PROF. DR. SEBASTIAN SAGER, BENJAMIN PETERS
INSTITUT FÜR OPTIMIERUNG
OTTO-VON-GUERICKE–UNIVERSITÄT MAGDEBURG

November 19, 2020

Webpage for the lecture: `https://mathopt.de/TEACHING/2020OMML/`

Optimization Methods for Machine Learning

WS 2020 – 4. exercise sheet

Exercise 4.1 (Text classification and kernel trick)
Goals: *Use different text preprocessing options and write your own kernel.*

1. Get the exercise template `ex04_temp.py` from our webpage `https://mathopt.de/TEACHING/2020OMML/` and go through the provided lines.
2. Complete the function `text_prep`:
   - Required arguments are the text-data and the number of consecutive words that should be used to build features. (Have a look at the `ngram_range` parameter of `CountVectorizer`.)
   - There should be optional arguments for usage of frequency and inverse document frequency transform that should be used by default.
   - The function should build a dictionary of features and transform the text documents into feature vectors. (Hints: Use `.fit_transform` from `CountVectorizer` and `TfidfTransformer`.)
   - Return the feature vectors.
3. Analyze the `example_text` with your function and have a look at different optional parameter settings.
4. Use the `20newsgroups` dataset to fit a *support vector machine*.
   - You can split the data into train and test set and see how the svm performs.
   - The Scikit Learn Algorithm Cheat Sheet suggests a *Naive Bayes* classifier. Import the `MultinomialNB` classifier from `sklearn.naive_bayes` and use this to predict the labels of your test set.
5. Test your results with sentences from the web or yourself. (You can relearn your classifiers on the complete data set.)
6. Write your own hyperbolic tangent `kernel`:
   - Return $k(x, x') = \tanh(< x, x' >)$
   - Compare the results using your kernel and using the `sigmoid` kernel provided in sklearn.