Introduction of SVM and NLP tokenization tools

Lingyun Yang
University of Nottingham
Psxly4@nottingham.ac.uk

ABSTRACT

This paper is an extended review of the research Using machine learning in a chatbox for heart disease prediction. In the research, two machine learning methods, k-Nearest Neighbors (KNN) and Support Vector Classifier are used to build for heart disease prediction. Another tool used in the research is NLTK. So as an extended review of the research, the topic of this paper includes two parts: optimize the parameters of SVM and introduce the tool of tokenization for natural language processing.

1. INTRODUCTION

Optimizing hyperparameters is vital in Machine learning. Building a machine learning model approximates a function that maps examples of inputs to examples of outputs. An approximation problem can be solved by framing it as an optimization problem. A machine learning algorithm defines the parameterized mapping function. (Brownlee, June 2, 2021). To discover the parameter values that will minimise a function's error, an optimization algorithm is applied to map inputs to outputs. Given this, We are solving an optimisation issue whenever we fit a machine-learning algorithm to a training dataset. The significant strengths of SVM are that the training is relatively easy. SVM has the advantage of being relatively simple to train. Contrary to neural networks, there is no local optimum. Additionally, classifier complexity and error may be tweaked manually. It scales reasonably well to high-dimensional data. That is also why the Support Vector Classifier is selected to build the model in the research.

Natural Language Processing (NLP) is a field of research and industry that examines the use of computers to interpret and handle natural language. To empower computers to interpret and handle natural languages, researchers are seeking to collect data on how people comprehend and manage language in order to build the right tools and strategies.(Chowdhary, 2020). The chatbot built in the research is a simple application based on natural language processing.

2. SVM AND OPTIMIZATION

One of the most widely used kernel learning algorithms is the Support Vector Machine (SVM). It uses well-established optimization theory concepts to achieve reasonably strong pattern recognition performance.(Awoke, 2012).

2.1. Kernel

What is kernel? In SVM, a kernel is a function that assists in problem solving. They offer shortcuts to complicated calculations. The beauty of kernel is that it enables people to go to higher dimensions and do smooth calculations. Kernels enable people to go up to an unlimited number of dimensions. The Gaussian RBF kernel has the following kernel parameter. Here is the function expression for the Gaussian kernel. The standard deviation is the width of the Gaussian distribution. With a greater sigma value, the decision boundary is likely to be flexible and smooth. The downside is it increases the probability of making the wrong decision, but it can also help prevent overfitting if it is used correctly. A smaller sigma, on the other hand, has a sharp decision boundary, so it has the disadvantage of overfitting.

$$k(x, x') = exp(-\frac{||x - x'||^2}{2\sigma^2})$$

2.2. Hard margin，soft margin and C

The dataset used in the project of chat box with machine learning have two classes with overlapping features. This section will talk about fit such dataset with hard-margin SVM and soft-margin SVM and the C parameter.

As shown in the Figure - 1 SVM margin and boundary below, it was initially assumed that the training data for the hard-margin SVM were linearly separable in mapped area. Furthermore, researchers seek to maximise the margin without sacrificing generality. However, unless the training instances are linearly separate, there is no solution. When data is linearly separable and no misclassifications are needed, SVM with a hard margin is a viable option. Alternatively, adopt soft margin SVM for the classifier if the linear boundary is not practical or if researchers wish to permit certain misclassifications for more generalisation.

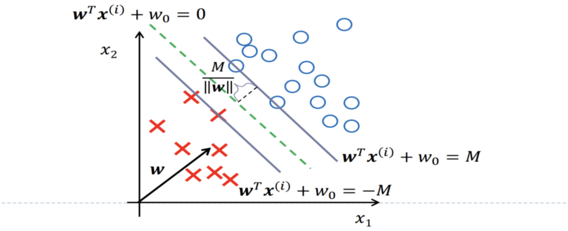

Figure – 1 SVM margin and boundary (Chen, 2021)

When the training set is not linearly separable, optimal hyperplanes are not useful (Bottou and Lin, 2007). Kernel machines are able to describe complex decision boundaries that accept any training set. However, this is impractical when the condition contains too many noisy. (Bottou and Lin, 2007). When dealing with noisy issues, it is appropriate to allow certain instances to break the margin limits of the primary problem. Slack variables ξ = (ξi . . . ξn)are used to represent these potential violations. In addition, parameter C governs the compromise between wide and slight margin breaches. (Bottou and Lin, 2007).

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \mathcal{P}(\mathbf{w},b,\boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \begin{cases} \forall i & y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ \forall i & \xi_i \geq 0 \end{cases}$$

Regarding what value of C must be set to get a hard-margin SVM, the objective of Slack variables is tolerating the violation of certain limitations. In other words, particular data points may be included in the margin. However, the number of points in the margin should be as little as feasible, as should the margin's distance.

2.3. Cross Validation

This section introduces the importance of nested cross-validation necessary and discusses inner and outer nested cross-validation?

