2021/2022 UNIVERSITÉ TOULOUSE I CAPITOLE

Mathematics of Deep Learning - Homework 2

November/December 2021

Team of 2 are allowed

To produce a good homework, you are asked to upload a Python notebook and a pdf document as well.

- The Python notebook should be illustrated with graphical outputs, and I strongly encourage the use of colab.
- The pdf document should describe the principle of each algorithm you are using to solve some datascience issues. For example, if you are solving a prediction task with the Lasso in a part of your notebook, you should explain in the pdf file:
  - What kind of energy is involved by the Lasso and why?
  - What kind of optimization algorithm is used and if it works rapidly or not (in terms of numerical complexity).
  - How the parameters are tuned.

Of course, you may try to use some algorithms that have not been introduced during the course. You are encourage to provide the best description you can even for an algorithm that is discovered in www.

1 Dataset

Coronavirus Tweets: available at `kaggle.com/datatattle/covid-19-nlp-text-classification`.

This dataset contains tweets from 2020, related to Coronavirus, with date and location. Each tweet was manually labelled regarding the sentiment expressed, hence this dataset is fitted for sentiment analysis but also for classification and topic identification.

I invite you to address a personnal study on each of the following problems. You are free to use any algorithm and any ressource you want.

2 Document content exploration

A basic problem consists in exploring directly the content of any document by collecting the terms and sequences of terms (n-grams) and performing some counting and clever proportions.

You are asked to

- Download the dataset (train and test)
- Represent the "sentiment" distribution over the dataset, and its evolution with time.
- Produce some quantitative elements that describe the dataset
- Clean the dataset with nltk
- Produce a cloud of words
- Shows the most frequent words used in tweets according to the sentiment expressed
- Obtain a final tractable dataset with tf-idf on the cleaned dataset.

Keywords: TF-IDF, Seaborn, Panda, nltk, ...

3 Clustering and topic identification

Given a collection of documents, one might want to group them in separate clusters and analyze the clusters, for example looking at the words used in each clusters.

Given a collection of articles, one might want to discover underlying topics reached in the different articles. This is a way to explore more deeply the contents of different documents in a corpus. Note that this is slightly different than clustering, as here a given document reaches several topics with a given percentage, rather than belonging to only one. Often, when we look at the terms used depending on the topics, the interpretability is relatively easy.

- Cluster the corpus
- Identify some profiles of texts and interpret them.
- Explain an analogy between NMF and soft-clustering (aka mixture models). Below, you will use the parametrization obtained after this NMF to produce some inference algorithms.
- Represent the sentiment according to the clusters/profiles

Keywords: Feature engineering, Kmeans, Silhouette, Latent Dirichlet Allocation, Non-negative Matrix Factorization

4 Sentiment analysis and inference

- Using the training and test set, you will have to *try several algorithms for predicting the sentiment.*
- *You can either choose to use classification or regression.* You will have to explain your choice!
- *Several metrics are possible:* the one of the classification, the one of the regression, an ad-hoc metric that accounts for some balanced or relative differences.

5 Generate text sequence (optional)

Generating text is a difficult but very interesting problem in natural language processing. It can help us create content such as film scripts, reformulate text sentences (without losing the meaning), write nicer emails, suggest text corrections and future words for efficient text processing.
Keywords: Markov model, Neural Network with memory system

Additional ressources

Python packages: NLTK, TextBlob, spaCy, ...

- Start kit on NLTK: `textminingonline.com/dive-into-nltk-part-i-getting-started-with-nltk`
- Starter kit on spaCy: `nicschrading.com/project/Intro-to-NLP-with-spaCy`