Comparisons of three sets of similar terminologies in data science

Nan Du z5245818

Abstract

In this essay, three main issues will be discussed: comparisons of machine learning and data mining, comparisons of linear and logistic regression models, and comparisons of neural networks and logistic regression models. Divided in three parts, in each part, first it will introduce the terminologies; then comparisons will be made.

Introduction

In the field of data science, there exist a large amount of data analyzation methods, and it is important for students to fully understand each term of those analysis methods. In short, the purpose of this essay is to explore the similarities and differences between some terms which may confuse most people. By referring to other essays, the relations between the terms will be discussed.

Machine Learning and Data Mining (410 words)

From materials given in the lectures, machine learning is a subsection of artificial intelligence which mainly focuses on 'the design and study of algorithms to build mathematical models based on data set'. There is a more general definition given by Samuel (1959), machine learning is a process of giving computers the ability to learn automatically. That is, no explicit programming is required for computers to learn and analyze data. Pattern classification, control, regression and prediction, and clustering are some major problems of machine learning. Lastly, machine learning has four major categories, which are supervised learning, unsupervised learning, semisupervised learning, and Reinforcement Learning (Géron, 2019). Some examples of applications include cancer prognosis and prediction (Kourou, 2015), flagging off spam emails, and analyzing images of products to classify the products automatically (Géron, 2019).

Similarly, from the materials given in the lectures, data mining mainly focuses on the extraction of information from large data sets into useful structure for data analysis. A more specific definition was given by Gupta (2014). According to Gupta (2014), data mining, also known as knowledge discovery in databases (KDD), is a set of exploration techniques including machine learning, statistics and database systems based on advanced analytical methods and tools to extract information from a large amount of information. The applications of data mining include educational data mining (EDM) (Romero and Ventura, 2013), data mining in health care and business areas to obtain predictive models.

Since both data mining and machine learning involves data, one similarity they share is that both need to develop methods and procedure to process data (Mirkin, 2011). As for the differences, they lie in different perspectives.

One important feature that distinguishes data mining and machine learning is that data mining employs machine learning to analyze the data. As stated in the essay before, the purpose of data mining is to subtract useful information from large data sets. The tools used include machine learning. Whilst machine learning, the 'tool', focuses more on designing algorithms to build mathematical models according to the data sets. Another difference, given by the lecture slides, is that data mining needs human interference, while machine learning is completely automated once design self-implemented.

Lastly, some other differences between machine learning and data mining are given in the lecture slides. For example, machine learning has a wider application scope than data mining. It can be applied in other tasks such as creating a chatbot or personal assistant (Géron, 2019), whilst data mining can only be applied in a limited data-based area.

Linear and Logistic Regression Models (431 words)

In the field of data science, linear and logistic regression models are statistical techniques which can be implemented to find and model the relationship between a series of variables (Montgomery, 2021). The following paragraphs will first introduce both models, then discuss the similarities and differences between them.

As the name suggests, linear regression is the process of modelling the relationship of variables into a straight line. Montgomery (2021) gives a graph showing the model created by linear regression.

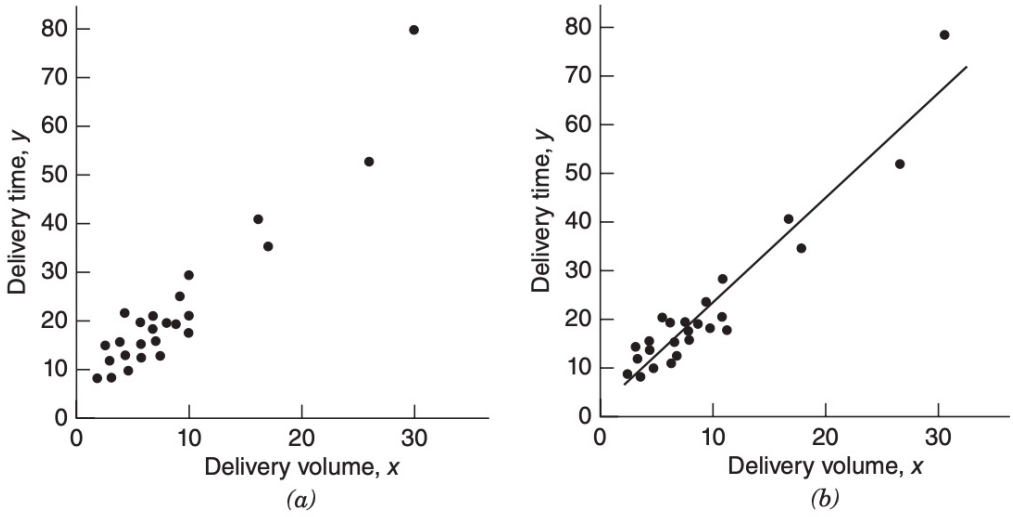

Figure 1.1 Scatter diagram given in the book written by Montgomery (2021).

The equation used by linear regression is $y = \alpha x + \beta + \varepsilon$, where x is the independent value, y is the dependent value, α is the slope of the model, β is the intersect point with y axis, and ε is the difference between the observed value of y and the straight line. For logistic regression, also known as perceptron, it's similar to the linear regression model with the difference as using a logistic function as the activation function. The activation function of logistic regression is $y = \frac{e^x}{e^x+1}$.

