Artificial neural network project for artificial intelligence

1. President Armando

P.A. is a project from our team composed of: Armando Di Stasio, Giulia Motta, Federico Di Bari, Gaetano Daniele Calcina, aimed to create an ANN for Parkinson's disease prediction. The dataset we have used is a compendium of the voice recordings of 31 patients, some with Parkinson's disease and some without, for a total of 195 entries, represented by the rows of our dataset. The columns are the features considered when analyzing the recordings; they are:

Name: ASCII subject name and recording number.

MDVP:Fo(Hz): Average vocal fundamental frequency.

MDVP:Fhi(Hz): Maximum vocal fundamental frequency.

MDVP:Flo(Hz): Minimum vocal fundamental frequency.

MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP: Several measures of variation in fundamental frequency.

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA: Several measures of variation in amplitude.

NHR, HNR: Two measures of the ratio of noise to tonal components in the voice.

Status: The health status of the subject that we have taken as the output of our neural network. One= Sick (has P.D.) Zero= Healthy.

RPDE, D2: Two nonlinear dynamical complexity measures.

DFA: Signal fractal scaling exponent.

Spread1, spread2, PPE: Three nonlinear measures of fundamental frequency variation.

We have divided the dataset, following the 80%-20% rule, into a training set of 155 data points and a test set of 40. This is to ensure that we have some actual data to confront the performance of the system in the phase of evaluation.

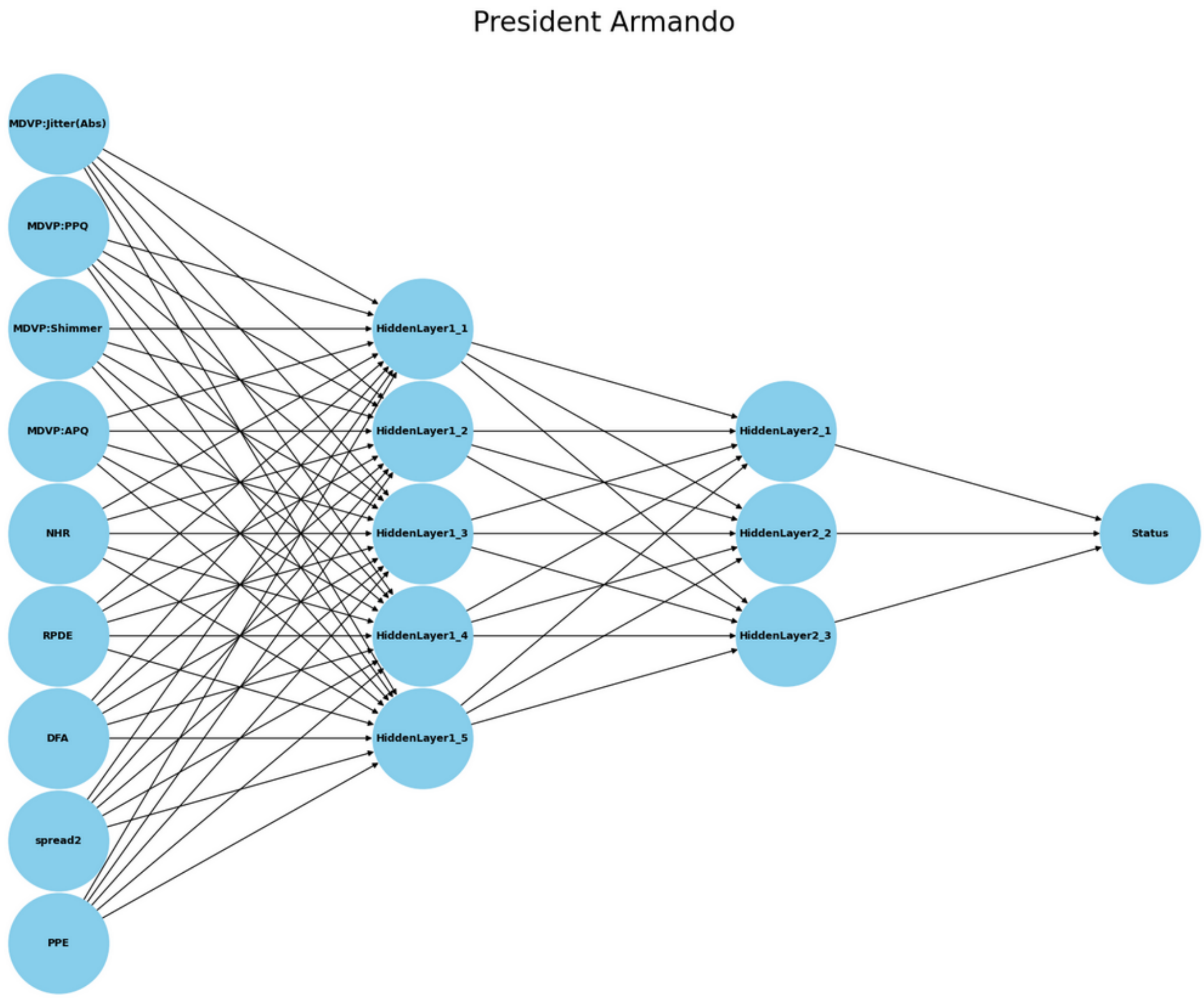
From here, we developed the graphs and, by looking closely at the data, we took some choices in the sense of:

- Cutting all features with relevance <10%
- Cutting redundant features

We then normalized the data and programmed the Neural Network. President Armando has:

- 9 input neurons (one for each relevant feature);
- A first hidden layer composed of 5 neurons;
- A second hidden layer composed of 3 neurons;
- 1 output neuron (Has Parkinson? Y/N);

Here’s a graphical representation of our network:



Finally, we have programmed a UI for a better user experience.

