Analyzation and application of PyMC3

Zhenghong Lei

University of Liverpool, Liverpool, L69 7ZX, UK

Abstract. This article studies and analyzes the data of “student study hour”. The purpose of this article is to analyze the relationship between students' learning hours and the score. Data analysis involves machine learning and algorithms. Therefore, these two parts and their related knowledge need to be learned and mastered. In the following part, the article will first introduce the utilization methods, which are machine learning, Bayesian method and PyMC3. Then, the article introduces the details of the analysis steps for the “student study hour” data. The conclusion will be presented at the end of the article, which concludes the linear relation of the data, and the pros and cons of PyMC3.

Keywords: Machine Learning; Bayesian Method; PyMC3.

1. The introducing of the dataset

Research on deep learning is constantly deepening. While exploring, we also studied many meaningful models and sequences. Such as ARIMA and grey prediction models.

ARIMA: Autoregressive comprehensive moving average model. The homogeneous non-stationary time series is a non-stationary time series that can be transformed into a stationary time series after the process of difference. ARIMA model refers to non-stationary time series, while ARMA model refers to stationary time series.

Grey prediction model: a method for predicting systems with uncertain factors. By recognizing the level of difference between the development trends of system, generating and processing the data. We can find the regular pattern of system change, and then establish an equivalent differential equation model to predict the development trend of factors in the future.

However, both of them have some drawbacks when using them. ARIMA requires that the data of time series should not be mobile or unstable after being differentiated. In addition, it cannot generate the nonlinear relationship but linear relationship. Besides, grey forecasting model is only suitable for medium and short-term forecasting.

Hence, this article focuses on a new model based on PyMC3, which is more accurate and faster than ARIMA and grey forecasting model.

Project dataset: “Student Study Hours” The data is obtained from Kaggle (the link above). Written and uploaded by Himanshu Nakrani [1].

(https://www.kaggle.com/datasets/himanshunakrani/student-study-hours)

The dataset consists of two columns. That is the study hours of the students and the scores they got. we can apply simple linear regression to predict the marks of the student given their number of study hours [1].

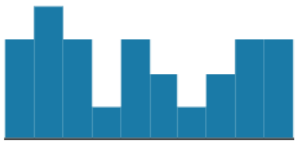
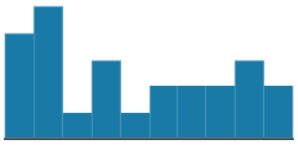
# Hours number of study hours	# Scores marks of the student
 1.1 9.2	 17 95
2.5	21
5.1	47
3.2	27
8.5	75
3.5	30
1.5	20

Fig. 1, List of data

This article is structured as follow: Section 2 (Data processing) presents the work before build the PyMC3-based model and calculate a simple result. Then, show the details of building the model during Section 3 (model). Next, Section 4 (Experiment of the model) will show how to analysis the relationship between the data in specific. Meanwhile, evaluating will also be presented at Section 4. Finally, the conclusion will be given at Section 5.

2. Data processing

2.1 Import the data

Before using the data, we first deleted the duplicate data and outlier data. In addition, some processes are exerting on the data, such as converting time into hours.

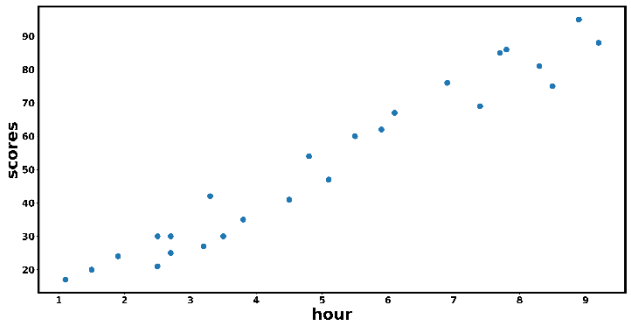


Fig. 2, Visual point diagram of data

2.2 Simple relationship calculates

Use LinearRegression imported from sklearn.linear_model to calculate the simple linear relationship between hours and score. Assumed that hours is the independent variable and score is the dependent variable. Meanwhile, assumed that the relationship between hours and score is linear. Hence,

$f(x_i) = wx_i + b_i$,

where “w” is weight (reg.coef_), “b_i” is bias (reg.intercept), $f(x_i)$ is the score and x_i is the study hours in the code.

