Discussion on classification problems in machine learning

Han Yao

Hefei University of Technology, Hefei 230009, China

observer_y@sina.com

Key words: A new observation belongs, category membership, classification.

Abstract. In the research about machine learning, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. In this study, the procedure and principle of resolving classification problems in machine learning is summarized by cookie quality testing.

1. Introduction

Till now, there has been no agreement on the definition of Machine Learning, one ancient definition is by Arthur Samuel: field of study that fives computers the ability to learn without being explicitly programmed. Samuel wrote a checkers playing program that played 10s of thousands games against itself and recorded what board positions lead to what result, the computer learned what board position were leading to success and was able to beat Samuel himself over time .Other studies consider that Machine Learning is an international forum for research on computational approaches to learning. The journal publishes articles reporting substantive results on a wide range of learning methods applied to a variety of learning problems.

Machine Learning (ML) is coming into its own, with a growing recognition that ML can play a key role in a wide range of critical applications, such as data mining, natural language processing, image recognition, and expert systems. ML provides potential solutions in all these domains and more, and is set to be a pillar of our future civilization. The supply of able ML designers has yet to catch up to this demand. A major reason for this is that ML is just plain tricky. This tutorial introduces the basics of Machine Learning theory, laying down the common themes and concepts, making it easy to follow the logic and get comfortable with the topic [1].

Classification and clustering are examples of the more general problem of pattern recognition, which is the assignment of some sort of output value to a given input value. Other examples are regression that assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values; parsing, that assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

In classification problem, as an important part in ML, inputs can be divided into two or more groups, and the learner must produce a model that assigns unseen inputs to one or more of these groups. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email messages or others, and the groups are “spam” and “not spam”. Classification problems in machine learning and statistics is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

Machine learning is used for a range of computing tasks where designing and programming explicit algorithms is not feasible. The applications of ML include spam filtering, optical character recognition, search engines and computer vision [2]. Focusing more on exploratory data analysis, Machine learning is sometimes conflated with data mining [3]. Machine learning and pattern recognition can be considered as two facets of the same field. When used in industrial contexts, machine learning methods may be referred to predictive analytics or predictive modelling [4].

In the terminology of machine learning, classification is viewed as an instance of supervised learning. The corresponding unsupervised procedure is well known as clustering, and which involves classifying data into categories based on some measure of inherent similarity or distance. In general, the individual observations are viewed as a set of quantifiable properties, known variously as explanatory variables or features.

In this paper, the procedure and principle of resolving classification problems in machine learning is summarized by a typical example: cookie quality testing.

2. Cookie quality testing study in Machine Learning.

Under supervised ML, classification machine learning systems can be considered as systems where we seek a yes-or-no prediction, such as “Is this tumor cancerous?”, “Does this cookie meet our quality standards?” and so on [1].

The major differences between Regression machine learning systems and classification machine learning systems are the design of the predictor $h(x)$ and the design of the cost function $J(\theta)$. One classification problem-cookie quality testing study we need to master is as follows:

The results of a cookie quality testing study is shown in Figure1. The training examples have all been marked as either “good cookie” ($y=1$) in black or “bad cookie” ($y=0$) in red.

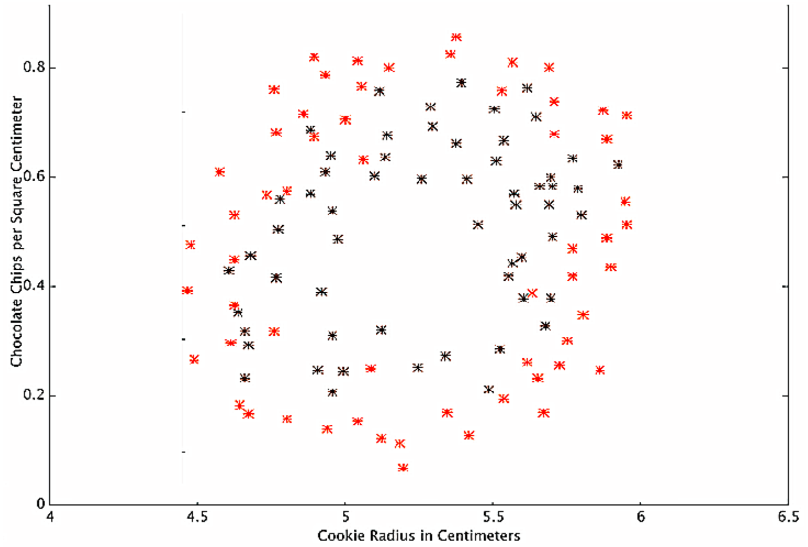

Fig.1 The results of a cookie quality testing study

A predictor that makes a guess somewhere between 0 and 1 is always needed. In a cookie quality classifier, a prediction of 1 represents a very confident guess that the cookie is good and utterly mouthwatering, while a prediction of 0 represents high confidence that the cookie is an embarrassment to the cookie industry. Values falling within this range refer to less confidence.

