Krannert School of Management

Individual Assignment 3

Total Points: 10

Part 1. Text Classification (total 5 points)

There are 1000 reviews for restaurants and films in a collection under the attached csv file on Brightspace. All those reviews are labeled with its category (either restaurant review or movie review). You are required to develop classifiers that could automatically determine whether a future text body is a restaurant review or a movie review. Please follow the steps listed below:

1. Build a training dataset with the first 400 restaurant reviews and the first 400 movie reviews (ID = [1:400, 501:900]), and use the rest as a test dataset.

2. Following step 1, transform those reviews into a TFIDF matrix, lemmatize all the words, remove the stop-words and punctuations, set the minimal document frequency for each term to be 5 and include 2-gram.

3. Following step 2, train a Naïve Bayes model, a Logit model, a Random Forest model, (with 50 trees), a SVM model, and a simple ANN models with 4 neurons in one single layer. Calculate the accuracy rate for each of the models, report them and discuss which model performs the best according to the accuracy rate.

4. Following step 1, represent each raw document using index-encoding with maximum length of 100 padding.

5. Following step 4, train a deep learning model with the sequential architecture of: 1 embedding layer with 20-element length for the output; 1 LSTM layer with 40 units, dropout of 0.2, and recurrent dropout of 0.2; one dropout layer of 0.1 rate; and 1 dense layer for the output. Setting batch size to be 100 and do 10 epochs. Report its accuracy rate and discuss whether this model performs better than models in step 3 and why.

Please submit 1 file:

A word file includes python code with your comments #, and one screenshot on your PyCharm showing that your code has run through successfully for **each of the steps** (5 screenshots in total). Also, **report** the accuracy rates and your discussions for question 3 & 5 at the end of the word file.

Part 1. Image Recognition (total 5 points)

CIFAR10 dataset in keras provides of 50,000 32x32 color training images and 10,000 test images, labeled 10 classes. You are required to develop classifiers that could automatically recognize the object(class) in the images. Please follow the steps listed below:

1. Load all data using from keras.datasets import cifar10, splitting the dataset into training set (50,000 documents) and test set (10,000 documents). Visualize the first 20 images of test set and compare them with their labels. Discuss what objects/labels are in this dataset.

2. Following step 1, train a CNN model using the sequential architecture:
    a. 1 conv2d layer with 32 3 by 3 filters and "ReLU" activation function
    b. 1 dropout layer of 0.2 rate
    c. 1 conv2d layer with 32 3 by 3 filters and "ReLU" activation function
    d. 1 maxpooling layer with a size of 2 by 2
    e. 1 conv2d layer with 64 3 by 3 filters and "ReLU" activation function
    f. 1 dropout layer of 0.2 rate
    g. 1 conv2d layer with 64 3 by 3 filters and "ReLU" activation function
    h. 1 maxpooling layer with a size of 2 by 2
    i. 1 flatten layer
    j. 1 fully connected layer with 256 neurons and "ReLU" activation function
    k. 1 dropout layer of 0.2 rate
    l. 1 fully connected layer for the output using softmax activation function.

    Setting batch size to be smaller than 500 to prevent overheating and do 5 epochs. Report the accuracy.

3. Following step 2, between layer h and i, add three more layers as the following:
    a. 1 conv2d layer with 128 3 by 3 filters and "ReLU" activation function
    b. 1 dropout layer of 0.2 rate
    c. 1 conv2d layer with 128 3 by 3 filters and "ReLU" activation function

    Train the new model with 5 epochs. Report the accuracy and compare it with the result of step 2. Discuss the difference.

4. Using the model in step 3 and change the epoch to 20. Report the accuracy and compare it with the results of step 3. Discuss the difference.

5. Flatten the representation of each image as a vector, train a Naïve Bayes model and a Random Forest model (with 100 trees and max depth of 10). Report the accuracy and compare it with the results of steps 2, 3 and 4. Discuss which model performs the best according to the accuracy rate.

Please submit 1 file:

A word file includes python code with your comments #, and one screenshot on your PyCharm showing that your code has run through successfully for **each of the steps** (5 screenshots in total). Also, **report** the accuracy rates and your discussions for question 2, 3, 4 & 5 at the end of the word file.