The two models have a lot of similarities. Firstly, as discussed previously in this essay, both methods are used to find the relationship between labelled variables. Both linear and

logistic regression models utilize activation functions to model the relationship and possibly draw the model's diagram. Another similarity they share is that both models have error measurements. For the linear regression model, ε is the error; it indicates the distance between the actual point and the straight line. Other error measurement methods can also be applied (to the linear regression model and logistic regression model). Some of the popular measurement functions include MSE, MAPE, MAE and RMSE (Pascual, 2018).

The major difference between those two models is the activation function. As stated before, the activation function of linear regression model is a linear function; and the activation function of logistic model is a logistic (or sigmoid) function.

Another difference is that linear regression is easier to implement, while it is less accurate than the logistic regression model. In *Comparison of Logistic Regression and Linear Regression in Modeling Percentage Data*, written by Zhao (2001), in their experiments, at least 78% of the observations have better prediction when logistic regression is applied. What's more, the logistic regression model has a smaller deviation. The reason for this is that when used for binary data, linear regression has limitations.

Lastly, a difference lies in the applications. In short, linear regression is more often implemented with continuous data, while logistic regression can be used to make a business prediction. Logistic regression models can be used to model small-business credit scoring (Bensic, 2005), analyze customer satisfaction data, failure prediction and much more.

Neural Networks and Logistic Regression (302 words)

The neural network is a machine learning model that was "inspired by the networks of biological neurons" that exists in human brains (Géron, 2019). Picton (1994) defined a neural network as a process of pattern classification: that produces output patterns when given input patterns. Having said that, once a problem can be simplified to pattern classification, it can be solved by the neural network. The application of neural networks includes autonomous driving systems (Mitchell, 1997), healthcare and earthquake prediction.

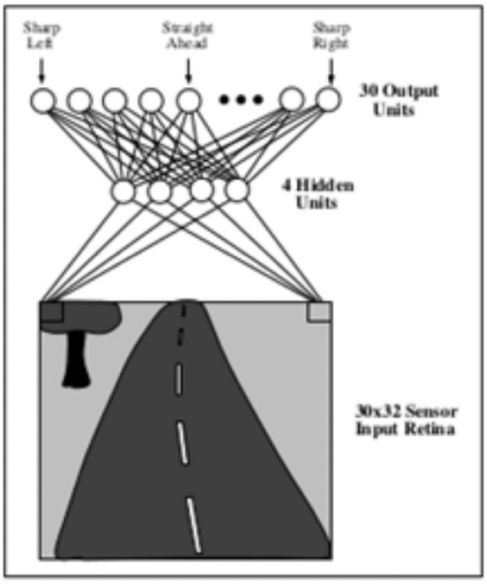

Figure 2.1 Neural network applied in autonomous driving (Mitchell, 1997).

Since the definition of logistic regression has already been given in the previous section, this essay will skip this part and move on to discuss their similarities and differences.

The logistic regression model, also known as perceptron, is a core building block of neural networks. As a simple neuron of neural network, logistic regression is based on a threshold logic unit (TLU) with numbers rather than binary as inputs and outputs (Géron, 2019). According to Dreiseitl (2002), one difference is that neural network models are more likely to be affected by overfitting. The reason is that neural networks are more flexible than logistic regression models. Similarly, in the experiment made by Kumar (1995), they tried to implement both models to estimate the decisions made by supermarket customers on whether to add a new product to their cart or not. The results also indicate that neural networks behave better than logistic regression models. According to Kumar (1995), once a neural network is successfully trained, it will have a better performance than logistic regression. The presence of the hidden layer is the reason behind this.

Another difference lies in the training process. Neural networks generally end up solving non-convex optimization methods; whilst logistic regression models will be used to solve convex optimization problems efficiently.

Conclusion

By citing other resources and making comparisons, this essay illustrated the relationships between those three terminologies. All sets of terminologies share common properties; and they also differ in many perspectives. For the first set, they are both used to process data, while machine learning can be seen as a tool in data mining; for the second set, they are similar estimation models with different activation functions and different applications; for the third set, logistic regression is the simplest form of neural networks, and neural networks tend to have better performances than logistic regressions in most cases.

References

[1] Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2), 206-226.

[2] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc."

[3] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.

[4] Gupta, G. K. (2014). *Introduction to data mining with case studies.* PHI Learning Pvt. Ltd.

[5] Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 3(1), 12-27.

[6] Mirkin, B. (2011, December). Data analysis, mathematical statistics, machine learning, data mining: Similarities and differences. In *2011 International Conference on Advanced Computer Science and Information Systems* (pp. 1-8). IEEE.

[7] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). I*ntroduction to linear regression analysis.* John Wiley & Sons.

[8] Pascual, C. (2018). Tutorial: Understanding Regression Error Metrics in Python. https://www.dataquest.io/blog/understanding-regression-error-metrics/

[9] Zhao, L., Chen, Y., & Schaffner, D. W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. *Applied and environmental microbiology*, *67*(5), 2129–2135. https://doi.org/10.1128/AEM.67.5.2129-2135.2001

[10] Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3), 133-150.

[11] Picton, P. (1994). What is a Neural Network? In: *Introduction to Neural Networks.* Palgrave, London. https://doi.org/10.1007/978-1-349-13530-1_1

[12] Mitchell. (1997). Machine Learning, Maidenhead; U.K: McGraw Hill.

[13] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics,* 35(5-6), 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0

[14] Kumar, A., Rao, V. R., & Soni, H. (1995). An empirical comparison of neural network and logistic regression models. *Marketing letters*, 6(4), 251-263.