Data Mining

Homework 3

Due: 12/1/2020, 23:59

Instructions

You must hand in the homework electronically and before the due date and time.

This homework has to be done by each **person individually**.

Handing in: You must hand in the homework by the due date and time by an email to Andrea (`mastropietro@diag.uniroma1.it`) that will contain as attachment **(not links to some file-uploading server!)** a .zip file with your answers. The filename of the attachment should be
`DM_Homework_1__StudentID_StudentName_StudentLastname.zip`;
for example:
`DM_Homework_1__1235711_Robert_Anthony_De_Niro.zip`.
The email subject should be
`[Data Mining] Homework_1 StudentID StudentName StudentLastname`;
For example:
`[Data Mining] Homework_1 1235711 Robert Anthony De Niro`.
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit then contact Andrea.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

For information about collaboration, and about being late check the web page.

Problem 1. This homework will be about neural networks. In particular, you have to develop a neural network able to perform text classification. You have to use PyTorch. The techniques seen in class during the lab session, such as feedforward neural networks and convolutional neural nets can be used (especially CNNs), even though the literature demonstrated RNNs (Recurrent Neural Networks) to be the most powerful neural nets to work with text. So, for the most adventurous, I advise to develop an RNN for the homework (not compulsory anyway). The task you are asked to perform is genre music classification using the lyrics of the songs. The dataset your are going to use is the MetroLyrics datasets. It contains lyrics from over 380K songs divided into 11 genres, and its available here: https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics/data.

The homework is divided into the two parts:

Part 1: Feature Extraction and Classification

1. Download the dataset and perform feature extraction from the lyrics. You can use any feature you want (bag of words, word count, tf-idf, word embeddings obtained by a neural network, etc.). You can also combine different features together. Do whatever you think can represent the dataset in the best way possible.

2. Build a neural network that is able to perform text classification. The inputs of the networks will be the features extracted from the songs and the output the most probable genre the song belongs to.

Part 2: Transfer Learning using BERT

1. For the second part of the homework you will download, using PyTorch facilities, a pretrained and very recent neural network developed by Google, called BERT, that is able to perform a very powerful text embedding taking into consideration information such as semantics, syntax, morphology and so on. So, you will do a bit of transfer learning. You will feed the network with your data and you will select the proper output layer for the task. I suggest you check at: `https://github.com/nlptown/nlp-notebooks/blob/master/Text%20classification%20with%20BERT%20in%20PyTorch.ipynb` to see how to use BERT.
2. Compare the results obtained using the method developed in Ex 1.1 and BERT. Can you do better? ;)

Write a SHORT report (max 4/5 pages) in which you describe all the steps (plots are welcome). Describe the features used and why you thought they could represent your data properly. Describe the neural net, the layers, the loss function and the optimizer used (not deeply) and why you think your network can model the problem properly. Comment the results obtained with any observation you think to be important.

Hint: The dataset is highly unbalanced, with some classes appearing more frequently than others. Thus, try to balance the dataset if your results are not satisfying or even remove those classes with too few samples (but please do not do just a POP vs ROCK classification). If you think this operation to be necessary, write it in the report and compare the results obtained before and after balancing the dataset.