To inquiry President Armando you need to input 9 different features taken from the voice recording of a patient. Evidently, you need a machine capable of detecting the same 9 features the network requires to run.

2. Graphs

In our project we made considerable use of graphs. Here we give the user a quick guide to correctly read them.

Graph A: a heatmap[1] of the whole dataset correlation matrix. This heatmap shows the correlation between all features in the original dataset. The colors represent the strength and direction of correlations:

- Lighter colors indicate a positive correlation;
- Darker colors indicate a negative correlation;

[1] A heatmap is a graphical representation of data where values are depicted with colors. The color intensity is directly proportionate to how strongly the different variables are correlated with each other.

- White represents the correlation of each column with itself.

That is, the lighter the color, the stronger the correlation.

The diagonal, which is always 1, indicates the perfect correlation with itself. This visualization helps understand and identify which features are closely related to each other but also it is of great help in identifying potentially redundant features.

Graph B: shows us the correlation of all columns with 'status'. In this case, positive correlations suggest that the feature can be a good indicator for a P.D. diagnosis, negative correlations suggest that it is not relevant.

Graph C: Correlation of all the columns with each other. This graph is similar to the first one but it is made on the cleaned dataset after the initial feature selection, 'status' column excluded. Again, here we find the visualization of the correlations between features, if some of them are similar or redundant and the possible relationships between them.

Graph D: as the name says – Correlation of all the columns with each other, filtered – this graph shows the correlation of all the columns with each other, same as Graph C, but this time, it shows only correlations below 0.95 threshold. This graph is helpful in feature selection as it shows only moderately correlated features, making the correlation matrix clearer and, more importantly, allowing us to find redundant values in order to allow us to ease the computation by dropping them for further cleaning.

Graph E: 'Accuracy over epochs' shows how the model's accuracy improves during training, which is a visualization of how the NN learns and improves its predictions over time. Generally speaking, you want to see the accuracy increasing and stabilizing. In this graph, the x-axis represents the training epochs while the y-axis represents the model accuracy.

Graph F: 'Loss over epochs' shows the loss (error) and how it changes during training. A good training process shows the loss decreasing over time so to minimize prediction errors. Here, the x-axis represents training epochs while the y-axis presents the loss value.

Graph G: Confusion Matrix shows the model's predictions performance. In this graph, rows represent actual classes (Healthy/Sick), and columns represent predicted classes (Healthy/Sick). The numbers in the cells represent the number of predictions:

- Top-left (true negatives): correctly predicted healthy cases;
- Bottom-right (true positives): correctly predicted Parkinson's cases;
- Top-right (false positives): healthy cases incorrectly predicted as Parkinson's;
- Bottom-left (false negatives): Parkinson's cases incorrectly predicted as healthy.

3. Issues

In the process of developing the network we have encountered a few issues. Firstly, regarding the dataset. We previously had a more exhaustive and diverse dataset, but it was lacking in clarity. It had over 200.000 datapoints but it was hard to decipher the meaning of the feature selected for the inquiry and a lot of the data was not taken homogeneously. For these reasons, we opted to look for a better and more reliable backbone for our network. The one we are using required little manipulation since it was very clear and tidy in the first place. The features are set out comprehensively, the data are gathered coherently with one another. Even though we must point out that this dataset is far from being considered complete and exhaustive. For the predictions of the network to be considered reliable, the network should be fed with a lot more data we are lacking. Still, we found the results to be satisfying enough for this project.

A second type of obstacle we faced has to do with designing the network itself. Initially we struggled choosing the number of neurons and hidden layers composing the ANN as well as deciding the type of activation function for the neurons (sigmoid, linear etc.). To overcome this difficulty, we used a heuristic approach. That is to say that we kept in mind some specific parameters given by the machine (loss function, accuracy, learning rate) to evaluate the performance of the machine in relation to the changes we made. We found the current build to be the most stable and best performing.

The third notable difficulty was found in the choice of graphs. Easily overlooked, the graphical representation molds our perception of the data and their respective salience. Throughout the project we decided to change dispersion graphs for histograms or cartesian diagrams, as they gave us a better picture.

4. Conclusions

In the end, we think we accomplished a very satisfying project. Despite the epistemic problem posed by the range of the dataset, our neural network is a good display of the capabilities and advantages of ANN.

Dataset taken from:

Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering.