

Support Vector Machines for Binary Classification

“Support Vector Machines should be in the tool bag of every civilized person.”

– Patrick Henry Winston ¹

Introduction

Support Vector Machine (SVM) learning is a supervised learning technique that may be used as a binary classification system or to find a regression formula. Support vector machines are a “maximum margin classifier. SVM finds the separating hyperplane with the maximum margin to its closest data points.” ²

Although there are many mathematical approaches to solving SVM, many of which include kernels, let us first describe the Linear SVM used for binary classification. Consider a situation where we need to distinguish between two circumstances, or classes.

In an imaginary biochemistry laboratory, researchers discover a novel enzyme found in different tissues throughout the human body. Biochemists purify the enzyme from several cadavers and several different tissue types. Literature suggests that this newly found enzyme has two isozymes, 1) Alpha has a high reaction rate and 2) Beta has a lower reaction rate. It now seems like a simple task to learn which isozymes you possess. Carry out kinetic enzyme analysis on the purified samples then attempt to classify them.

Once you have carried out the kinetic analysis you then determine the Michaelis–Menten constant, K_M . The K_M constant is plotted on a single axis and produces the graphic below.

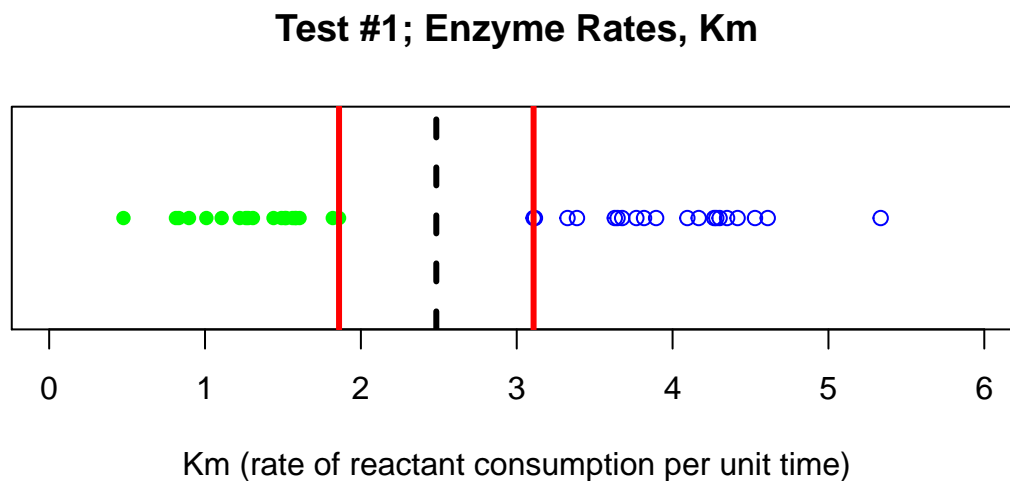


Figure 1: Is linearly separable data.

Linearly Separable

In test #1, we can see the two isozymes can easily be separated by activity alone. Figure 1 demonstrates that the data is *linearly separable*. The dataset is linearly separable if a single straight line can partition the data. In more general terms, “if the classes are linearly separable, it can be shown that the algorithm converges to a separating hyperplane.” ³ As Cortes and Vapnik indicate the hyperplane is the decision boundary of any

¹Patrick Henry Winston, 6.034 Artificial Intelligence, Fall 2010, Massachusetts Institute of Technology: MIT OpenCourseWare, <http://ocw.mit.edu/6-034F10>

²Nika Haghtalab & Thorsten Joachims, CS4780/5780 - Machine Learning for Intelligent Systems, Fall 2019, Cornell University, Department of Computer Science, <https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html>

³Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning; Data Mining, Inference, and Prediction, <https://web.stanford.edu/~hastie/ElemStatLearn/>, 2017

high dimension feature space, considering a hyperplane has one less dimension than its n-dimensional space. Incidentally, in Patrick Winston’s lecture on SVM, he calls SVM the “widest street approach.”⁴ Why does Professor Winston use this term? There are many possible streets which can be traced but the *goal* is to find the *widest street*. Many streets may be drawn in our example, but requiring the *widest street* leads to one. In fact, “an optimal hyperplane is here defined as the linear decision function with maximal margin between the vectors of the two classes.”⁵

