1 ML Practical Test

1.1 Timing

This practical problem must be solved in four hours at most.

1.2 Introduction

Decision trees are very useful tools to extract relationships between data. They are broadly used today in the context of data-mining activities.

In short, decision trees predict the value of a given attribute providing the values of the rest of the attributes for the same instance.

For example, one very popular data set is the one usually referred as *Census Income* (http://archive.ics.uci.edu/ml/datasets/Adult), which presents information from the US census (age, education, marital status) along with the income for those people (<= $50K/year, > $50K year).

Age	Workclass	education	Marital-status	Occupation	Relationship	Race	Sex	Native-country	Income
39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	United-States	<=50K
50	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	<=50K
38	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	United-States	<=50K
53	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	United-States	<=50K
28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	Cuba	<=50K
37	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States	<=50K
49	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	Jamaica	<=50K
52	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
31	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	United-States	>50K
42	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
37	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	Black	Male	United-States	>50K
30	State-gov	Bachelors	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	India	>50K
23	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	United-States	<=50K
32	Private	Assoc-acdm	Never-married	Sales	Not-in-family	Black	Male	United-States	<=50K
40	Private	Assoc-voc	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	?	>50K
34	Private	7th-8th	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	Mexico	<=50K

The algorithm to create a decision tree for a given **output attribute** given an **input dataset** is rather simple[1]:

1. If all the output values are the same in dataset, return a leaf node that says "predict this unique output"
2. If all input values are the same, return a leaf node that says "predict the majority output"
3. Else find attribute X with the highest information gain (IG).
4. Suppose X has n_x distinct values (i.e. X has arity n_x).
    a. Create and return a non-leaf node with nx children.
    b. The i-th child should be build recursively with a dataset whose records are the ones for which X= i-th distinct value of X.

The information gain of an attribute Y given X, $I(^{Y}/_{X})$, is defined as

$$IG(^{Y}/_{X}) = H(Y) - H(^{Y}/_{X})$$

[1] https://www.autonlab.org/_media/tutorials/dtree18.pdf

Where $H(Y)$is the entropy of attribute Y and is defined by the expression:

$$H(Y) = -\sum_{j=1}^{m} p_j \log_2 p_j$$

Meaning the mínimum number of bits to transmit a stream of symbols drawn from Y's distribution. A high value of the entropy means Y is a very uniform distribution, whereas a low value indicates Y has a varied (peaks and valleys) distribution.

$H(Y/X)$ is the conditional entropy of Y given X, defines as

$$H(Y/X) = \sum_{j} P(X = V_j) H(Y/X = v_j)$$

1.3 Examples

Probably, some examples will help clarify the above expressions.

Suppose the following simple dataset[2]:

X (College Major)	Y (Likes "Gladiator")
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

$H(Y/X = v)$ is the entropy of Y among only those records in which X has value v. Then:

$$H(Y/X = Math) = 1$$
$$H(Y/X = History) = 0$$
$$H(Y/X = CS) = 0$$

Computing some probabilities from the above table and expressions, it turns out that

[2] From http://www.autonlab.org/tutorials/infogain11.pdf

v_j	$P(X = v_j)$	$H(Y/X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y) = 1$$

$$H(Y/X) = 0.5 \cdot 1 + 0.25 \cdot 0 + 0.25 \cdot 0 = 0.5$$

And therefore

$$IG(Y/X) = 1 - 0.5 = 0.5$$

1.4 Expected Outcome

The candidate should:

1. Write a program, preferably in C++, with the following command-line:

   ```
   dectree input-file output-attribute-index
   ```

   Where *input-file* is the path of the input file in CSV format and attribute-index is the index of the attribute (starting at 0) whose output is to be predicted. All the fields in the input file should be handled as categorical (non-numerical) attributes.
2. The program should build a decision tree using the rules provided above. The program should print out using the standard output a set of rules defining the built decision tree in the format:

   ```
   Attr1 = Val1 & Attr2 = Val2 … => AttrN = ValN
   ```

3. In order to have partial results should the candidate not finish the code completely, some partial results may be delivered:
   a. Functions/classes to parse the input file.
   b. Functions/classes to compute the information gain for a set of records.
   c. Functions/classes for computing base cases in the process of building the decision tree.

1.4.1 Constraints

1. The program should be bug free.
2. The candidate can use STL containers, built-in functions and any other construct he/she may find useful.
3. Whenever justified, the candidate can use open source libraries from third parties.

4. As a plus, the candidate may provide a parallel version of the program in which the decision tree is computed using several threads.

1.4.2 Testing

1. The candidate can use several test files provided as input in order to verify the program being built.

adult.zip