Homework 2: Information Gain

CS 447: *Machine Learning*

Instructor: Collin Engstrom

Due: Friday, Oct 27th, by **11:59 p.m.**

Directions: Complete the following programming question using the slides and textbook to help you. You may do this programming assignment in pairs. You may use the code we wrote together in lab3 to get started. You should be able to copy it into the notebook I've provided for this homework. There is no written portion for this assignment.

Finally, as usual, ***DO NOT Google for code!!!*** *Use your own resources (textbook, slides, notes, etc.) or talk to me if you need help. If I catch plagiarized work, it will be handled accordingly.*

1. Information Gain

Using Jupyter notebooks and the Python programming techniques we've been covering in class, complete the Information Gain programming assignment. The skeleton code is available on D2L. Please note the following in your implementation:

1. You will fill in all the necessary code to make the `InfoGain` code run. Note that in certain places I've left hints in the code. All input and return values are also stated in the function headers. Example outputs are currently in the Jupyter notebook to help check your work.
2. We will stick with *binary classification* (i.e., two class labels: "0" and "1" for survival status).
3. The code will expect both the `titanic_new.csv` and `attribs` files in the same directory as your notebook. I have included sample files for you to work with in the zip archive. You should be able to run your `InfoGain` implementation on these files and get the same output as what's currently in the Jupyter notebook. Please let me know if yours doesn't match, as I may have a bug.
4. You'll also want to closely inspect the `titanic_new` file, since I had to modify it a bit for our information gain task. For instance, a few non-categorical features have been converted to categorical (e.g., age is now "A" for adult or "C" for child). The possible values each attribute can take on are listed in the `attribs` file, along with the class label (at the very end). Note that missing attribute values mean that they weren't found in the titanic dataset (e.g., "7" for Siblings/Spouse).
5. You shouldn't need to convert any numeric feature values to integer or float type. They should all be fine if you leave it in string format. (You will still need numbers to calculate entropy, conditional entropy, and information gain.)
6. You may **NOT** import any libraries (e.g., Pandas, Sci-Kit Learn, etc.) beyond what is already included in the skeleton code.

7. I will be calling all of your functions with my own test code and input files, so you must **NOT** change the function headers from what is in the skeleton code. Doing so will break the execution of your code, and I will deduct points.
8. You must also take care to return the expected value(s) from all of your functions.
9. There is test code in a separate cell beneath each of the functions. This will allow you to individually test the functionality of the functions as you build them. The functions should build on one another until you've reached the full information gain calculation in the `get_info_gain()` function.
10. You will likely need to refer to various APIs and tutorials online related to using Python. Please be certain that you are **NOT** copying code, though. Doing so constitutes cheating, as does copying code from others in class. If I catch instances of either of these, I will handle them in accordance with university policy.
11. You should also **NOT** consult other implementations of Information Gain, as the final product you submit must be your own. You may, however, use notes, labs, and any other material we've used for class.
12. Since there are likely some details I've forgotten to include in this document, I've created a D2L discussion called "HW2." To access it in D2L, use the "Communications" menu and select "Discussions." You are free to post questions and discuss your ideas with each other and myself. Please be aware, though, that you must **NOT** discuss things at the implementation level. So, for instance, a discussion at the algorithmic level is permissible and encouraged, while a discussion of specific loops, if/else-statements, and so on should not be included. (You may privately talk to me about these matters if you're stuck, however.)

When you are finished, please submit your Jupyter notebook (.ipynb) file on D2L. I don't need any other files of yours.

Please make sure you are submitting the correct version of your code!!!!