SUPPORT-VECTOR NETWORKS

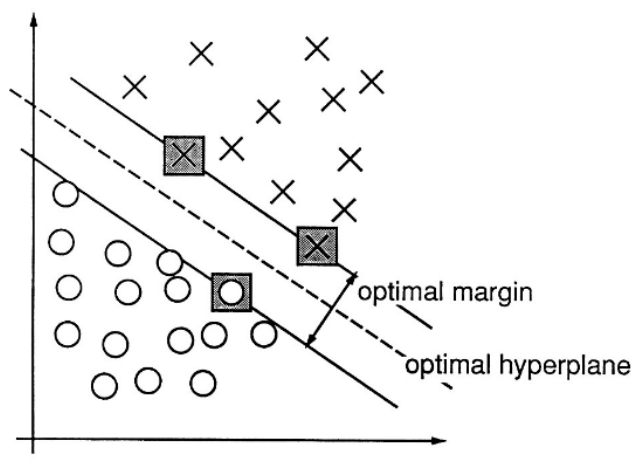


Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

6

Adding the prosaic phrase *widest street* smartly leads to the idea that a widest decision boundary also has the greatest ability to generalize.

Bias-Variance Trade-Off

The ability to generalize is a key idea in predictive modeling. This idea harkens back to freshman classes where one studied Student’s t-test and analysis of variance.

$$E \left[\left(y_0 - \hat{f}(x_0) \right)^2 \right] = Var(\hat{f}(x_0)) + \left[Bias(\hat{f}(x_0)) \right]^2 + Var(\epsilon) \quad (1)$$

The bias-variance dilemma can be stated as follows.⁷

- Models with too few parameters are inaccurate because of a large bias: they lack flexibility.
- Models with too many parameters are inaccurate because of a large variance: they are too sensitive to the sample details (changes in the details will produce huge variations).
- Identifying the best model requires controlling the “model complexity”, i.e., the proper architecture and number of parameters, to reach an appropriate compromise between bias and variance.

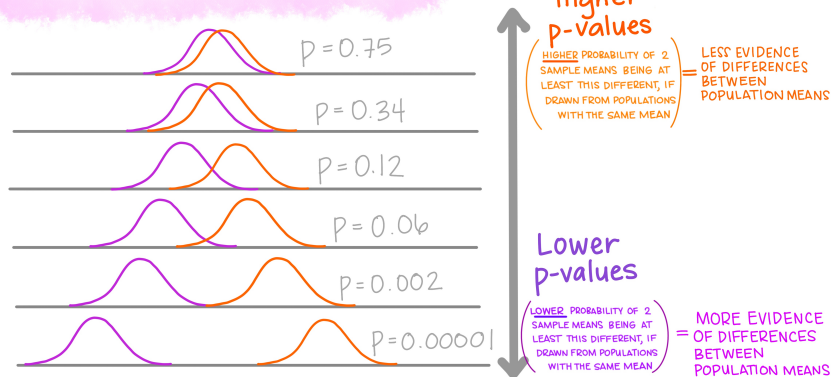
⁴Patrick Winston, 6.034 Artificial Intelligence, Fall 2010, Massachusetts Institute of Technology: MIT OpenCourseWare, <http://ocw.mit.edu/6-034F10>

⁵C. Cortes, V. Vapnik, Machine Learning, 20, 273-297, 1995

⁶C. Cortes, V. Vapnik, Machine Learning, 20, 273-297, 1995

⁷Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning; Data Mining, Inference, and Prediction, <https://web.stanford.edu/~hastie/ElemStatLearn/>, 2017

P-VALUES, SCHEMATICALLY:



8

However in real life, linearly separable data is rarely the case. Most often the activities are mixed as shown in test #2.

Test #2; Enzyme Rates, Km

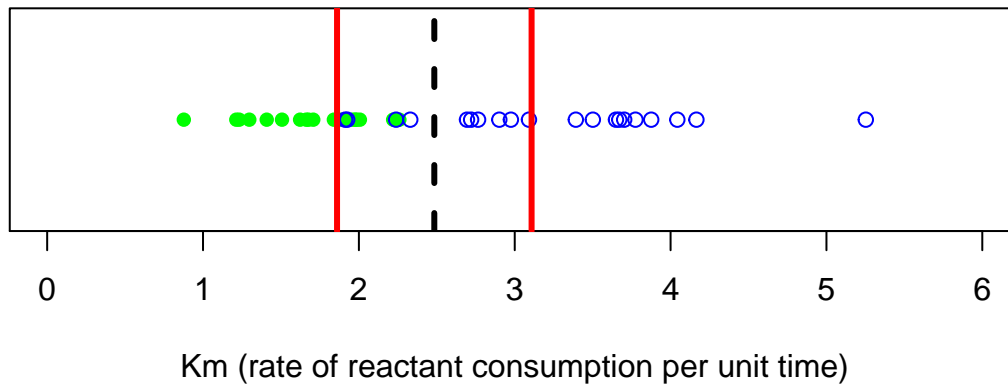


Figure 2: Is not linearly separable data.

