Lab Course Machine Learning
Exercise Sheet 11

Prof. Dr. Dr. Lars Schmidt-Thieme, Shayan Jawed
Information Systems and Machine Learning Lab
University of Hildesheim

January 21st, 2021
Submission on January 27th, 2021 at 12 noon, (on learnweb, course code 3116)

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. You should submit a jupyter notebook detailing your solution.
2. Please set the seed(s) to 3116.
3. Please explain your approach i.e. how you solved a given problem and present your results in form of graphs and tables.
4. Please submit your jupyter notebook to learnweb before the deadline. Please refrain from emailing the solutions except in case of emergencies.
5. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
6. Please refrain from plagiarism.

Exercise 0: Preprocessing Text Data (5 Points)

In this exercise, you are tasked with implementing a Text Classifiers to categorize news items. The dataset name and link:

- 20newsgroups dataset (A collection of 20,000 news items across 20 categories)
- Available via Scikit-Learn Datasets API.
- Subset the dataset to only the following two categories named as 'sci.med' and 'comp.graphics'

The preprocessing tasks are as follows:

1) Preprocessing textual data to remove punctuation, stop-words (list available via external libraries such as NLTK and spaCy).
2) Implementing a bag-of-words feature representation for each text sample
3) Implementing a TF-IDF feature representation for each text sample
4) Split the dataset randomly into train/validation/test splits according to ratios 80%:10%:10%

Please refer to the following resource for explanation regarding the above two preprocessing schemes: `https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html#training-a-classifier`

Exercise 1: Implementing Naive Bayes Classifier for Text Data (5 Points)

In this exercise, you are tasked with implementing a Naive Bayes Classifier to categorize news items. The Naive Bayes assumption is conditional independence of the features when modeling for the label.

1) For both preprocessing types train and validate the Naive Bayes Classifier to classify each news item to the two categories named above.

2) Report the test set accuracy.

Exercise 2: Implementing SVM Classifier via Scikit-Learn (5 Points)

Please refer to the Scikit-Learn library for implementation of SVM classifiers. You are required to replace the classifier in Exercise 1 with SVMs.

1) Tune the different SVM kernel choices provided by Scikit-Learn, and other associated hyperparameters for validation set.

2) Report the test-set accuracy.

Exercise 3: Hands-On Tutorial for Submitting jobs on ISMLL Cluster (5 Points)

Please note that the last 5 points for this exercise sheet are reserved for an in-class hands-on tutorial where we shall be looking how complex and compute-intensive tasks can be delegated to the ISMLL cluster. In order to qualify for these 5 points, your attendance is thus mandatory in the next week's tutorial.