CSCI 4520 - Introduction to Machine Learning

Homework 1

Due Date: 01/31/2020

Instructions: There is one written question and a coding question on this assignment. You need to turn in the hardcopy of your answers (excluding the programming question) at the beginning of the class. Also, submit both your written answers (as a pdf) and your implementation to the Dropbox Folder on Folio. Please show your work for all questions.

Your written answers may be typed. You're welcome to type your solutions in LaTex if you know how. If you don't know LaTex but want to type, there a number of markdown editors with real-time preview and equation editing. Here are two: stackedit, marxi. Writing your solutions by hand is also fine as long as they're neat.

Decision Tree [20 Points]

Decision tree learning is a form of supervised inductive learning. A set of training examples with their correct classifications is used to generate a decision tree that, hopefully, classifies each example in the test set correctly. In this problem you are asked to learn the decision tree by applying the ID3 algorithm on the dataset summarized in the Table below.

Example	Type	Price	Buy
CD1	HipHop	Expensive	Yes
CD2	Rock	Cheap	Yes
CD3	Rock	Expensive	Yes
CD4	HipHop	Cheap	Yes
CD5	Jazz	Cheap	Yes
CD6	Rock	Expensive	No
CD7	Jazz	Expensive	No
CD8	Jazz	Cheap	No
CD9	HipHop	Expensive	Yes
CD10	Jazz	Expensive	No
CD11	Rock	Expensive	No
CD12	Jazz	Cheap	Yes
CD13	Rock	Expensive	No

In this problem, you are learning the concept of whether or not to purchase a music CD. The dataset consists two attributes `Type, Price` and a binary label `Buy` to describe the examples:

`Type possible values: Rock, Jazz, HipHop`

`Price possible values: Cheap, Expensive`

Apply the decision tree ID3 algorithm using the concept of ***information gain*** to train the dataset, and draw the final decision tree. Show all your work.
After training is done, evaluate your decision tree and measure the accuracy using the test dataset shown in the Table below.

Example	Type	Price	Buy
CD1	Rock	Cheap	Yes
CD2	Jazz	Cheap	No
CD3	Jazz	Expensive	No
CD4	Rock	Expensive	Yes
CD5	HipHop	Expensive	Yes

Programming Decision Tree [30 Points]

In this Assignment, you will utilize `scikit-learn`, a python machine learning library, to train a decision tree on the training-dataset and test it on our "buy-CD" test-dataset.

Before starting this programming exercise, you will need to make certain that you are working on a computer with the following particular software:

- python 3.8
- numpy
- scikit-learn

One of the easiest way to install all of the aforementioned packages is to use Anacanda.

To make sure that you have the libraries, run the following code in the
python interpreter (you should just be able to cut & paste the code):

```
from sklearn import tree
X = [[0, 0], [1, 1]]
y = [0, 1]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X,y)
clf.predict([[2.,2.]])
```

If this code runs without error and gives you the following output:

```
array([1])
```

For this question, you have to either write a python program or use Jupyter Notebook and perform the following tasks:

- load the training dataset
- train a decision tree
- predict the label for the test dataset
- evaluate your model
- plot the decision tree using `graphviz` and save it as a `pdf` file