Assignment 1

CS-541: Artificial Intelligence

Spring 2022

In this assignment you will implement a logistic regression model and a K-Means clustering model and apply them on two real-world datasets.

In order to get full credit, you should complete both the problems without using any python scientific packages including `sklearn`, `scipy` etc. You can only use `pandas` or `numpy` for reading the data or for matrix operations.

1 Abusive Swearing Detection

Details The dataset [2] is a set of annotated tweets based on whether the tweet contains abusive swearing or not.

You are given two input files; *swad_train.csv* & *swad_test.csv*. You are also given two additional files which will be used during pre-processing; *punctuations.txt* & *stopwords.txt*.

Part 1. [5 pts] Textual data needs to be pre-processed before it can be used for feature extraction. Carry out the following pre-processing steps:

1. **Lowercasing:** Convert all the tweets into lowercase
2. **Tokenization:** Add space before and after punctuation present in *punctuations.txt* **Use tfidf**
3. **Stopword removal:** Remove all the stopwords present in *stopwords.txt*

Part 2. [7.5 pts] Extract the features from the dataset. One idea is to use tfidf features. You can also use any other sets of features as well. Details of tfidf are mentioned in Appendix A.

Part 3. [10pts] Implement the logistic regression model. You will need to implement the loss function as well as a parameter optimzation algorithm, e.g. gradient descent or stochastic gradient descent.
Note: In order to complete the question, you can use a python package, e.g., `sklearn`. This will ensure that you at least get graded for other parts of the question. However in that case, you will **NOT** get points for this part.

Part 4. [**2.5pts**] Evaluate the model on the test data. You should use accuracy score as your evaluation metric.

2 Clustering demographics

Details This partial dataset [1] contains census information including, age, workclass, education, marital status etc. The original task is to classify the people into salary brackets.

You are provided with a single file (*income.csv*). The task is to group the data into optimum number of groups.

Part 1. [**5pts**] The dataset provided is not very clean. Therefore you have to perform a number of pre-processing steps before you can input the dataset to the machine learning model.

1. **Missing features:** You should replace every such feature with other values, e.g., minimum, maximum, or average of that feature.
2. **Normalization:** Make the range between values similar. Larger differences can make model convergence difficult.
3. **Categorical features:** Replace them with numerical values. e.g. male/female can be replaced with 0/1 and so on.

Part 2. [**10pts**] Implement the K-Means clustering algorithm to cluster the data into K cluster.
Note: In order to complete the question, you can use a python package, e.g., `sklearn`. This will ensure that you at least get graded for other parts of the question. However in that case, you will **NOT** get points for this part.

Part 3. [**5pts**] You should optimize the value of K using elbow method.

Part 4. [**5pts**] Create visualizations of the distortions computed by the elbow method.

Submission

This is an individual assignment. Each person should submit as a single zip file named with assignment number and the username (e.g. *HW1_akhan4.zip*). The zip file should contain the required code file and a readme file. The readme should include the following:

- one line descriptions of the code file
- final test accuracy score (for Q1)
- elbow method plot (for Q2)
- optimal K value (for Q2)

Remember that after general discussions with others, you are required to work out the problems by yourself. All submitted work must be your own, though you can get help with others, so long as you cite the help. Please refer to the Stevens Honor System for clarifications.

References

[1] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[2] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.765.

A Appendix

Term Frequency-Inverse Document Frequency (tf-idf) is one idea to convert text into a set of features. It is commonly used in information retrieval. It is intended to reflect how important a word is to a document in a collection or corpus. The tf–idf is the product of two statistics, term frequency and inverse document frequency.

Term Frequency:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Inverse Document Frequency:

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where,
D is the total documents
$N = |D|$ is the total number of documents
$f_{t,d}$ is the raw count of a term in a document

Text

Dont remove missing val

Label encode the values- countries 1—9