Intro to DL Y2022 Semester B

Project part 1

Submission

Submission is in **pairs or singles**.

Data format and description

The data appear in moodle in three files:

- training_ex1_dl2022b.csv
- validation_ex1_dl2022b.csv
- test_ex1_dl2022b.csv

The training and validation data contain rows with ID, text, and labels (0,1,2 or 3):

id	text	**label**
1	blah blah blah	3
…	…	…

The test data contains row ID, 12 attributes (valued 0 or 1), and no label:

id	text
1	2
…	…

The task

Your goal is to **predict the label** for test data values, based on the text (do not use the ID column! If you do, your score will be 0, and you will be on the bottom of the kaggle table).

Represent the text using either a tf-idf or n-gram text models of the sklearn. The choice of a model is up to you, and the choice of n as well. You can use character or word n-grams.

Then use any of the suitable traditional ML methods in sklearn.ensemble to train your classifier on the training data and to evaluate it on the validation data. Your job is to get accuracy as high as possible for the validation set.

Then run your trained classifier on the test set and save the label you produce for every item in .csv file, as described below.

Note: you can use any of the normalization and data analysis methods in sklearn to improve your scores.

Result submission

Your result should include item IDs from the test set and predicted label, and to be saved as csv file:

id	label
1	1
2	2
…	…

How kaggle works

Your results will be compared with the actual test dataset labels, and the resulting accuracy will be reported on the scoreboard of the competition. Note that public scoreboard will show accuracy on 50% of the test set, and private (i.e., my) scoreboard will show accuracy on the whole test set. The final scoreboard will be published after submission & code checking is over, and your grade will be determined by your place in the competition.

Code submission

Submit your code and your results on moodle, as a single <id1>_<id2>.py file (do not submit python notebooks!).

Note of warning: all code will be automatically checked for copying. If cheating is discovered, you will get grade 0 automatically and go on to face the scholarly committee.