Machine Learning Squared: Methods of Pre-Processing data using machine learning

Jakolbia Shipmon William Carey University Department of Mathematics

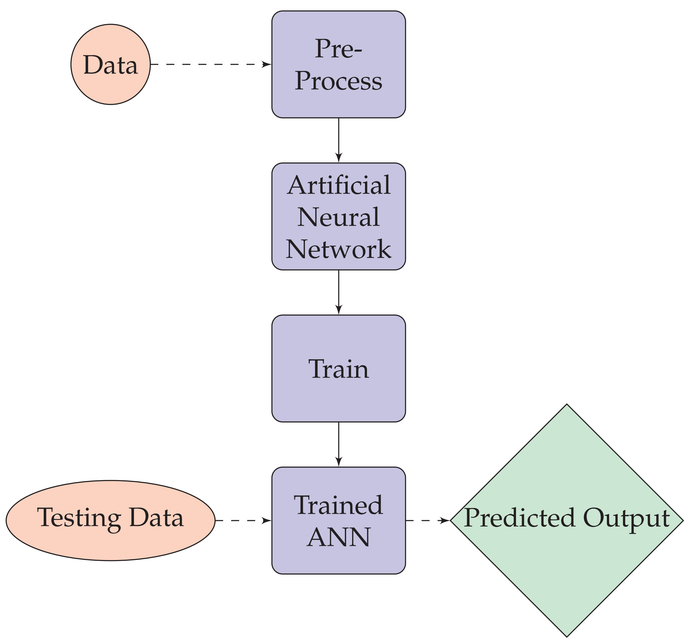
Pre-Processing Methods

Pre-processing datasets is a subset of research in the field of machine learning. One elementary method of pre-processing datasets that have missing entries is to average the values of the corresponding known entries and fill in the missing entries with that average. However, we will take a more advanced approach to this problem by using the neural network to recursively predict the value of each missing entry. One particular dataset that is well-posed for this type of pre-processing is the Pima Indians Diabetes dataset.

We used four different ways to pre-process data:

1. **No preprocessing** (not filling in any missing entries)
2. **Normalize and recenter** (put each data type on same scale)
3. **Fill with averaging** (fill in missing values with naïve averages)
4. **Fill with machine learning** (fill using neural networks)

After we have pre-processed this data in the different ways mentioned above, we design neural networks and train/test the neural networks to predict its accuracy. Machine learning is done by training and testing artificial neural networks (ANNs) as we described above. The following flow chart maps out how and when we train and test a network, and which data is used to do each part to produce the output.

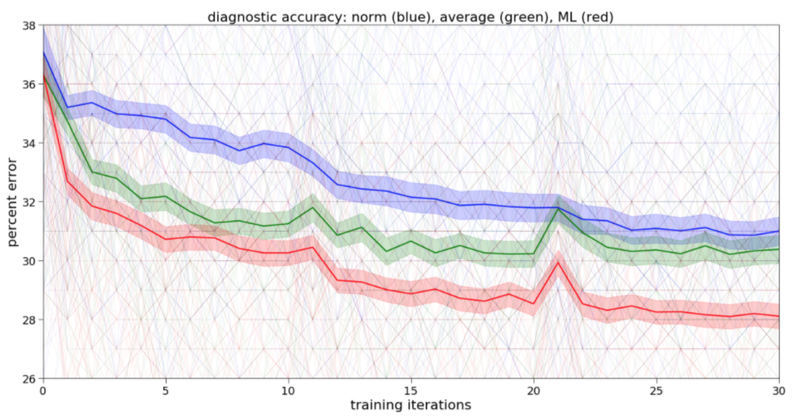


Example Dataset: Pima Indians

The Pima Indians Dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset is comprised of health characteristics:

1. Pregnancies
2. Glucose
3. Insulin
4. Body Mass Index
5. Skin Thickness
6. Age
7. Blood Pressure

that either play a role in or lead to diabetes. One problem when using a neural network to predict a particular outcome is that the data will most likely require some type of pre-processing. In the case of this dataset, we will have to normalize the data and fill in missing values. Columns 7 and 8 have no missing values; therefore, we begin to predict missing values in other columns by using these columns to train the network. In the Pima Indians Dataset, there are 652 missing values. Training a neural network that produces a high accuracy of prediction requires an accurate dataset. By pre-processing this dataset, the missing entries are determined. Once the data has been normalized and a model has been created, the data is utilized to train neural networks. By training neural networks, we can efficiently predict if patients have diabetes.



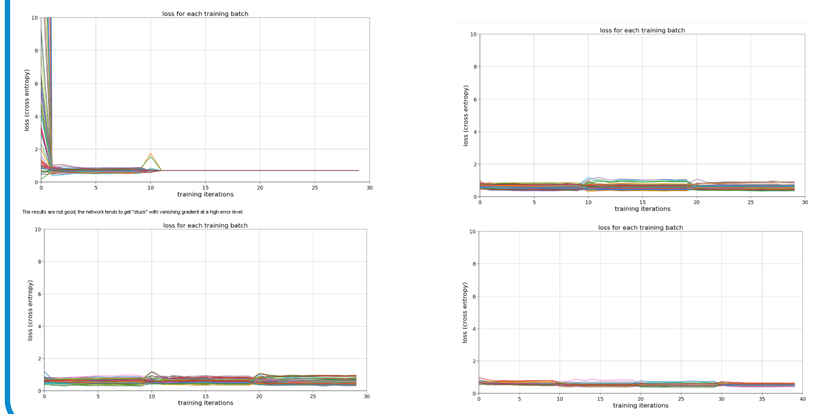
The above error plot shows three different types of pre processing:

1. The blue line shows results from the data being normalized.
2. The green line shows results from the averaging method discussed earlier.
3. The red line shows results from using machine learning to preprocess.

Testing Pre-Processed Data:

A network architecture was designed to run the pre-processed data (nothing, normalize, averaging, and machine learning) and test its accuracy. The following error charts show these results.

1. The top-left image shows training without pre-processing.
2. The top-right image shows training using data that is normalized and recentered.
3. The bottom-left image shows training where the missing data is filled in by the average values.
4. The bottom-right image shows training where machine learning is used to fill in the missing values.



References

[1] M. E. J. Newman. *Networks*. Oxford University Press, 2nd edition, 2018.

[2] Charu C. Aggarwal. *Neural Networks and Deep Learning*. Springer, 2018.

[3] Francois Cholletl. *Deep Learning With Python*. Manning Publications Co, 2018.

Contact Information

Web www.wmcarey.edu
Email jshipmon380103@student.wmcarey.edu
Phone (601)318-6172