



Advanced machine learning methods in psychiatry: an introduction

Tsung-Chin Wu ¹, Zhirou Zhou,² Hongyue Wang,³ Bokai Wang ², Tuo Lin,⁴ Changyong Feng,² Xin M Tu^{5,6}

To cite: Wu T-C, Zhou Z, Wang H, *et al.* Advanced machine learning methods in psychiatry: an introduction. *General Psychiatry* 2020;**33**:e100197. doi:10.1136/gpsych-2020-100197

Received 21 January 2020
Accepted 03 February 2020

ABSTRACT

Mental health questions can be tackled through machine learning (ML) techniques. Apart from the two ML methods we introduced in our previous paper, we discuss two more advanced ML approaches in this paper: support vector machines and artificial neural networks. To illustrate how these ML methods have been employed in mental health, recent research applications in psychiatry were reported.

INTRODUCTION

Machine learning (ML) methods have been increasingly used in mental health and related research. In our last paper, we discussed two ML methods, the logistic regression and the k-means clustering.¹ In this report, we focus on two more advanced ML approaches, support vector machines (SVMs) and artificial neural networks (ANNs), along with their applications in psychiatry.

SVM is a supervised learning method to classify labelled outcomes. SVM applies a small number of samples called ‘support vectors’ from each class to build a classifier which separates samples into the different classes.² SVM is an extension of linear discriminant function, a popular statistical method of supervised learning, as it attempts to accommodate non-linear discriminant functions for more precise classification.³ SVM has been widely used, including in the field of psychiatry. For example, in studies on major depressive disorder (MDD), SVM was employed to identify patients with MDD from healthy controls by using demographic and clinical variables such as age, gender, education level, medication and so on.⁴ It is also a popular technique in neuroimaging.^{5,6} We will further discuss SVM in the next section.

ANN is composed of many simple units called ‘artificial neurons’. The main components of ANN are input layers, hidden layers and output layers. Computer scientists have developed ANNs to imitate biological neural networks by building models that mimic the learning process of human brains from training data without any prior knowledge of the data.⁷ For example, in

artificial intelligence, ANNs can be used to learn patterns such as numbers or letters from images (labelled data), then identify these numbers or letters from other images. There is no need to specify the patterns of these numbers or letters a priori or any other characteristics before learning. Thus, ANNs can be used in either supervised learning or unsupervised learning.⁷ We will focus on its usage in one type of supervised learning—classification—in this paper.

SUPPORT VECTOR MACHINE

SVM algorithms are classifiers defined by a particular linear decision boundary called ‘hyperplane’. For simplicity, we assume our data have two classes and are linearly separable, in which case a hyperplane that can perfectly separate the classes exists. In the specific case of two-dimensional feature space, a hyperplane becomes a straight line. Given training data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$, where n is the sample size, d is the number of predictors, and \mathbb{R}^d denotes the d -dimensional Euclidean space, the goal of SVM is to find a hyperplane that can split the data points into the two classes. Note that the hyperplane can be written as $w^T x + b = 0$, $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, so we want to find w and b such that $y_i (w^T x_i + b) > 0$ for all i . By scaling w and b , it is equivalent to finding w and b such that $y_i (w^T x_i + b) \geq 1$ for all i . There are infinitely many hyperplanes that can achieve our goal if the data are linearly separable (figure 1). To increase the possibility that the hyperplane is a ‘good’ classifier not only for the training data, we desire to have a hyperplane that can be as far as possible from the both classes, which means the best hyperplane is the one that lies ‘in the middle’ of the two classes (figure 2). In order to find such hyperplane, the problem becomes to maximise the distance from the hyperplane to the nearest data points in each class. It turns out that these nearest data points,



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Mathematics, University of California San Diego, La Jolla, California, USA

²Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, USA

³Departments of Biostatistics and Computational Biology and Anesthesiology, University of Rochester, Rochester, New York, USA

⁴Clinical and Translational Research Institute, University of California San Diego, San Diego, California, USA

⁵Family Medicine and Public Health, University of California San Diego, La Jolla, California, USA

