COMP10200: Assignment 3 – Part 2

Overview

For this assignment, you will obtain some machine learning data for classification, test various SKLearn implementations of Naïve Bayes, and report your findings.

This is Part 2, in which you will obtain a corpus[1] for text classification and apply the Multinomial Naïve Bayes classification algorithm to it using various representations of the text.

The Data

The UCI Machine Learning Repository has 26 data sets marked as "text" and "classification".

https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=&area=&numAtt=&numIns=&type=text&sort=nameUp&view=table.

There are data sets here for spam filtering, sentiment analysis, sentence classification, etc. You will have to write your own code to read and convert them into a usable form.

If you want to use the Reuters-21578 corpus, there is a zip file on Canvas with a cleaned-up version of the corpus, some helper code for reading it, and a handout explaining it. Feel free to use this code.

Whatever data set you use, choose at least 5 binary classification tasks from it (i.e. pick five labels and use them to create five binary classification tasks – in each label categorize the texts as 1 or 0, meaning the text gets the label or does not get the label).

The Code

The code you use for this assignment should be written in Python using Numpy and SKLearn. It is expected that you will have to write some of this code yourself, but you are not expected to write everything from scratch. Feel free to adapt the code from Canvas or other sources to suit your needs. You could also explore other Python packages for natural language processing, such as **nltk** (Natural Language Toolkit).

The most important requirement for the code you hand in is that it be correctly sourced and documented. You must make it clear where you got the original code from and what modifications, if any, you made to adapt it to your needs.

The Task

Your task is to test the Naïve Bayes classification algorithm against at least 8 different representations of text. Create these 8 different versions by choosing 2 values for each of 3 parameters (words vs. stems, bag vs. tf-idf, Multinomial vs. Complement NB, word removal vs. no word removal, etc.) and then trying all the permutations. At least one of the 3 parameters you vary should require some manipulation of the list of words (e.g. stemming, word removal, 2-grams, etc.). Create a standard training/testing split and

[1] A corpus is a large collection of natural language text. The plural is **corpora**.

use the same split for every run. For each run, you should report a combined confusion matrix that summarizes all 5 tasks together, then compute the **micro-averaged** accuracy, precision, and recall (micro-averaged just means it's created from the combined confusion matrix). Here are some ideas for different text representations:

- Bag of words (using Multinomial NB and/or Complement NB)
- Set of words (using Bernoulli NB)
- Bag or Set of Stems
- Bag or Set of words with some removed (stop words, frequent words, infrequent words, etc.)
- Bag or set of words and N-Grams (e.g. add a selection of 2-word phrases to the feature set)
- The tf-idf representation (using Multinomial NB and/or Gaussian NB)

Whatever text representations you choose, make sure that you use Multinomial NB for word counts and Bernoulli NB for binary data.

The Report

You should write a short report, using word processing software, that contains the following sections:

1. **Data Description** (data set name, source, description of classification task, and a description of how you generated the training and test set split, plus some statistics – number of features on each run, number of items in each class, number of items in training and test set, etc.)
2. **Description of the Text Representations** (describe the 5 representations of text you experimented with, and explain how you created them)
3. **Results** (confusion matrix, accuracy, precision, recall, and for all 5 runs)
4. **Discussion** (are there clear winners or losers? Give some solid ideas for why some text representations might be better or not better than others. Do the results make sense to you? Why or why not? Be as specific as you can.)
5. **Future Work** (If you had more time, where would you go next? What other variations of text representation would you like to explore? What other algorithms or data sets would you like to use? What other tests would you like to do? Etc.)

Throughout your report, make sure you are using standard grammar and spelling, and make sure you make proper use of correct machine learning technology wherever appropriate. If you quote or reference any information about Naïve Bayes or issues in Text Classification that were not explicitly covered in class, you should cite the source for that information using correct APA format.

Handing In

Zip up your report, your data files (if not using Reuters-21578), and the code for both parts of this assignment and hand them in together. It should be possible for me to unzip your folder and run your code easily (i.e. without having to move files around or make changes to the code) to see the same results you are reporting in your report. If necessary, include instructions at the top of each code file with the correct command to run the code. See the drop box for the exact due date.

Evaluation

This assignment will be evaluated based on: 1. the quality of the report you produce; 2. how well you met the requirements of the assignment; and 3. the quality of the code you handed in (including quality of documentation and referencing within the code).

See the rubric in the drop box for more information.