Finally, the result is shown as figure 3:

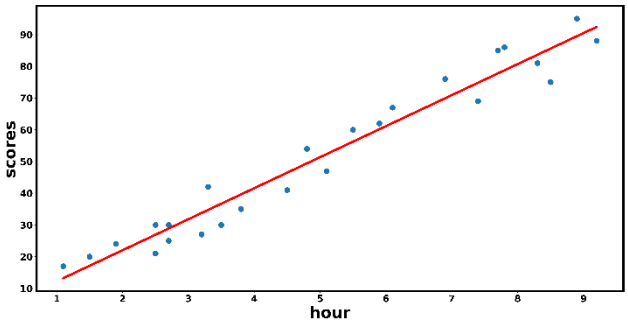


Fig. 3, Simple linear relationship analysis

By printing the result of reg.intercept (b_i) and reg.coef_[0] (w), the result are 2.48367 and 9.7758033 respectively, which are

$score = 9.7758033 \times hours + 2.48367.$

Obviously, there are some deviations between the obtained linear relationship results and the actual results. Furthermore, we need to use PyMC3 method to help us optimize the result.

3. Model

```
[ ]  model = pm.Model()

     with model:

          m = pm.Normal('m', mu=0.0, sd=1.0)

          c = pm.Normal('c', mu=0.0, sd=1.0)

          sd = pm.Exponential('sd', lam=1.0)

          maximum value = pm.Exponential('maximum value', lam=1.0/10.0)

          mu = pm.math.minimum(maximum_value, m * x + c)

          Y_obs = pm.Normal('Y_obs', mu=mu, sd=sd, observed=y)
```

Fig. 4, The code for modeling

The model definition language in PyMC3 is very similar to the formula in statistics. Generally, one statement corresponds to one formula [2]. Take the code in the "student study hours" project as an example, which are shown in the figure 4:

model = pm.Model()

Define a new object of the model, which is the stochastic variable container of the model.

with model

After that, use the “with” statement to define a context manager. The “model” is used as the context, and the variables of definition in this context are added to the “model”.

m = pm.Normal('m', mu=0.0, sd=1.0)

c = pm.Normal('c', mu=0.0, sd=1.0)

sd = pm.Exponential('sd', lam=1.0)

maximum_value = pm.Exponential('maximum_value', lam=1.0/10.0)

Define Four stochastic random variables with normal distribution prior.

Y_obs = pm.Normal('Y_obs', mu=mu, sd=sd, observed=y)

Finally, define the observed value, which is a special observed random variable, refers to the possibility of the data of model. The observed parameter is used to tell this variable that its value has

been observed (i.e., variable Y), and will not be changed by the fitting algorithm. Both numpy.ndarry and pandas.DataFrame can be used to pass in as a parameter.

4. Experiment of the model

4.1 Find the maximum a posteriori (MAP) of parameters using optimization method

Using the optimization method to find the point estimation is fast and simple. PyMC3 provides "find_MAP" function which returns an estimated value (point) of the parameter. Moreover, "find_MAP" uses the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm to optimize the maximum value of the log posterior distribution [2].

It is worth noting that map estimation is not always valid, especially in extreme model cases. In the high-dimensional posterior, the probability density of occurrence may be large, but the overall probability is very small, which is common in the hierarchical model.

4.2 Use the sampling method to obtain the posterior distribution to calculate (NUTS)

This method is used in this project, because this kind of sampling method can obtain a posteriori distribution close to the real sampling value. NUTS algorithm is based on the gradient of logarithmic probability density and uses the region information of high probability density to obtain faster convergence speed than traditional sampling algorithm [2].

Take the code in the "student study hours" project as an example, which are shown in the following figure:

with model:

trace = pm.sample(draws=2000)

Sample the posterior distribution of the model 2000 times.

with model:

pm.traceplot(trace, var_names=['m', 'c', 'sd', 'maximum_value']);

Use the trace plot function provided by PyMC3 to plot the trend of posterior sampling. (As shown in the figure 5):

