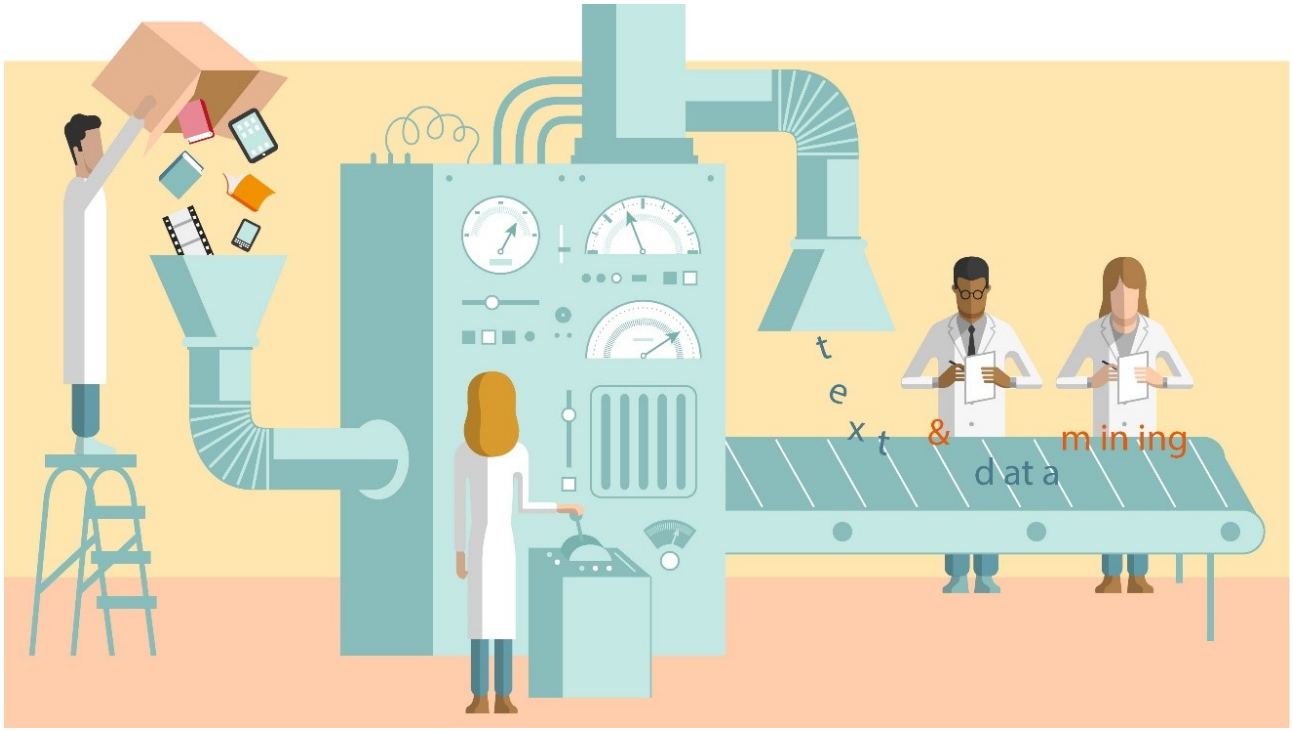
Amirkabir University of Technology

Computer engineering and it department

Principles of Data minig

Assignment 3



Authors:

Mohammad Navid Shahsavari

Yasaman Mirmohammad

Sina Malakouti

Mohammad Hossein Goharinejad

Under Supervision of:

Prof. Ehsan Nazerfard

April 28, 2019

1 Theory

Question 1

Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?

Question 2

What is boosting? State why it may improve the accuracy of decision tree induction.

Question 3

It is important to calculate the worst-case computational complexity of the decision tree algorithm. Given data set:D, the number of attributes:n, and the number of training tuples:|D|, show that the computational cost of growing a tree is at most n.|D|.log.|D|

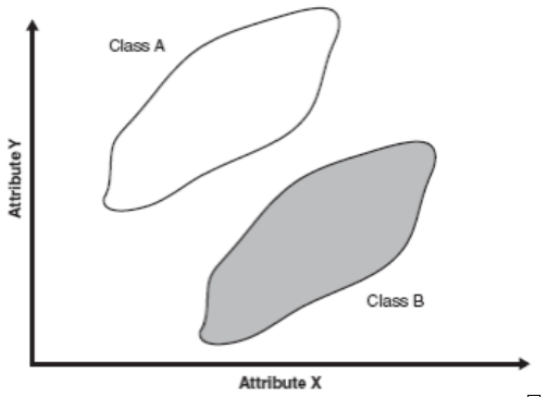
Question 4

The following table consists of training data from an employee database. The data have been generalized. For example, "31 ... 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

Department	Status	Age	Salary	Count
Sales	Senior	31...35	46K...50K	30
Sales	Junior	26...30	26K...30K	40
Sales	Junior	31...35	31K...35K	40
Systems	Junior	21...25	46K...50K	20
Systems	Senior	31...35	66K...70K	5
Systems	Junior	26...30	46K...50K	3
Systems	Senior	41...45	66K...70K	3
Marketing	Senior	36...40	46K...50K	10
Marketing	Junior	31...35	41K...45K	4
Secretary	Senior	46...50	36K...40K	4
Secretary	Junior	26...30	26K...30K	6

(a) How would you modify the decision tree algorithm to take into consideration the count of each generalized data tuple

(b) Consider the following the dataset, explain how the decision tree classifiers would perform on this data set. How we can change the splitting idea so decision tree would fit the data well? (Hint : Oblique Split idea)



Question 5

Suppose we have a dataset consists of 2 features (X and Y) and 3 different Classes (A, B ,and C). we want to train a decision tree on this dataset. So, It has to decide whether to split a node

$$\mathrm{N} = \begin{pmatrix} A & 100 \\ B & 50 \\ C & 60 \end{pmatrix}$$

into one of two possible splits. The first candidate split splits the node N based on feature X :

$$N_{X_1} = \begin{pmatrix} A & 62 \\ B & 8 \\ C & 0 \end{pmatrix}$$

$$N_{X_2} = \begin{pmatrix} A & 38 \\ B & 42 \\ C & 60 \end{pmatrix}$$

The second candidate split splits the node N based on feature Y :

$$N_{Y_1} = \begin{pmatrix} A & 65 \\ B & 20 \\ C & 0 \end{pmatrix}$$

$$N_{Y_2} = \begin{pmatrix} A & 21 \\ B & 19 \\ C & 20 \end{pmatrix}$$

$$N_{Y_3} = \begin{pmatrix} A & 14 \\ B & 11 \\ C & 40 \end{pmatrix}$$

1. Compute the gain in information gain for the two candidate splits. Then choose the appropriate Feature for splitting.
2. Compute the gain in Gini index for the two candidate splits.Then choose the appropriate Feature for splitting.

You can get more information here.

Question 6

Suppose that we have the following data set. Where A, B, and C are the features and Y is the class attribute :

A	B	C	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	0

1. Compute Information Gain for each possible split at the root. Which attribute would ID3 select at this point?
2. Write down the complete decision tree generated using the ID3 algorithm for this data set. (Note that over-fitting is not important for this part. i.e. no need to prune the tree)
3. One way to prevent over-fitting is pre-pruning(top-down pruning). The idea is to start at root and prune splits when the criterion (e.g. information gain) gets smaller than some small threshold ϵ . What is the decision tree returned for $\epsilon = 0.0001$?
4. Another idea is post-pruning(bottom-up pruning). The idea is to start pruning after growing the tree. In this way, we start at leaves, and prune subtrees for which the information gain of a split is less than ϵ.What is the tree returned for $\epsilon = 0.0001$?
5. Discuss when would you choose post-pruning over pre-pruning and vice versa.
6. Does ID3 guarantee a globally optimal decision tree? Explain your answer.

Question 7

Show that accuracy is a function of sensitivity and specificity.

$$accuracy = sensitivity\frac{P}{(P+N)} + specificity\frac{N}{(P+N)}.$$

2 Implementation

Classification with Decision Tree and Random Forest Classifier:

1) Preparing Data:

In this part you will classify and predict data using two strategies. The implementation of Decision tree and random forest tree in sklearn packages expects digit encoded categorical data, for example for age data one could have digit 0 for ages from 0 to 10, digit 1 for ages from 11 to 20 and so on.The Provided Datasets have both numerical and categorical data.
Before feeding the the data into algorithm you need to turn numerical data into categorical data and then encode all the categorical data with digits. To turn numerical data into categorical data you can split the numerical data into ranges and assign categories to each range, as an example for age you can assign category "teenager" to ages from 10 to 20 and category "adult" to ages from 20 to 40 and so on. Next, you have to encode these categories. To do so you can use OneHotEncoder from sklearn packages. Visit here for more info on encoding categorical data.

2) Classifying the Data:

With the Data ready, it's time to start classifying and predicting the data. Before doing so, make sure you visit here to get a better idea on what you're gonna do.
First you need to split your dataset into a training set with 80 percent of data and a test set with 20 percent of data. The training set is used to actually train the algorithm and test set is used to see you well the trained algorithm may perform.
The Decision Tree and Random Forest Classifier use different parameters to classify the data. In this Part , using each dataset on at a time, you need to train the two algorithms with parameter "criterion" set to "gini" and then "entropy", visualize the Decision Tree, report the accuracy of each algorithm and classify the Unknown datasets provided to you. Next you need to find values for parameters "max_depth" and "min_samples_split" to increase the accuracy. you can do this manually or using grid search. After tuning the parameters repeat the previous steps and report your results.

3) Classifying with Weka (Bonus):

In this part you will use one of the famous tools for data mining called Weka. to classify and predict your data. You can visit here to learn how to use Weka. Weka expects the input data to be in ARFF format. So first you need to transform you CSV data into ARFF format. You can find instruction to do so here. After preparing your data, go to Weka explorer and load your data, next go to classify tab and choose J48 from tree section. Set the confidenceFactor parameter to 0.05 for first run and 0.35 for second run and classify the data each time. For each run visualize the learnt trees and report the accuracy as part of your report.

3 Caution!!!

•Report is an important part of your grade. So write it completely and explain your analysis. Your report is only accepted in 'pdf' format. Put it in "report" folder. (There is no force on the language of the report)
•Your codes should be written in python. Put them in "supporting material" folder.
•"income","poisonous" and "disease" are target features for Dataset1 , Dataset2 and Dataset3 respectively.
•You are provided with three classified datasets for training and testing the accuracy and three unknown datasets that you should classify. after classifying these unknown datasets you should save the vector of classes in a separate file for each dataset.
•Make sure that the visualized trees as well as the accuracy of trained models is presented both in your pdf report and in the python notebook when TAs run your code.
•Deadline of this assignment is til "22th of ordibehesht". you will lose 10 percent of your grade after that on each day of delay.
•Put all your folders and files like the sample format in a "zip" file and upload it on moodle(https://ceit.aut.ac.ir/courses/)
•Please upload your homework in this format:

```
9******_FirstnameLastname_HW1.zip
├── [directory] Report
│   └── 9******_FirstnameLastname_Report1.pdf
└── [directory] Supporting_Material
    └── codes.py
```