfect classification. This would be not generalizing well to unseen examples because the date doesn't talk about any conditions involved, at all. In order to avoid such an eventually, a potential method would be to select feature based on a metric different from information gain. Tom Mitchell's book mentions Gain Ratio as such alternative. Also, we are not covering any dependencies as such when we use the day as a feature for predicting whether or not to play tennis. This is also a factor that needs to be taken into consideration.

b.

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	High	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Suppose we do not have Outlook in the above diagram for D3, as mentioned in the question. They, while computing Information Gain, we actually had no contribution from Overcast part of Outlook feature when

considering it while deciding the hierarchy of the tree. In this particular case, we could still learn the decision tree by interpolating the missing Values. We could take a majority poll to fill the slot. We could take the value to be sunny, overcast, and rain. Sunny has 5 occurrences of the assumed 13 sample dataset (assuming D3's outlook is missing). Rain has 5 occurrences. Overcast has 3 occurrences. Now we can weigh these by the probability of playing that day. Overcast has high chances as all the other times we've played it has been overcast. We can compute the probabilities as the product of the probabilities of playing while raining and the probability for rain. (5/13*⅘ for sunny, 3/13 for overcast and ⅗*5/13 for rainy). We'd have more chances of overcast and rain. Overcast would leave us with the old decision tree itself (done in class). If we take rain or sunny, we'd have a net increase in information gain of 0.07 in the case of rain and by a similar amount for sunny. This would imply that we'd still have the same decision tree in this particular example as there isn't a lot of change in entropy. We'd have the same decision tree as earlier.
Note: This is in alignment with what Tom Mitchell discusses in the textbook. We're only considering probabilistically a little more information, which in this case is whether it is a 'Yes' or a 'No' for a particular day.

Q2 - Q7 - All have been answered in the Jupyter Notebook here.

References:

1. https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/
2. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
3. Tom Mitchell, Machine Learning
4. https://www.python-course.eu/Decision_Trees.php
5. https://www.python-course.eu/Regression_Trees.php