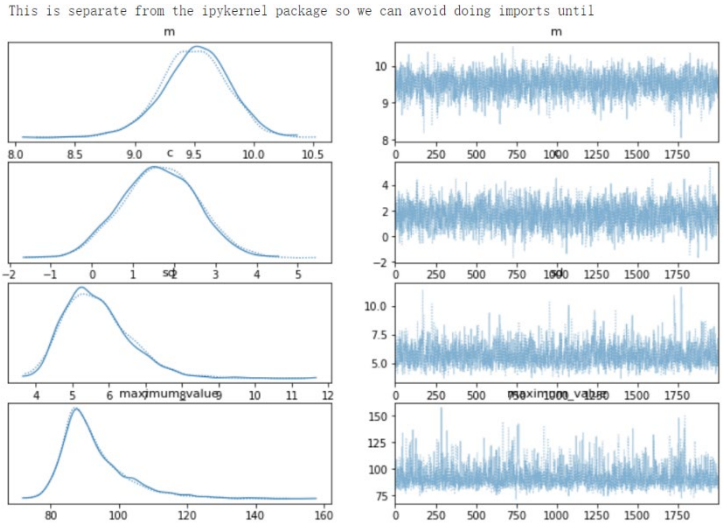


Fig. 5, Difference analysis results of PyMC3

The left side is the histogram of the marginal posterior of each random variable, which is smoothed using kernel density estimation. On the right is the Markov chain sample values plotted in order. For vector parameter beta, there are two posterior distribution histograms and posterior sampling values.

4.3 Result

After the training and testing by PyMc3 probabilistic machine learning tools, we get new analysis data which is more accurate. the result is shown as figure 6:

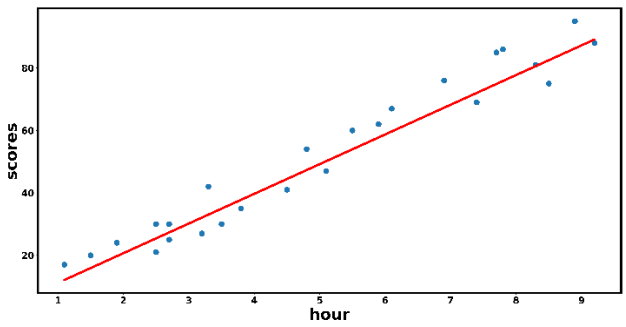


Fig. 6, Linear relationship analysis after optimize by PyMC3

By printing the result of reg.intercept (b_i) and reg.coef_[0] (w), the result are 1.65088 and 9.507238 respectively.

From the above analysis results, it is obviously that there is a linear relationship between students' learning hours and the scores, which is

$score = 9.5072 \times hours + 1.65088.$

That means, the longer time students study, the higher score the students may achieve.

4.4 Evaluating

In order to evaluate the method proposed, the data were divided into two parts, the number of training group and testing group are 7:3. Therefore, since the total number of datasets is 25, there are 18 training sets and 7 testing sets. Compare the relationship between the final results of the two groups and the results obtained above.

Training set

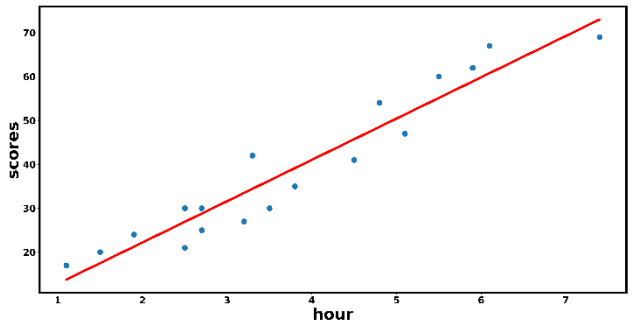


Fig. 7, Linear relationship analysis of the training set

By printing the reg.coef_[0] (w), which are 9.4061, the result of the relationship is

$$score = 9.4061 \times hours + 1.68175$$

Testing set

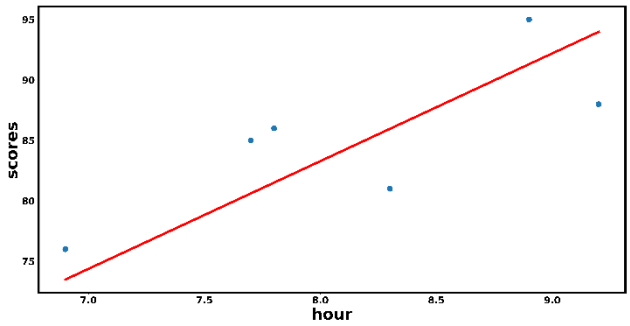


Fig. 8, Linear relationship analysis of the testing set

By printing the reg.coef_[0] (w), which are 7.916373, the result of the relationship is

$$score = 8.91041 \times hours + 1.59014$$

Obviously, there is small difference between training group and testing group in terms of results. However, the two results do not have very huge difference between each other and the final result shows in figure 6. Therefore, PyMC3 has better effects than other future predicting model.

5. Conclusion

In a nutshell, this article introduces the analysis of Student study hours datasets with PyMC3.