⁶Naval Health Research Center, San Diego, California, USA

Correspondence to

Dr Xin M Tu; x2tu@ucsd.edu

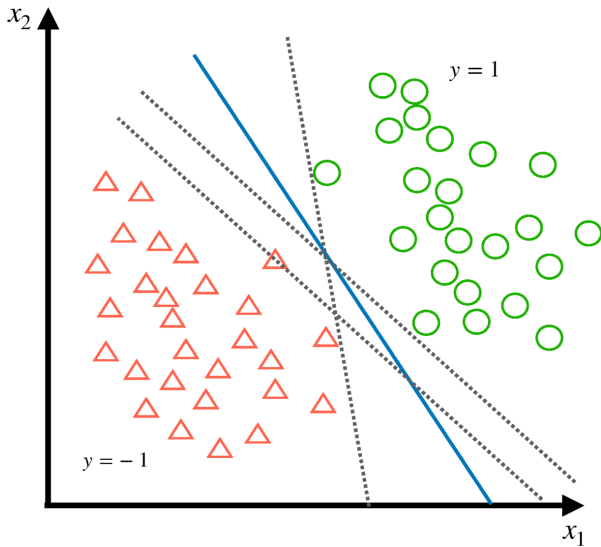


Figure 1 There are many hyperplanes that can split the data into two classes if the data are linearly separable.

or support vectors, can be found to completely determine such an optimal hyperplane.⁸

In practice, data are often not linearly separable. In this case, we can allow some points to lie on the wrong side of their ‘marginal hyperplane’—a hyperplane that is parallel to the optimal hyperplane and passed through support vectors. The region bounded by the two marginal hyperplanes is called the ‘margin’. We define the ‘slack variable’ of a data point as the distance of such point to its marginal hyperplane if the point lies on the wrong side of its marginal hyperplane, and zero if it lies on the correct side of its marginal hyperplane (figure 3). This approach may not work for classes that are truly non-linear. For example, if one class is inside and the other is outside a circle, this ad-hoc approach will not work. A qualitative improvement by SVM over the classic linear discriminant method is the ‘kernel trick’.⁸ In this case, we apply kernels, or mathematical transformations, to transform

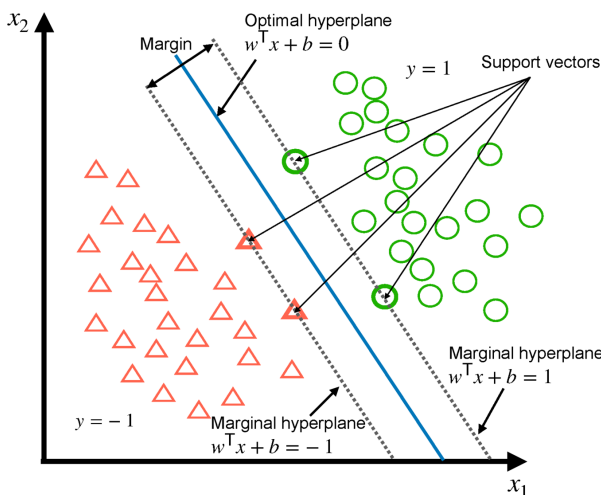


Figure 2 The linearly separable case where the optimal hyperplane separates the data into two-dimensional feature space.

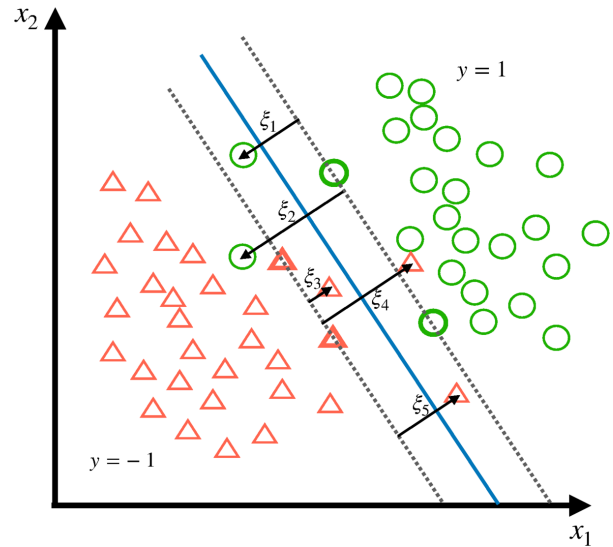


Figure 3 The non-linearly separable case where the optimal hyperplane separates the data into two-dimensional feature space. Note that $\xi_i, i = 1, \dots, 5$ are positive slack variables of the data.

data points observed in the study into data points in a higher dimensional Euclidean space and consider separation of classes by determining a hyperplane in such high dimensional space. There is no guarantee that there exists a kernel function to achieve perfect separation in every application. Even if such a kernel exists, finding it is no means straightforward. Nonetheless, SVM in theory offers new opportunities and more choices for better classifiers than its classic counterparts such as the linear discriminant function.

