Predicting Earning of an individual using Neural Network

Rushil Agarwal
Australian National University

Abstract: This paper aims to predict whether a person earns more than $50K per annum on the basis of various characteristics using Neural Network. Further it aims to improve the efficiency and of the model by improving on the input encoding and also applies network pruning techniques particularly via measuring distinctiveness between neurons.

Keywords: Neural Network, network pruning via distinctiveness, input encoding techniques.

1. Introduction

There are various techniques to encode the input data before feeding into the network, which depends on the type of data (to name a few - nominal categories, ordinal categories, continuous); and depending on the intra-relationship between the categories of the concerned feature, the structure of input neurons are formed in accordance. Further to increase the efficiency of the model in terms of computation, network pruning is required without a sacrifice in the models accuracy.

This paper uses the data from http://archive.ics.uci.edu under the heading 'adult.data' which contains anonymized information of individuals focused to predict their salary bracket between a division of under $50K or over it.

2. Dataset Information

The original data was extracted from http://www.census.gov from the 1994 census database and was split into training and test set using MLC++ GenCVFiles with a ratio of 2:1 for train-set:test-set. There are a total of 48842 instances with unknown values in some data instance. 45222 of those instances hold complete information and do not contain any unknowns. Therefore ~93.35% of the instance in the dataset are completely organic.

The dataset is discretized over the gross income of an individual into two ranges with threshold 50,000. Without removing the unknowns, the probability for the label '>50K' is 23.93%, hence the data contains more training examples for the label '<=50K'.

Information of other models used as reported that were built on this dataset:
C4.5 : 84.46+-0.30
Naive-Bayes: 83.88+-0.30
NBTree : 85.90+-0.28
The above stated models used instances without the unknowns.

Attribute Information

1. ‘age’
   Type: Continuous.
2. ‘workclass’
   Type: Nominal categories
   Categories: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. ‘fnlwgt’
   Type: Continuous
   Info: Also referred to as final weight, it is created by the census organization and in essence reflects socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights.
4. ‘education’
   Type: Ordinal categories
   Categories: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. ‘education-num’
   Type: Continuous
   Info: Represents the information in ‘education’ attribute in a integer based numerical form.
6. ‘marital-status’
   Type: Nominal categories
   Categories:Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. ‘occupation’
   Type: Nominal categories
   Categories: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. ‘relationship’
   Type: Nominal categories
   Categories: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. ‘race’
   Type: Nominal categories
   Categories: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10. ‘sex’
    Type: Nominal categories (Binary)

Categories: Female, Male.

11. ‘capital-gain’
    Type: Continuous
12. ‘capital-loss’
    Type: Continuous
13. ‘hours-per-week’
    Type: Continuous
    Categories: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
14. ‘native-country’
    Type: Nominal categories
    Categories: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad & Tobago, Peru, Hong, Holand-Netherlands

3. Pre-processing and Input Encoding

After extracting and cleaning the data, the unknowns which were marked by ‘?’ were replaced by the most frequently occurred (mode of data attribute) value in the concerned column.

The dataset contains 4 types of attributes and for each attribute, following measures have been implemented:

1. Continuous values - Normalized to between 0 and 1.
2. Nominal Categories - Distributed over multiple input neurons, where each input neurons represents a value of the category.
3. Ordinal Categories - There is only one such category ‘education’, which was removed as ‘education-num’ represents the same value as the structuring of ‘education’ as an ordinal category would represent.
4. Nominal categories (Binary) - Converted to integer type input by being either 0 or 1.

4. Neural network Information

- Total input neurons after input encoding - 104
- Output neurons - 2
- Learning rate - 0.001
- Number of Epochs - 300
- Optimizer - RMSprop
- Loss Function - Cross Entropy Loss
- Hidden neurons - 100

5. Original network Results

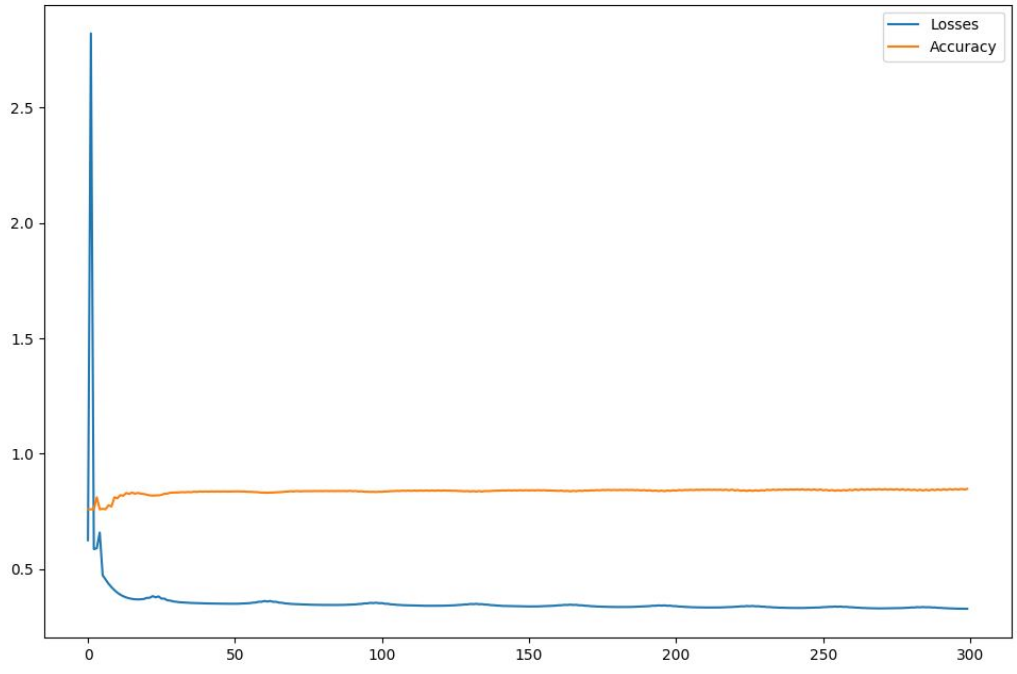
Confusion matrix for training:

3931 3910
1154 23566

Testing Accuracy: 84.45 %

Confusion matrix for testing:

1982 1971
576 11752



6. Network Pruning

In order to increase the computational efficiency of the model without sacrificing the accuracy of the same, pruning methodology of calculating distinctiveness between hidden neurons was used.

Calculating Distinctiveness

The distinctiveness between neurons is calculated by measuring the directional distance or the angle between 2 neurons. To mathematically calculate this, a vector is formed with the weights of the hidden neuron. For each such neuron vector, the angel is calculated by the following formula -

$$\cos \alpha = \frac{a \cdot b}{|a| \cdot |b|}$$

Where ‘α’ is the desired angle between the 2 vectors, If the angle is less than 15 degrees, the neurons are merged by simply adding the weights. The accuracy of the pruned model relative to the original model can be increased by reducing this angle with a trade-off of decreased efficiency as less number of neurons will be merged.

7. Pruned network Results

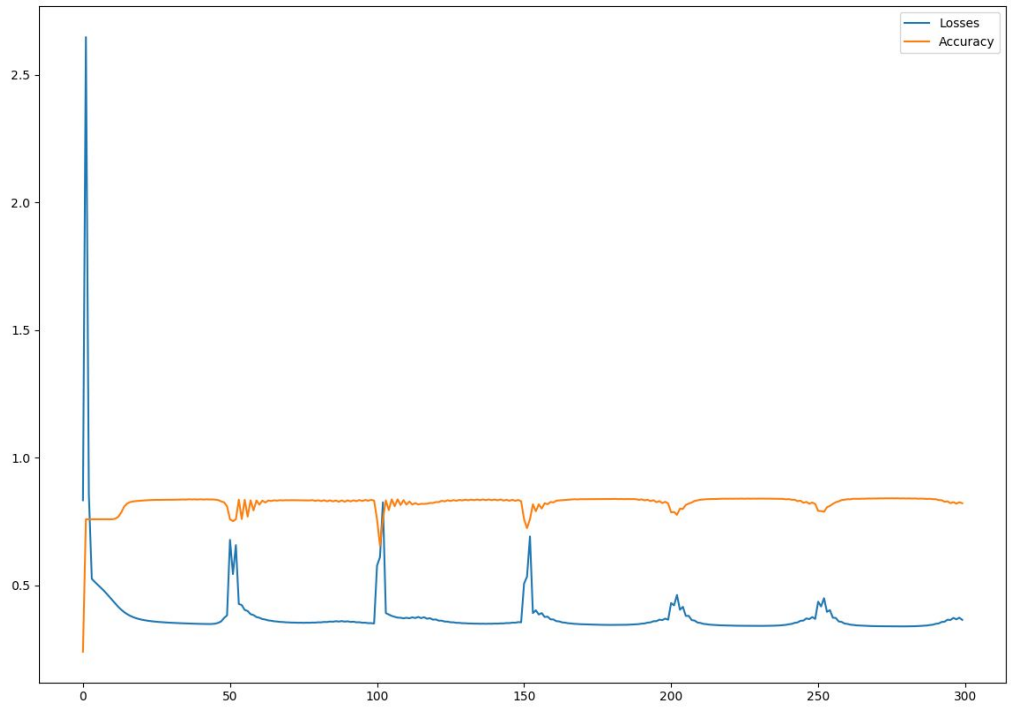
Confusion matrix for training:

1748	6093
231	24489

Testing Accuracy: 80.58 %

Confusion matrix for testing:

875	3034
109	12263



8. Future Work

The column ‘fnlwgt’ represents the socio-economic characteristics of a population with the essence that People with similar demographic characteristics will havel similar weights. Although the sample is actually a collection of 51 state samples, each with its ownl probability of selection, and the procedure of calculating this weight makes the above statement only applicable within a state. Therefore, additional information of state of an individual or dropping the stated factor will help in the accuracy of the model. Since 2 instances from different will represent a different demographic even though they have a similar value.

The effect and extent of pruning can be further examined by using various threshold of angles while keeping the test and train data same for all such models and creating multiple pruned networks and test it against the test data.

9. References

Archive.ics.uci.edu. (2018). *UCI Machine Learning Repository: Adult Data Set*. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Adult [Accessed 1 Jun. 2018]
Jmlr.org. (2018). [online] Available at: http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf [Accessed 1 Jun. 2018].
Zhu, H. (2018). *Predicting Earning Potential using the Adult Dataset*. [online] Rstudio-pubs-static.s3.amazonaws.com. Available at: https://rstudio-pubs-static.s3.amazonaws.com/235617_51e06fa6c43b47d1b6daca2523b2f9e4.html [Accessed 1 Jun. 2018].