Then, the dataset is imported into the system, and the system simply calculates and print the relationship between the study hours and score. Next, the system builds a PyMC3-based model and train over 2000 epochs. Meanwhile, the system gets a new result that is more accurate than the first one. Finally, the data is divided into two parts, which are training part and testing part. And result confirms that the PyMC3-based model is more accurate than the simple one.

On the other hand, the data has some shortcomings. Although the relationship between study hours and score is linear, students' study time cannot reach infinity. When the hours increase enough, the improvement of score will gradually become smaller, which cannot be analyzed through this set of data.

In conclusion, PyMC3 is much faster in analysis and algorithm than other speeds, and it is easy to implement the models. However, it is not particularly perfect in the analysis of linear regression [3]. In general, the future development of PyMC3 is worth looking forward to.

6. Relevant knowledge

6.1 Machine Learning

Introduction

Machine learning is a computer science subject that study problems by modeling assumptions, which learn model factors from the training data, and finally predicts the data [2, 18]. Machine learning does not require writing any problem-relating code specifically. Generic algorithms can tell the analyzing result about the data. The generic algorithm will present the bottom logic of the data after importing.

Machine Learning - Supervised learning

The input data is composed of the input characteristic value and the target value. The output of a function can be regression or classification.

The characteristic of regression is that the annotated data set has variables of numerical target [3]. In other words, each observation sample has a numerical labeled true value to supervise the algorithm.

Regression

It refers to the classification of various things and is used for discrete prediction (Picture recognition of cats and dogs) [4, 5]. For example, Bayesian classification, logistic regression.

Classification

Predict continuous, specific values (the prediction of house prices) [4]. For example: linear regression, ridge regression.

Machine Learning - Unsupervised learning

Introduction

Unsupervised learning problem means that the model learns without any indication. Since there is no indication, it means that the only thing that the model can learn is what the sample itself contains, such as the similarity and pattern between the samples, or the distribution of the population where the samples are located [6].

Compared with supervised learning, unsupervised learning data is not labeled. We give the data to the machine and let it try to find some structure from the data.

Unsupervised learning enables us to deal with problems without knowing what the results should be. We can derive structures from data that do not necessarily know the effects of variables.

We can cluster the data according to the relationship between the variables in the data, thus obtaining this structure [7]. In unsupervised learning, there is no feedback based on the prediction results.

Characteristics

Unsupervised learning purpose is not clear.

Unsupervised learning does not require the data is labeled.

Unsupervised learning cannot quantify the effect [7].

6.2 Bayesian Method

We often hear from the weather forecast that the precipitation rate tomorrow will be 80%. What does this mean? It is difficult for us to explain this statement directly, especially from the perspective of the probability School: it is unrealistic to repeat rain / no rain experiments indefinitely (or not many times). Bayesian method can explain this [8].

Bayesian world view interprets probability as a measure of the reliability of an event, that is, how much confidence we have in the occurrence of an event [8]. This means that in Bayesian method, we can never absolutely determine our "belief", but we can definitely express how much confidence we have in the occurrence of relevant events. In addition, as more data are collected, we can have more confidence in our own beliefs.

Bayesian analysis

Bayesian analysis provides a method to support the calculating of probability of hypothesis, which is based on the prior probability of hypothesis, and observes the probability of different data under given assumptions [9].

The theory of Bayesian statistical

In the theory of Bayesian statistical, the correlation quantity in statistical inference is considered as a stochastic quantity, no matter whether it generates a stochastic value. Probability is considered to mean that the same probability of random events may have the same probability on the same level of broken knowledge [9].

According to Bayesian theory, even though each estimator is different from the correlation measured, it is used to estimate the measured undetermined true value [10].

For simplicity, the estimator, the value of the estimator and the measured value are all represented by the same symbol. This can be distinguished from the context without confusion, because the sample is a random variable and the sample value is some constant, which is different from the classical statistical theory [10].

MCMC (Markov chain Monte Carlo)

The basic idea of MCMC method is to construct a Markov chain, then the posterior distribution samples is generated through the Markov chain. When the Markov chain reaches a stable distribution, perform Monte Carlo integration based on valid samples [11].

Markov chain

Russian mathematician Andrey Andreyevich Markov studied and proposed a general law model that can explain natural changes by mathematical methods, which is named Markov chain [12, 13]. Markov chain is a stochastic process that transit by going through one state to another. Meanwhile, a process called "memoryless" is required, that is, the next state's probability distribution is only related to the current state [14], which is called Markov property. Random process, that is based on probability theory and mathematical statistics, and has Markov property is called Markov chain (MC) [12].