There’s a nice function that captures this behavior well, which is called the sigmoid function, g (z), which is shown in formula (1). The graph of this function is shown in Figure 2.

$$h(x)=g(z) \tag{1}$$

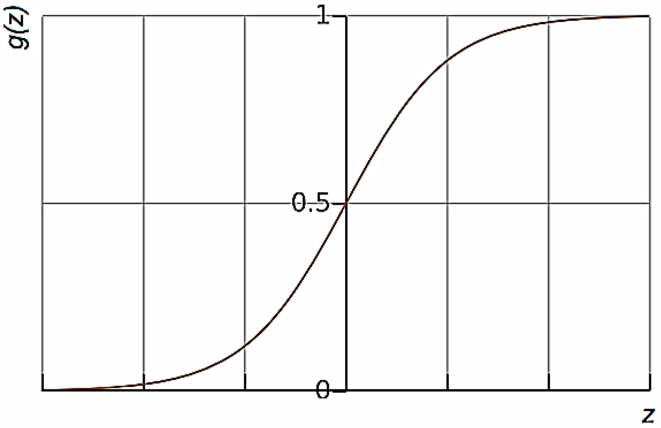

Fig.2 The graph of $h(x)=g(z)$

Where z represents our inputs and coefficients, such as $z = \theta_0 + \theta_1 x$. So the predictor can be written as $h(x) = g(\theta_0 + \theta_1 x)$.

The logic behind the design of the cost function is also different in classification problems. Again we ask "what does it mean for a guess to be wrong?" and this time a very good rule of thumb is that if the correct guess was 0 and we guessed 1, then we were totally wrong, and vice-versa. Alternatively, if the correct guess was 0 and we guessed 0, our cost function should not add any cost for each time when this happens. If the guess was right, but we weren't entirely confident, this should come with a small cost, and if our guess was wrong but we weren't entirely confident, this should come with some important cost, but not as much as if we were entirely wrong[1].

This behavior is captured by the log function, shown in formula (2)

$$\text{cost} = \begin{cases} -\log[1 - g(\theta_0 + \theta_1 x)] & \text{if } \ \text{y} = 0 \\ -\log[g(\theta_0 + \theta_1 x)] & \text{if } \ \text{y} = 1 \end{cases} \tag{2}$$

Again, the average cost over all of our training examples can be given by the cost function $J(\theta)$. Therefore, here we've described how the predictor $h(x)$ and the cost function $J(\theta)$ differ between regression and classification, however, gradient descent still works fine [1].

A classification predictor can be visualized by drawing the boundary line. With a well-designed system, cookie data can generate a classification boundary, shown in Figure 3:

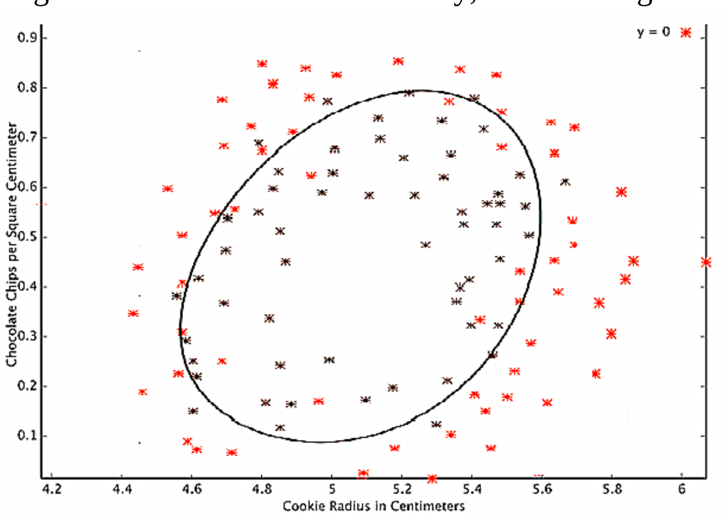

Fig.3 The results of cookie quality testing

3. Conclusion

Machine Learning is an incredibly powerful tool. In this paper, the procedure that how to solve classification problems in machine learning is acquired during solve cookie quality testing. In the coming years, Machine Learning promises to help solve some of our most pressing problems, as well as open up whole new worlds of opportunity.

References

[1] Nick McCrea. An Introduction to Machine Learning Theory and Its Applications: A Visual Tutorial with Examples. Software Engineer, September, 2015.

[2] Wernick, Yang, Brankov, Yourganov and Strother, Machine Learning in Medical Imaging, IEEE Signal Processing Magazine, vol. 27, no. 4, July 2010, pp. 25-38

[3] Friedman, Jerome H. Data Mining and Statistics: What's the connection. Computing Science and Statistics. 1998, 29 (1): 3-9.

[4] C.M. Bishop. Pattern Recognition and Machine Learning. Springer. 2006, ISBN 0-387-31073-8.