MM-811 Multimedia Reading Topic

Report

Deep Learning

Prabhjot Singh Mansa
3-6-2016

Interpretation of Dataset: Small Slice

1. Description of the Dataset

The dataset used is 'Yeast'. In this data set, there are nine fields which has eight deciding fields as input to calculate whether yeast is nuclear or non-nuclear. The ninth field in the dataset is deciding factor of yeast as there are two values 0 and 1 only.

1_value	2_Value	3_value	4_value	5_value	6_factor	7_value	8_Value	9_Value
Sequence	MCG	gvh	Alm	mit	erl	pox	vac	Nuc

Note:

1. Header in above table is not completely identical to dataset naming convention.

2. Description of Problem

The problem is to find out the yeast whether it has nuclear. To achieve the goal there are eight variables used. To analysis nuclear localization signals of nuclear and non-nuclear.

3. Condition for input

There is one condition to check the nuclear which is obtained by 'AND' operation of two layers in iterative way.

Parameters:

inputs = data[0:,0:8].astype(np.float32) //Input range from the dataset
outputs = data[0:,8:9] .astype(np.int32) //output range defined

Interpretation of different Architecture

4. Architecture Details

I have used three architectures for the Deep Learning.

S.NO	Input	Hidden	Layer
1.	8	4	2
2.	8	7	3
3.	8	2	2

5. Condition for Input

Yeast dataset has been used which has total nine fields. The input information has eight fields which provides the information regarding the following in table. The ninth field provides the information regarding the information of neutron of yeast.

Declaration of Input: inputs = data [0:,0:8].astype(np.floats32).

Defines the range for input which is used as range of field which it will get from dataset

1. Sequence Name: Accession number for the SWISS-PROT database
2. mcg: McGeoch's method for signal sequence recognition.
3. gvh: von Heijne's method for signal sequence recognition.
4. alm: Score of the ALOM membrane spanning region prediction program.
5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.
6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.
7. pox: Peroxisomal targeting signal in the C-terminus.
8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.
9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

6. Description to interpret the Output

The performance of the output actually depends upon the hidden layer if it is less then there is high possibility of achieving good result. The output will may reference between two classes of neutron and non-neutron.

Outputs = data[0:3:4].astype(np.int32)

7. Description of Performance

S.NO	Architecture	True Positive	True Negative	False Positive	False Negative
1.	(8,4,8,2)	24	07	06	00
2.	(8,6,5,8,2)	24	7	3	2
3.	(8,2,2)	21	0	16	0

8. Discussion to improve the result

The result can improve from different techniques like K nearest neighbor. This can be calculated from Euclid distance. Another way is to try regression technique which can be explored to improve the result.

9. Summary

1. First dataset is loaded from the data file which is divided for testing and training in 9:1.
2. Then, data is divided into two parts for validation and training.
3. An array is used to store the data in form of true and false.
4. A classifier is used define the architectures.
5. To calculate K-nearest mean if required. Then, Euclid distance is used (if using classifier slice).
6. Finally, trained data is used to check the classifier.
   Note: In hidden layer, iteration uses the result of pervious for learning.

10. References

UCI Machine Learning Repository: Yeast Data Set