A Study on the Differences of Dimension Reduction Methods Combined with Different Neural Network Activation Functions.

By Diep Nguyen and Roice Johnson, University of North Carolina at Charlotte

2018

Abstract

This research composed of experiments with three specific dimension-reduction methods: principal component analysis, factor analysis, and linear discriminant analysis. After dimension reduction, neural networks used the new datasets to create prediction models. A total of nine different neural networks were created as a result of combining the three dimension-reduction methods with three activation functions: tanh, relu, and sigmoid. Using the scikit learn library, the accuracy results were compared to find the best combination of dimension reduction , linear regression, neural network activation function to predict how well particular colleges prepare students for financial stability based on their median household income after graduation.

Introduction

The motivation behind this project was the need to understand the differences between three dimension-reduction methods and how they can affect the results of the neural network prediction score. We also wanted to see how the three new datasets would affect the prediction score in combination with the three popular choice of activation function learned in class. Using a very large dataset from the college scorecard which had more ninety-two columns, we reduced the size to two columns and chose a specific target column to train the model.

Data Set Description

This data was gathered and produced by the United States department on education as a public tool for deciding which college should be attended by prospects. Our target feature was the median household income graduates from said university earned prior to graduation.

Our features data used was numerical representations of college demographics and average and median quarterly earnings for a given instance.

Related Work

The Bureau of Labor Statistics also known as BLS, has many studies published and produced each year to monitor the trends of the economy as it pertains to education race and demographics. The following link is a comprehensive study completed by the BLS to report on the impact education and demographics play on the financial stability of individuals.

https://www.bls.gov/cps/earnings.htm#demographics

Dimension Reduction

To start the experiment, Roice used the three dimension reduction methods on the original dataset. After looking at the dataset, Roice picked a column as a target for the prediction. He then split the dataset into six parts for each dimension reduction method . Each method received a training and testing datasets. He then scaled the data so that the outliers did not overwhelmingly affect the training. The outliers were prevalent in the original dataset, which consisted of 7175 instances. Then Roice imported libraries from scikit learn for each dimension reduction method. Each method consisted of two lines of code to fit and transform the datasets. He tested the new datasets by using them to train linear regression models. The scores for each model using the test datasets.

Reduction Method	Prediction Score
Principal Component Analysis	0.4080340818290376
Factor Analysis	0.3964442468415504
Linear Discriminant Analysis	0.6347963064084712

Neural Network

For Diep's part, all she had to was to use scikit-learn's neural network library to train nine regression models. The regression models used the previous dimensionally reduced dataset from Roice's part. He created three reduced dataset. Each dataset was used three times in combination with the three of the activation functions mentioned previously. Below are the scores for each combination of dimension reduction method and activation function. All model used stochastic gradient descent and a max iteration of 1000.

Combination	Prediction Score
Principal Component Analysis and Sigmoid	0.3826017301670845
Factor Analysis and Sigmoid	0.0015637694633382493
Linear Discriminant Analysis and Sigmoid	0.7547775925129232
Principal Component Analysis and Tanh	0.5757658742720374
Factor Analysis and Tanh	0.392948179473934
Linear Discriminant Analysis and Tanh	0.7238271957639453
Principal Component Analysis and ReLU	0.5099716576135929
Factor Analysis and ReLU	0.4400805863850834
Linear Discriminant Analysis and ReLU	0.7754775300442343

Conclusion

Although we received some unexpected results with our models, Linear Discriminant Analysis constantly performed best for the given data set with dimension reduction. Dimension reduction increased the accuracy of our prediction for neural network model using only LDA and decreased for PCA and FA. There is too much data loss and not enough data separability to use dimension reduction for FA and PCA. Thus dimension reduction cause a sharp performance decline if accuracy is the priority then the two methods should not be used for this data set if size is the main priority then linear discriminant analysis with either neural net or linear regression is acceptable.

Model with original Dataset	**Score**
Linear Regression	0.9837408949617984
Neural Network with Tanh	0.5296111947642794
Neural Network with ReLU	0.6784948607565653
Neural Network with Sigmoid	0.681853868289383