In a recent study, Mikolas *et al.*⁶ applied linear SVM to brain fractional anisotropy (FA) data consisting of 77 first episode of schizophrenia-spectrum disorder (FES) and 77 healthy controls (HC). They were interested in whether SVM would identify the FES from the HC based on the preprocessed skeletonised FA images. If we allow some data points to be misclassified by SVM (slack) and assume the label of FES is 1 and the control is -1, then the optimisation problem can be formulated as follows:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

Subject to $y_i (w^T x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i , where x_i is a vector of voxels of the i th image, $\xi = (\xi_1, \dots, \xi_n)$ is a vector of the slack variable for each point, and C is a parameter that controls the trade-off between the margin size and the total error $\sum_{i=1}^n \xi_i$. Note that, by definition, slack variables are always non-negative. If data are linearly separable, then $\xi = 0$, and the problem is the same as the linear separable case described above. For the parameter C , a larger C means a larger weight of slack, or a smaller margin, resulting in less misclassified points. A smaller C provides a larger margin and allows more points to be misclassified. To determine what C to use, one can use a grid search method and cross-validation to choose a C

that has the smallest cross-validation error.⁹ After estimating w and b , we determine a hyperplane for classification. Given $x^* \in \mathbb{R}^d$ as a new image data point, if $w^T x^* + b$ is positive, we classify the new image as FES, otherwise we classify it as control.

ARTIFICIAL NEURAL NETWORK

Like a human brain, ANN comprised artificial neurons. Artificial neurons receive inputs, process the information (compute a weighted sum of the inputs and maps it into a non-linear function called activation function) and produce an output.⁷ From the viewpoint of biology, the inputs correspond to the signal received by the dendrites, the processing of inputs corresponds to the soma, and the outputs correspond to the signal transmitted from the axon. The outputs will be the inputs of the next artificial neurons, thus forming a network. Activation functions can be viewed as rules of transmitting the signal. In the soma, when the electrical potential reaches a certain threshold, a signal pulse will be transmitted by the axon. Common activation functions include step function, sigmoid (logistic) function and rectified linear unit.¹⁰ When describing an ANN model, each artificial neuron can be considered as a node. If the output of an artificial neuron is the input of another artificial neuron, we can connect these two nodes by an arrow. The connections can be assigned different weights to represent different importance. There is no restriction on the number of connections. One node can have multiple inputs from other nodes and multiple outputs transmitted to others. The collection of artificial neurons that receive inputs from training data is called the input layer, which is the first layer of the ANN model. Similarly, the collection of neurons that produce final outputs is called the output layer, which is the final layer of the ANN model. The other neurons that do not belong to the above two layers are in hidden layers. The number of hidden layers depends on the problem we are trying to solve, therefore the structure of connections varies from case to case. The first and the simplest type of ANN is a feedforward neural network (FNN) in which neurons in each layer feed their outputs forward to the next layer until the final outputs.^{11 12} FNN only has one direction of information transmission without any feedback such as cycles or loops (figure 4).

A common class of FNN is the multilayer perceptron (MLP). The perceptron is defined as a function f called Heaviside step function that maps input $x \in \mathbb{R}^d$ to a binary output

$$f(x) = \begin{cases} 1, & \text{if } w^T x + b > 0 \\ -1, & \text{otherwise} \end{cases}$$

where $w \in \mathbb{R}^d$ is a vector of weights and $b \in \mathbb{R}$ is the bias.¹³ MLP is composed of many perceptions, but the function defined in the perceptron can be placed by other activation functions such as sigmoid, Gaussian and so on.¹⁴ Different activation functions allow MLP

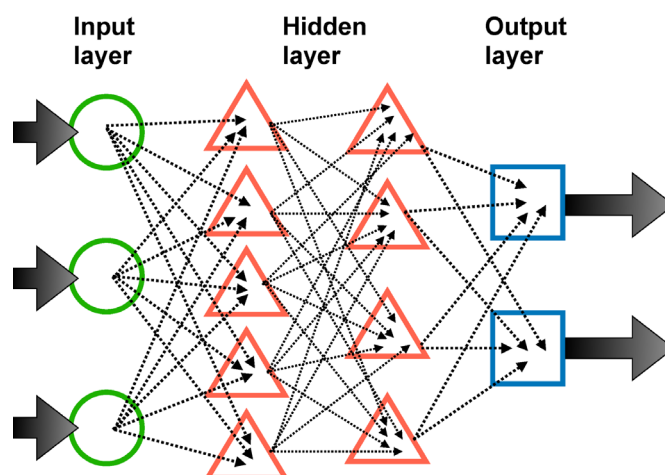


Figure 4 An example of a feedforward neural network.

