Neural Network That Classifies Diabetes Risk with original data set

Maroš Stredanský
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach
Košice, Slovenská republika
maros.stredansky@student.tuke.sk

***Abstract*—Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems ”learn” to perform tasks by considering examples, generally without being programmed with task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as ”cat” or ”no cat” and using the results to identify cats in other images. They do this without any prior knowledge of cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the examples that they process.**

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron that receives a signal then processes it and can signal neurons connected to it.

In ANN implementations, the ”signal” at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology.

***Index Terms*—fingerprint recognition,artificial intelligence, artificial neural network, recognition**

I. Dataset of Diabetes

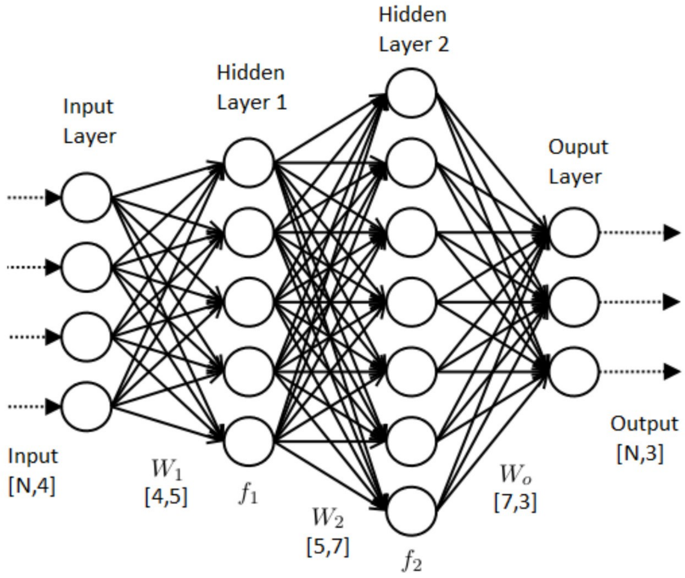
Dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Data atrribute information by columns
The first column represents Pregnancies: Number of times pregnan.t
The second column represents Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
The third column represents BloodPressure: Diastolic blood pressure (mm Hg).
The fourth column represents SkinThickness: Triceps skin fold thickness (mm)).
The fifth column representsInsulin: 2-Hour serum insulin (mu U/ml)).
The sixth column represents BMI: Body mass index weight in kg/height.
The seventh column represents DiabetesPedigreeFunction: Diabetes pedigree function).
The eighth column represents Age: Age (years)).
The ninth column represents Outcome: Class variable (0 or 1)).

Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing. Raw data (”unprocessed data”) is a collection of numbers or characters before it has been ”cleaned” and corrected by researchers. Raw data needs to be corrected to remove outliers or obvious instrument or data entry errors (e.g., a thermometer reading from an outdoor Arctic location recording a tropical temperature). Data processing commonly occurs by stages, and the ”processed data” from one stage may be considered the ”raw data” of the next stage. Field data is raw data that is collected in an uncontrolled ”in situ” environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording. Data has been described as the new oil of the digital economy.

II. Data as inputs

We are using our data belonging to each csv attribute and we are implementing them as input for our neural network, each data from csv file represents single input layer for our neura network.



Input layer-The neurons, colored in purple, represent the input data. These can be as simple as scalars or more complex like vectors or multidimensional matrices. Input layer-The neurons, colored in purple, represent the input data. These can be as simple as scalars or more complex like vectors or multidimensional matrices.

III. Code

Before we started doing out project, we needed to download a few python packages, and we used Python 3.5 as well.

At first we set random seed value , than we loaded our dataset into ”diabetescsv”. This dataset consist from data that are equal for past 5 year of medical history. After that we wanted to split our given data from csv into two variables, first inputs and second final. Inputs represents list of data that is strictly defined - data are divided by row. Final value represents our final decision, it means we are evaluating if person has diabetes or no.

For creating neural network model we chose Keras, but there is a lot of alternatives for creating model of NN model. With Keras we defined each one layer of neural network, we created input layers, hidden layers and outut layers as well. As input values we used our values from diabetes.csv (Pregnancies,Glucose, BloodPressure, etc.) When the data coming in our network, we are multiplying it by matrix of weights. Essentially we are taking in some data than we are performing a bunch different operations on it and then we are taking that big huge number and we are squishing it down, i the case of rail we are basically taking that big huge number we are either assigning it as 0 or a 1 to. We are using rel u function, but there is a lot of other activation functions that you can use, but in our case we decided to chose a rel u function, because the rel u is the most effective for training NN models in a efficient way.

References

[1] HRECHAK, Andrew K.; MCHUGH, James A. Automated fingerprint recognition using structural matching. Pattern Recognition, 1990, 23.8: 893-904.

[2] KARU, Kalle; JAIN, Anil K. Fingerprint classification. Pattern recognition, 1996, 29.3: 389-404.

[3] RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[4] LUK, Andrew, et al. A two-level classifier for fingerprint recognition. In: 1991., IEEE International Sympoisum on Circuits and Systems. IEEE, 1991. p. 2625-2628.

[5] O’GORMAN, Lawrence; NICKERSON, Jeffrey V. An approach to fingerprint filter design. Pattern recognition, 1989, 22.1: 29-38.

[6] LEUNG, M.-T.; ENGELER, W. E.; FRANK, P. Fingerprint image processing using neural networks. In: IEEE TENCON’90: 1990 IEEE Region 10 Conference on Computer and Communication Systems. Conference Proceedings. IEEE, 1990. p. 582-586.

[7] FEIGIN, G.; BEN-YOSEF, N. Line thinning algorithm. In: Applications of Digital Image Processing V. International Society for Optics and Photonics, 1983. p. 108-112.

[8] COWAN, Jimmy Cripe, et al. Untitled-International Institute of Informatics and Systemics.