Individual Study: Basic concepts of Machine Learning

Die Hu

Zheng Gao(Mentor)

Abstract

An independent study on basic concepts of Machine Learning and entry level of data analysis including data visualization by python.
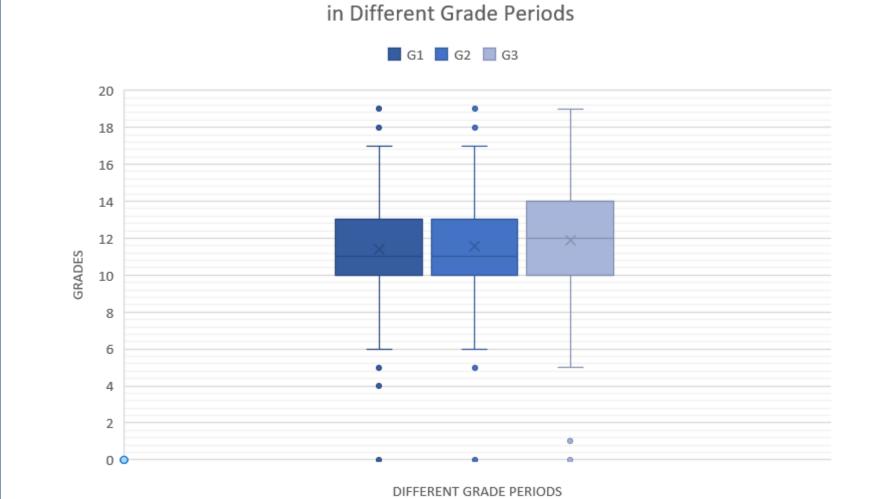
In this study, an artificial neural network model from python package 'sklearn' is used to predicting the likely performance of a student on the first period grade based on datasets from UCI Machine Learning Repository that the data consisted approach student achievement in secondary education of two Portuguese school for Portuguese language class. (Reference 1)

The data contains attributes, for example, include information of the 395 students' grades, demographic and social that collected by school reports and questionnaires. (Figure 4)

Students' academic performance may likely to be influenced by various factors, such factors were considered as attributes in the dataset and input for the prediction model. The prediction model could be used by advisors to predict the students' performances based on the student's information more precise and efficient that save times from reading transcripts and basic information of the student and compare with the other students' information that the advisor knows.

Introduction

- Machine Learning
  - Machine Learning is the scientific study of algorithms and statistical models that the computer system use to perform a specific task without using explicit instruction but rely on the patterns and inference instead. (Reference 2)
- Artificial Neural Network (ANN)
  - Perceptions (Reference 4)
    - Artificial Neural Networks (ANN) are inspired by the human neural network architecture but instead of billions of neurons, the ANN can only consist one neuron called a perceptron that has one input layer to receive inputs. (Figure 2) As simple as that could be, the ANN can also contain multilayer perceptions as the Human Nervous System. (Figure 3)
  - The Human Nervous System (Reference 4)
    - Human nervous system consists of billions of neurons and collectively process input received from sensory organs, process the information and make decision. The human nervous system contains different types of neuron to receive, process and pass information to others via the axon. (Figure 1)
- Python: mostly used in the field of Machine Learning
  - Sklearn: a package of Python
    - sklearn.neural_network.MLPClassifier() : Method for Multi-layer Perceptron classifier
      - class sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000) (Reference 3)

Students Academic Performance in Different Grade Periods

Steps

1. Make the datasets to be encoded that all the category data originally (Figure 4) transformed to be numerical data (Figure 5).
2. Select the amount of variables as input for the model before train and test the model.
3. Make the prediction and the accuracy rate and adjust the input values by the accuracy.

Conclusion

The purpose of the predication model is to predict the Final Grade (G3) that is 0 out of 20 points.

The first model that contains 32 variables as input, the accuracy of the prediction is 46%.

The second model that contains 2 variables (G1 and G2) as input, the accuracy of the prediction is 53% with some changes in the hidden_layer_sizes parameter from sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000).

Result

```
      precision  recall  f1-score  support                    precision  recall  f1-score  support
   0      0.25    1.00      0.40        1            0            0.00    0.00      0.00        6
   6      0.00    0.00      0.00        1            6            0.00    0.00      0.00        1
   7      0.00    0.00      0.00        2            7            0.00    0.00      0.00        3
   8      0.40    0.67      0.50        6            8            0.29    0.62      0.40        8
   9      0.33    0.11      0.17        9            9            0.00    0.00      0.00        6
  10      0.65    0.61      0.63       28           10            0.62    0.59      0.60       27
  11      0.57    0.73      0.64       22           11            0.61    0.95      0.74       21
  12      0.40    0.38      0.39       16           12            0.44    0.36      0.40       11
  13      0.36    0.38      0.37       13           13            0.50    0.69      0.58       13
  14      0.17    0.10      0.12       10           14            0.67    0.20      0.31       10
  15      0.25    0.50      0.33        6           15            0.88    0.78      0.82        9
  16      0.50    0.25      0.33        8           16            0.75    0.30      0.43        9
  17      0.57    0.67      0.62        6           17            0.22    0.50      0.31        4
  18      0.00    0.00      0.00        2           18            0.33    1.00      0.50        1

accuracy                    0.46      130       accuracy                        0.53      130
macro avg   0.32    0.38    0.32      130       macro avg    0.38    0.43        0.36      130
weighted avg 0.45   0.46    0.44      130       weighted avg 0.51   0.53        0.49      130
```

Figure 6. The Prediction Accuracy with 32 variables input   **Figure 7.** The Prediction Accuracy with 2 variables input

Conclusions

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13

5 rows × 33 columns

Figure 4. Example of Original Dataset

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2
0	0	0	3	1	0	0	4	4	0	4	...	0	4	3	2	3	0	0	4	0	7
1	0	0	2	1	0	1	1	1	0	2	...	0	4	2	2	0	0	2	2	6	7
2	0	0	0	1	1	1	1	1	0	2	...	0	3	2	1	1	2	2	6	9	9
3	0	0	0	1	0	1	4	2	1	3	...	1	2	1	1	0	0	4	0	11	10
4	0	0	1	1	0	1	3	3	2	2	...	0	3	2	1	0	1	4	0	8	9

5 rows × 32 columns

Figure 5. Example of Encoded Dataset

Figure 2. Single Layer Artificial Neural Network

Figure 3. Multiple Layers Artificial Neural Network

Figure 1. The Human Nervous System

References

1. Database(Source data): P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. https://archive.ics.uci.edu/ml/datasets/student+performance
2. Wikipedia contributors. (2019, December 5). Machine learning. In Wikipedia, The Free Encyclopedia. Retrieved 05:23, December 6, 2019, from https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=929436005
3. sklearn.neural_network.MLPClassifier¶. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
4. Robinson, S. (2018, February 6). Introduction to Neural Networks with Scikit-Learn. Retrieved from https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/.