AMIRKABIR UNIVERSITY OF TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING AND IT

Principles of Data mining

Assignment 3

Kian Eliasi

MohammadMahdi Heydari

Armin Kazemi

Saeedeh Sadeghpour

Under the supervision of: Dr. E. Nazerfard

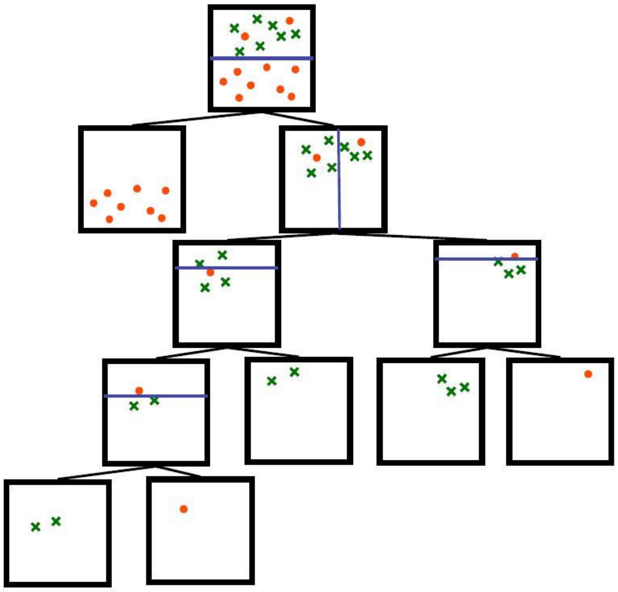
Spring 2020

Question 1

Explain what is “Gradient Tree Boosting” and “Random Forest” and compare them.

Question 2

Describe the problem with the following decision tree and explain how it can be fixed. What would be a general solution for this issue in decision trees?



Question 3

Describe a situation in which “accuracy” is not a good evaluation metric for a classifier. What other evaluation metric would you suggest?

Question 4

Suppose there are 25 base independent classifiers and each has error rate = 35%. What is the probability of wrong prediction for the ensemble classifier? Explain your answer.

Question 5

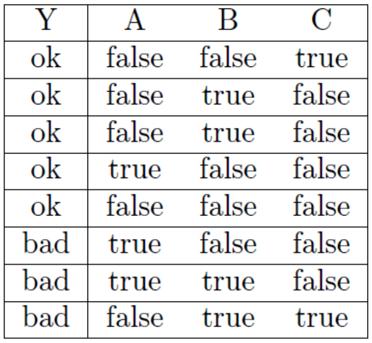
We want to make a decision tree for Super Mario which predicts if a mushroom is edible or not based on its characteristics. Here is our data:

Habitat	Cap Color	Cap Shape	Odor	Edible?
Woods	Red	Flat	None	**Poisonous**
Woods	Red	Flat	Foul	**Poisonous**
Grasses	Red	Flat	None	**Edible**
Leaves	Green	Flat	None	**Edible**
Leaves	White	Convex	None	**Edible**
Leaves	White	Convex	Foul	**Poisonous**
Grasses	White	Convex	Foul	**Edible**
Woods	Green	Flat	None	**Poisonous**
Woods	White	Convex	None	**Edible**
Leaves	Green	Convex	None	**Edible**
Woods	Green	Convex	Foul	**Edible**
Grasses	Green	Flat	Foul	**Edible**
Grasses	Red	Convex	None	**Edible**
Leaves	Green	Flat	Foul	**Poisonous**

A) Show which feature will be selected at each step by calculating information gain.
B) Draw the final decision tree.

Question 6

Suppose you are given a dataset with three Boolean features A, B, C and class attribute Y. You are going to train a Naïve Bayes classifier for this dataset.



Y	A	B	C
ok	false	false	true
ok	false	true	false
ok	false	true	false
ok	true	false	false
ok	false	false	false
bad	true	false	false
bad	true	true	false
bad	false	true	true

1. Write an expression for P(y|a,b,c) in terms of P(y), P(b|y), and P(c|y).
2. Given the observation: *A = false, B = true, C = false*, what prediction would the Naïve Bayes classifier make? Write down your calculations. What would be the posterior probability of this prediction?

Implementation 1: Decision Tree

In this part, you will classify and predict heart diseases using decision trees. You can see the dataset with description in /dataset/heart.csv.

1) **Preparing the data:** The implementation of decision tree in sklearn package 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. So, you have to convert numerical features to categories each representing possible value ranges. Then you need to encode non-numerical values with digits. The dataset has a number of categorical and binary features that require encoding.
2) **Splitting the data:** To see how well your classifier performs you need to test it with new entries. First you need to split your dataset into a training set with 80% of data and a test set with 20% of data. The training set is used to actually train the classifier and test set is used to measure it
3) **Training and classification:** Finally, your decision tree will classify the test dataset. The Decision Tree in sklearn has various parameters. Try different values for *criterion, max_depth and min_sample_split* parameters and train different classifiers. Visualize each decision tree, test them with the test set and report accuracy for each one.

Implementation 2: Decision Tree with Weka

In this part, you will use one of the most famous tools for data mining called Weka. 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 your CSV heart data (from the previous section) to 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 the first run and *0.35* for the second run and classify the data. For each run visualize the learnt trees and report the accuracy as part of your report.

Implementation 3: Naïve Bayes Text Classification

In this part, you will implement a Naïve Bayes classifier for classifying book reviews as positive or negative. The dataset is a collection of Amazon user reviews about a book available in /dataset/reviews_train.csv.

1) **Preparing the data:** You have to convert each review's text to a bag of words representation. So, the first task is to go through all reviews and construct a vocabulary of all unique words. In this step you have to ignore all the stop words given in the 'sw.txt' file. The words must be extracted by splitting text off the whitespace characters and it is recommended to convert all letters to lowercase. The next step is to get counts of each individual word for positive and negative classes separately to get *P(word|class)*.

2) **Classification:** In this step you need to go through all the negative and positive samples in the test set (/dataset/reviews_test.csv) and classify each one according to the parameters learned earlier. The classification should be done by comparing the log posterior (*P(X|Y)P(Y)* for both classes).

3) **Laplace Smoothing:** An issue with Naïve Bayes classifier is that if a test sample contains a word which is not present in the dictionary, *P(word|class)* equals to zero. To mitigate this issue, one solution is to employ Laplace Smoothing (it has a parameter α). Enhance your classifier by employing Laplace Smoothing. Compare classifier results with and without Laplace Smoothing and using different values for α.

CAUTION

- For each part, write your codes in a .py (or .ipynb) file naming with the number of parts, and put it in the “supporting material” folder.

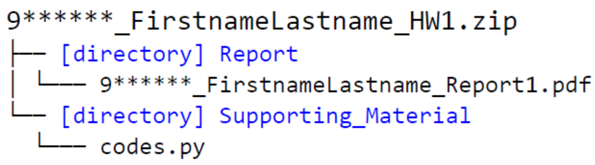
- Deadline is on 31th of Ordibehesht and you will lose 10% of your grade after that on each day of delay.

- 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)

- Put all your folders and files like the sample format in a “zip” file and upload it on moodle (http://courses.aut.ac.ir/)

- If you have any question regarding the assignment contact kian.elbo@gmail.com

Please upload your homework in this format:



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