ID3

Deadline: 16 november 2023, 23:59
Total number of points: 1.5

1. Preprocessing

a. Provide a brief description of the dataset. What are the attributes, what is the target attribute? What is the purpose of the dataset? Specify which attributes are discrete and continuous.
b. Identify the NaN's (Not a Number) in your dataset. Remove the rows that contain such values.
c. Calculate the mean and variance for each numerical attribute.

2. Probabilities, Information Theory

a. Write a function `compute_probabilities` that calculates the probability mass function of a discrete attribute. Apply the function to the discrete attributes from your dataset.
b. Write a function `calculate_entropy` that computes the entropy of a random variable given its probability distribution. Calculate the entropy for each discrete attribute.
c. Write a function `calculate_conditional_entropy` that computes the conditional entropy of two random variables. Calculate $H(Y|X)$, where Y denotes the target attribute, and X one of the discrete attributes from the dataset.
d. Write a function `calculate_information_gain` that computes the information gain of two random variables. Calculate the information gain in the previous example.

3. ID3

a. Write a function `find_root_node` that finds the attribute picked by ID3 as root. The function should return a tuple with the name of the attribute and the information gain. Use the functions created at point 2. What is the attribute identified as a root node?
b. Write a function `id3_discrete` that implements the ID3 algorithm for the discrete attributes. The function should return a dictionary following this structure

```
{"node_attribute" : {
        "n_observations" : {value1 : v1, value2 : v2, ... , valuen : vn},
        "information_gain" : ig_value,
        "values" : {
                node_attribute_value1 : {
                        "node2_attribute" : ...
                },
                .
                .
                .
                node_attribute_valuen : {
                        "node2_attribute" : ...
                }
        }
}}
```

c. Run `id3_discrete` on the dataset containing only discrete attributes. Compare the results with the ones from `sklearn`. (make your comparison as thorough as possible)
d. Write a function `get_splits` which, given a continuous attribute and the labels, will identify the splits that could be used to discretization of the variable. Test your function on an example.
e. Write a function `id3` that implements ID3 on the entire dataset, both continuous and discrete attributes. The function should return a dictionary similar with the one above. Compare the results with the ones from `sklearn`.
f. Modify the two ID3 functions such that they will allow pruning. Use TWO methods of pruning, one of which should be based on the depth of the tree.
g. Aim to avoid overfitting and find the best pruning values that will lead to the best tree. Use cross-validation in your approach and justify your reasoning thoroughly.

Notes:

- make sure you include the functions implemented in the previous points! (for example, for calculating the information gain use the entropy function you defined earlier)
- the ID3 implementation should be done by you, do not use framework such as `sklearn` to build the tree (except for the comparison).
- the Assignment should be written in a Jupyter Notebook that will be sent via email