Monte Carlo

Monte Carlo method is basically describe various stochastic phenomena when the equipment is operating. This method can comprehensively reflect the influence of stochastic factors on the operating process of the equipment [15]. More accurately shows the active process of application.

Simulating the random phenomenon in the random service system and calculating its digital characteristics; Some complex equipment operation actions are simplified into a series of connected events through reasonable decomposition, and then each event is simulated by random sampling method to finally achieve the purpose of simulating equipment operation activities or operation processes [15].

6.3 The introducing of PyMC3

PyMC3 has advanced next-generation MCMC sampling algorithms such as No-U-Turn sampler (NUTS) and Hamiltonian Monte Carlo (HMC) [2], which has a great positive effect on the high-dimensional and complicated posterior distribution, supporting the fitting of complicated models when the lacking the particular knowledge of the fitting algorithm is occurred. NUTS and HMC algorithms converge much faster than the traditional one, especially for the large models. In addition, NUTS contains a set of self-adjustment processes, so that the system do not requiring users to know whole algorithm specifically. [2].

PyMC3 is a Python-based open-source probabilistic programming framework. Through PyMC3, we can flexibly create a self-defined probability model and perform Bayesian inference to achieve insight and learning in the data. At the same time, because it is based on Bayesian method, it is often necessary to specify a priori distribution to constrain our model in the process of use, so as to obtain the uncertainty estimation of the posterior distribution of the unknown quantity [17].

References

[1] NAKRANI, H., 2022. Student Study Hours. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/himanshunakrani/student-study-hours> [Accessed 2 September 2022].

[2] Xu, J., 2022. PyMC3 - Introduction and introduction. [online] Blog.csdn.net. Available at: <https://blog.csdn.net/jackxu8/article/details/71080865> [Accessed 2 September 2022].

[3] Zhihu.com. 2022. How to evaluate PyMC3? [online] Available at: <https://www.zhihu.com/question/49044755/answer/203024580> [Accessed 2 September 2022].

[4] Zh.wikipedia.org. 2022. supervised learning-Wikipedia. [online] Available at: <https://zh.wikipedia.org/wiki/%E7%9B%91%E7%9D%A3%E5%AD%A6%E4%B9%A0> [Accessed 2 September 2022].

[5] Artificial intelligence learning library. 2022. Supervised learning. [online] Available at: <https://easyai.tech/ai-definition/supervised-learning/> [Accessed 2 September 2022].

[6] Shui, Y., 2022. [Machine learning note] Unsupervised Learning. [online] Blog.csdn.net. Available at: <https://blog.csdn.net/shuiyixin/article/details/80185207> [Accessed 2 September 2022].

[7] Zhihu. 2022. Machine learning: what is unsupervised learning. [online] Available at: <https://zhuanlan.zhihu.com/p/355600736> [Accessed 2 September 2022].

[8] Nguye, T., 2022. PyMC3 introduction: Python package for probability programming. [online] Blog.csdn.net. Available at: <https://blog.csdn.net/tMb8Z9Vdm66wH68VX1/article/details/112597594> [Accessed 2 September 2022].

[9] Yang, X., 2022. Artificial intelligence and machine translation. 1st ed. Southwest Jiaotong University, p.page 233.

[10] Wang zhong yu, 2010. Small sample non statistical analysis principle of precision instruments. Beijing: Beijing University of Aeronautics and Astronautics Press.

[11] Zhu, X., 2022. Review of Markov chain Monte Carlo method.

[12] Brémaud, P. . Markov chains: Gibbs fields, Monte Carlo simulation, and queues (Vol. 31) ： Springer Science & Business Media，1999：Chapter 2-4, pp.53-156

[13] Zhihu. 2022. Brief introduction to Markov chain. [online] Available at: <https://zhuanlan.zhihu.com/p/448575579> [Accessed 2 September 2022].

[14] Serfozo, R. . Basics of applied stochastic processes. ： Springer Science & Business Media.，2009：pp.1-99, 241-242.

[15] En.wikipedia.org. 2022. Monte Carlo method - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Monte_Carlo_method> [Accessed 2 September 2022].

[16] "What is Machine Learning?". www.ibm.com. Retrieved 2021-08-15.

[17] Li, X., 2022. PyMC3 module. [online] Cnblogs.com. Available at: <https://www.cnblogs.com/lixin2011/p/13975487.html> [Accessed 2 September 2022].

[18] Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

[19] Machine learning and pattern recognition "can be viewed as two facets of the same field."