CSCI6515 - Machine learning for Big Data: Assignment 1

Publish date: Aug 4, 2019

Due date: Oct 25, 2019 11:59PM ADT

In this assignment, you are going to read some text files and classify them according to their labels. The Reuters corpus is one of the most famous datasets for text categorization tasks. We provide a subset of this dataset on Brightspace. You apply these files to make your classifier. There is more information about this dataset available on http://disi.unitn.it/moschitti/corpora.htm. (Total 64 points)

1- Download zip file and extract it. Consider this data is a subset of full Reuters corpus to make it possible for you to process without the need of a powerful server.

> You have access to the following assets using your CSID that might be useful:
>
> 1- Bluenose: Bluenose.cs.dal.ca (undergrad and grads)
>
> 2-Hector: hector.cs.dal.ca (only grad students)
>
> 3-Gitlab: https://git.cs.dal.ca

2- Each file contains some XML files. Explore XML files and find a list of all fields available there.

3- Write a function extract a Pandas's Dataframe containing: (1) **headline, (2) text, (3)** bip:topics,(4) dc.date.published, (5) itemid, (6) XMLfilename (4 points)

4- Write a python function to find all the possible values for bip:topics. Consider that each news can belong to more than one topic. (4 points)

5- Write a function to prepare your text data by methods such as removing stop words. You are allowed to use the NLTK library. You can find more information here: https://www.nltk.org/. (4 points)

6- Extract features from the text using any approach you like. Write a function that input the Dataframe in step 3 and generates a new Dataframe of your features and labels. (4 points)

7- Divide your data into a training and test set. You can use any method such as cross-validation. You need to provide a reason why you decide so here. (4 points function, 4 points explanation: 4+4=8 points)

8- Write a function to get the Dataframe of step 6 and a set of parameters to return a trained classifier to classify all labels that you get in step 4.(4 points)

9- Write a function to evaluate the quality of your classifier (like accuracy, F-score, AUC, ...). Explain why you think this function is the best choice. (4 points function, 4 points explanation: 4+4=8 points)

9- Generate five different classifiers (Random Forest, Decision Tree, Linear Regression, Neural Network, and SVM) using step 8. Tune them up for the best parameters. Find the best classifier. Explain why. (4 points each classifier, Tune up 4points, explanation of best classifier 4 points: 4X5+4+4=28 points)

10- Go to Brightspace and upload your notebook containing all of your work under the assignment 1 section.

11- Go to Brightspace and under the Quiz of Assignment 1 section, upload your answer for each question separately.

Notes:

- Delay to submit from 5 minutes to 24hours will deduct 25% of your mark. Between 24 hours and 48 hours deducts 50% of the mark and between 48 hours 52hours deducts 75%, and more than 52hours deduct 100% of the mark.
- each question will be evaluated by
    - **Excellent**: Material is clearly grasped and convincingly presented. (4/4points)
    - **Good**: Most of the key ideas are present and clearly presented. (3/4points)
    - **Satisfactory**: Much of the assignment may be a summary or paraphrase of the material. There may be inaccuracies in the presented content. (2/4 points)
    - **Below Standard**: Has not grasped the material enough to answer the question effectively. (¼ points)
    - **Wrong:** No Answer, unrelated, or does not make sense. (0/4 points)
- Assignments are individual works.
- If you get the help of TA in the learning center, explain which part.
- If you have any questions about this assignment, post your question on the discussion form of assignment 1 on Brightspace. Link to discussion on Brightspace here (you need to login first)

Goodluck