to perform either classification or regression. In psychiatry, MLP can be applied to brain imaging data. Wang *et al.*¹⁵ used MLP to identify Alzheimer's disease from brain MRIs. MLP is typically trained by backpropagation (BP) algorithm. The goal of BP is to find w and b that minimise the loss function (eg, squared loss) by gradient methods such as gradient descent. BP computes the gradient of the loss function with respect to w and b by the chain rule, and updates them in each of the iteration until the error between the network output and the actual output is smaller than a preassigned threshold.¹⁶

DISCUSSION AND CONCLUSION

In this paper, we introduced two advanced ML approaches and their applications in mental health and related fields. Compared with logistic regression and k-means clustering, the final models of SVM and ANN are usually more difficult to understand and interpret, which is especially true for ANN, as it does not directly correspond to any statistical methods. However, these two methods work well with unstructured data such as images and texts, for which statistical methods are seldom used. As imaging analysis and text mining all have applications in mental health and related research, we believe that an understanding of the components and capabilities of these and other advanced ML methods will help to develop an appreciation of their existing and potential applications and apply them to improve our understanding and treatment of mental disorders.

Contributors T-CW and ZZ conceived of the presented idea. T-CW wrote the manuscript with support from ZZ, HW, BW and TL. CF and XMT helped supervise the project.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which

permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

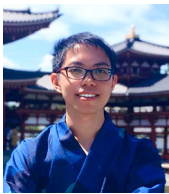
ORCID iDs

Tsung-Chin Wu <http://orcid.org/0000-0002-5224-6307>

Bokai Wang <http://orcid.org/0000-0002-2998-693X>

REFERENCES

- Zhou Z, Wu T-C, Wang B, et al. Machine learning methods in psychiatry: a brief introduction. *Gen Psychiatr* 2020;33:e100171.
- Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.
- Tu XM, Kowalski J, Randall J, et al. Generalized covariance-adjusted discriminants: perspective and application. *Biometrics* 1997;53:900–9.
- Wang J, Wei Q, Yuan X, et al. Local functional connectivity density is closely associated with the response of electroconvulsive therapy in major depressive disorder. *J Affect Disord* 2018;225:658–64.
- Orrù G, Pettersson-Yeo W, Marquand AF, et al. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev* 2012;36:1140–52.
- Mikolas P, Hlinka J, Skoch A, et al. Machine learning classification of first-episode schizophrenia spectrum disorders and controls using whole brain white matter fractional anisotropy. *BMC Psychiatry* 2018;18:97.
- Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal* 2000;22:717–27.
- Schölkopf B, Smola A. *Support vector machines*. Wiley StatsRef: Statistics Reference Online, 2014.
- Hsu CW, Chang CC, Lin CJ. *A practical guide to support vector classification*. Department of Computer Science, National Taiwan University, 2003.
- Zhang C, Woodland PC. Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling. *Proc Interspeech* 2015:3224–8.
- Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958;65:386–408.
- Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;61:85–117. ISSN 0893-6080.
- Panchal G, Ganatra A, Kosta YP, et al. Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *IJCTE* 2011;3:332–7.
- Shenouda E. A quantitative comparison of different MLP activation functions in classification. Lecture notes in computer science. *Advances in Neural Networks* 2006;3971:849–57.
- Wang S-H, Zhang Y, Li Y-J, et al. Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization. *Multimed Tools Appl* 2018;77:10393–417.
- Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, MA: MIT Press, 2016: 200–20.



Tsung-Chin Wu is a second year master's student studying statistics at University of California, San Diego (UCSD). He works at the Sam and Rose Stein Institute for Research on Aging in UCSD as a Graduate Student Researcher. He obtained both a Bachelor of Science degree in Mathematics and a Bachelor of Science degree in Horticulture and Landscape Architecture at National Taiwan University. Tsung-Chin has a strong interest in modern statistical applications for mental health data, and has applied many statistical methods for longitudinal data analysis, causal inference, and machine learning.