It is not possible to train the machine learning model with all of the data. Because there will be no data to assess and evaluate the performance of the model. The most straightforward way to this problem is to split the entire data set into two halves. One for training and another for testing, known as the training set and test set, respectively. However, this method has a serious disadvantage. Performance and parameter selection of the model rely heavily on the distribution of training and test data. Therefore, if the method of dividing the training set and test set is inappropriate, researchers may cannot select the optimal model and parameters. The purpose of nested cross-validation is to find the best parameters of the model by estimating the generalisation error of the model. As discussed above, traditional validation approaches just split the data into a training dataset and a test dataset, which does not fix the problem of choosing the model parameters and selecting the best model(Jones et al., 2008).

Cross-validation nested loops have two types: outer and inner loops. Generally, the inner cross-validation is responsible for hyperparameter tuning, in other words, selecting the model's parameters. The outer loop takes charge of the generalisation error, which means error estimation. Random search, for instance, or grid search. Data from the outer loop is used to train the inner loop, and part of the data is retained to test the inner loop(Fix and Hodges, 1989). As a consequence, data leakage can be minimised and model score deviation can be kept to a minimum.

Cross-validation includes training a new model on a fraction of the entire data set and then evaluating it on the rest of the data. By repeating the process several times and averaging the validation error, researchers can estimate the generalisation performance of the model.

3. NLTK

3.1. Tokenization using NLTK

Why tokenize is important?

NLP(Natural Language Processing) problem is a multi-stage procedure issue (shubham.singh, July 18, 2019). Before even considering the modelling phase, researchers must cleanse unstructured text data. A few essential stages include in data cleansing are: Word tokenization, predicting sections of speech for each token, Text lemmatization, Identifying and eliminating stop words and so on (shubham.singh, July 18, 2019).

It is common procedure to tokenize text data while dealing with it. During tokenization, a phrase, sentence, paragraph, or full text is divided into smaller components, like words or terms.

Each of these smaller units is called a token. Various tokens can be used, including words, numbers, and punctuation marks. By identifying word boundaries, tokenization creates smaller units. Word boundaries are the points at which one word ends and another begins. As a first step towards stemming and lemmatization, these tokens are considered. It is important to identify the words in the natural language before processing them. Tokenization is, therefore, one of the most important steps in text data Natural language Processing. Analyzing the words in the text allows one to interpret the meaning of the text quickly. This method can be used for many different purposes. This tokenized form can be used by researchers to count the number of words in the text and to count how many times a word appears.

3.2. Tokenization tools

It contains a module called tokenize() that further categorizes it into two subcategories. NLTK, short for Natural Language ToolKit, is written in Python. The two sub methods are as below:

Tokenize a word: This method separates a sentence into words using the word_tokenize() method.

Sentence tokenization: This method breaks up a paragraph or document into sentences using the sent_tokenize() method.

The spaCy library can also be used to handle tokenization in NLTK. "spaCy" is a freely available NLP library. More than 49 languages are supported, and its computation speed is very fast. Keras is another useful framework for tokenization. Keras is one of the most popular frameworks in the world of deep learning. It is a Python-based open-source neural network library. Keras is simple to use and compatible with TensorFlow. Keras may be used to cleanse unstructured text data in the context of Natural Language Processing . Using the Gensim library is our last way of tokenization here. It is a library for natural language processing and unsupervised topic modelling. The library automatically derives semantic concepts from a text.

4. CONCLUSION

This paper introduced the technologies and methodologies used in the research: Using machine learning in a chatbox for heart disease prediction. For the SVM parameters. 'kernel' selects the type of hyperplane used to separate the data. In 2D data, 'linear' will use a line as the hyperplane, 'rbf' will use a non-linear hyperplane, and 'poly' will use a non-linear hyperplane (Karim et al., 2019). 'C' is tolerance of the violation of some constraints. For the NLTK part, four tokenize tools are introduced. Those are 'nltk.tokenize' package, spacy library, Keras, and Gensim library. These technologies and methods are not all used in my research, but they provide a lot of reference and inspiration in the process of project realization. At the same time, it also provides a very good accumulation of knowledge for future project development.

REFERENCES

AWOKE, G. 2012. Predicting HIV Infection Risk Factor using Volantory Counseling and Testing Data. *Addia Ababa University, Addis Ababa, Ethiopa*.

BOTTOU, L. & LIN, C.-J. 2007. Support vector machine solvers. *Large scale kernel machines,* 3**,** 301-320.

BROWNLEE, J. June 2, 2021. *Why Optimization Is Important in Machine Learning* [Online]. Optimization. Available: https://machinelearningmastery.com/category/optimization/ [Accessed].

CHEN, X. 2021. COMP3009 Machine Learning Course Materials. *In:* NOTTINGHAM, U. O. (ed.).

CHOWDHARY, K. 2020. Natural language processing. *Fundamentals of artificial intelligence***,** 603-649.

FIX, E. & HODGES, J. L. 1989. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique,* 57**,** 238-247.

JONES, S., MURPHY, F., EDWARDS, M. & JAMES, J. 2008. Doing things differently: advantages and disadvantages of web questionnaires. *Nurse researcher,* 15.

KARIM, A., MISHRA, A., NEWTON, M. H. & SATTAR, A. 2019. Efficient toxicity prediction via simple features using shallow neural networks and decision trees. *Acs Omega,* 4**,** 1874-1888.

SHUBHAM.SINGH. July 18, 2019. *How to Get Started with NLP – 6 Unique Methods to Perform Tokenization* [Online]. Analytics Vidhya. Available: https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/?cv=1 [Accessed 2022].