Understanding the hyperplane equation

If we were trying to find:

$$\frac{1}{2} \widehat{W}(X_{\oplus} - X_{\ominus}) \quad (2)$$

Where $\widehat{W} = \left(\frac{x_1}{\|x\|}, \frac{x_2}{\|x\|} \right)$ and X_{\oplus} and X_{\ominus} represent data points that are labeled either positive or negative.

Suppose that X_{\oplus} and X_{\ominus} are equidistant from the decision boundary:

⁸Allison Horst, University of California, Santa Barbara, <https://github.com/allisonhorst/stats-illustrations>

Where a represents the region above the hyperplane;

$$W^T X_{\oplus} + b = a \quad (3)$$

and where $-a$ represents the region below the hyperplane or decision boundary.

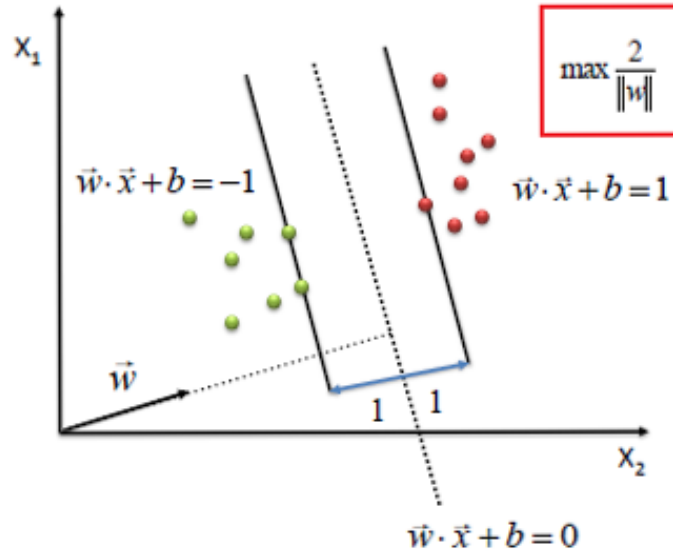
$$W^T X_{\ominus} + b = -a \quad (4)$$

Subtracting the two equations:

$$W^T (X_{\oplus} - X_{\ominus}) = 2a \quad (5)$$

Divide by the norm of w :

$$\widehat{W}^T (X_{\oplus} - X_{\ominus}) = \frac{2a}{\|W\|} \quad (6)$$



9

Soft Margins

In the case above, the activities overlap hence determining which isozyme is Alpha or Gamma is more difficult. In 1995, C. Cortes and V. Vapnik introduced the mathematics and ideas for “Soft Margins” or non-separable training data.¹⁰

The same is true of an n-dimensional system.

The first mention of an SVM like system is by Vapnik and Lerner in 1963, where the two described an implementation of a non-linear generalization called a Generalized Portrait algorithm.¹¹ As research has

⁹Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning; Data Mining, Inference, and Prediction, <https://web.stanford.edu/~hastie/ElemStatLearn/>, 2017

¹⁰C.Cortes, V.Vapnik, Machine Learning, 20, 273-297, 1995

¹¹V. Vapnik and A. Lerner, 1963. Pattern recognition using generalized portrait method. Automation and Remote Control, 24, 774–780

progressed, the types and complexity of SVM implementations have grown to encompass many circumstances. The ability of SVM to deal with different problems and handle different decision boundary shapes has made SVM a potent tool.

For example, this experiment has chosen to investigate three possible decision boundary shapes for the two-class protein data. The three mathematical constructs which will be tested are:

1. Linear hyperplane (also known as “plain-vanilla”),
2. Curvilinear or polynomial hyperplane and,
3. A radial basis function hyperplane,
4. Sigmoidal.

