Νευρο-Ασαφής Υπολογιστική

Χειμερινό Εξάμηνο 2018-2019

Δημήτριος Κατσαρός

Coding project

The original format of our data was in the form of a text file and had the following attributes:
#* --- paperTitle

#@ --- Authors

#t ---- Year

#c --- publication venue

#index 00---- index id of this paper

#% ---- the id of references of this paper (there are multiple lines, with each indicating a reference)

#! --- Abstract

<u>Processing of DataSet v1:</u>

1) Convert the .txt file to .csv with all the information given.

    a. We created a DataFrame through which we removed the title, authors, publication venue, abstract.

    b. We deleted records that were not dated.

    c. We converted the dataset into a table for better and faster processing.

2) Format of the new table: [Ids] lines and columns [years] up to 600,000 entries.

3) Each position is increased by 1 each time a paper refers to another paper in a particular year.

4) With this amendment we will use our data on the neural network.

5) Looking at the data, we see a rather sparse table so we do extra filtering to reduce it even further.

a. We deleted the first 50 years since they had little or no references at all.
b. We deleted the rows that were referenced less than 90 times.

Ending up with 1865 records to train in our neural network.

Choosing a neural net:

```
model1 = Sequential()
model1.add(LSTM( 5, activation = 'relu', input_shape=(5, 1), return_sequences = True))
model1.add(LSTM( 5, activation = 'relu', return_sequences = False))
model1.add(Dropout(0.3))
model1.add(Dense(1))
model1.compile(optimizer=”Adam”  metrics = ['accuracy'],loss='mse')
```

1) **Long short-term memory** (**LSTM**) is an artificial recurrent neural network (RNN) architecture[1] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data.

2) In a neural network, the activation function is responsible for transforming the summed weighted input from the node into the activation of the node or output for that input.

We chose **ReLu**. The rectified linear activation function is a piecewise linear function that will output the input directly if is positive, otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often achieves better performance.

3) A single model can be used to simulate having a large number of different network architectures by randomly dropping out nodes during training. This is called **dropout** and offers a very computationally cheap and remarkably effective regularization method to reduce overfitting and improve generalization error in deep neural networks of all kinds.

4) Adam is an optimization algorithm that can used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data.

Variable Selection:

We observed that more neurons didn’t make an improvement in our results.

Also, we used a dropout of size 0.3 in order to avoid overfitting.

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges. We decided to normalize out Data, in order to have a more concentrated Dataset.

Training error = 2.5

Testing error = 7

With our final Dataset (1860 records) our neural network took about 25 minutes to train the model.