Linear: $K(x, y) = w^T x + b$

- The linear kernel does not transform the data at all.

Three common SVM kernel formulae investigated are:

Polynomial: $K(x_i, y) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

- The polynomial kernel has a straightforward non-linear transform of the data.
- Such that γ , r , and d are kernel parameters.

Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i^T - x_j\|^2), \gamma > 0$

- The Gaussian RBF kernel which performs well on many data and is a good default

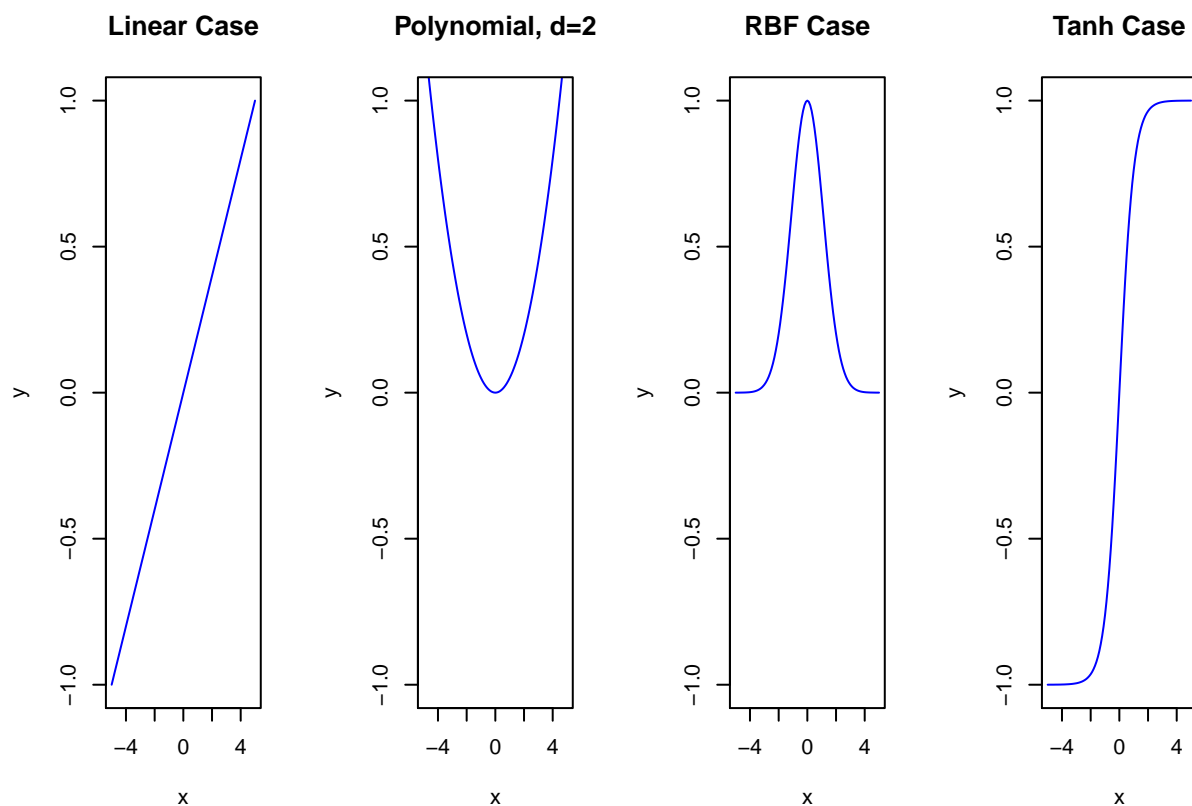
Sigmoidal: $K(x, y) = \tanh(\gamma x^T y + r), \gamma > 0$

- Incidentally, The sigmoid kernel produces an SVM analogous to the activation function similar to a [perceptron] with a sigmoid activation function.¹²

It is essential to note, at this time, there are no reliable rules for which kernel, i.e., boundary shape, to use with any given data set.

¹²(<https://data-flair.training/blogs/svm-kernel-functions/>)

Plots of 4 common SVM boundary shapes:



SVM-Linear

The simplest form of SVM utilizes a hyperplane as a separating element between the positive and control protein observations. This type of implementation is denoted as SVM-Linear (svm-lin) in this report. Here the mathematics is more easily described and can even be shown with a simple 2-dimensional graphic.

SUPPORT-VECTOR NETWORKS

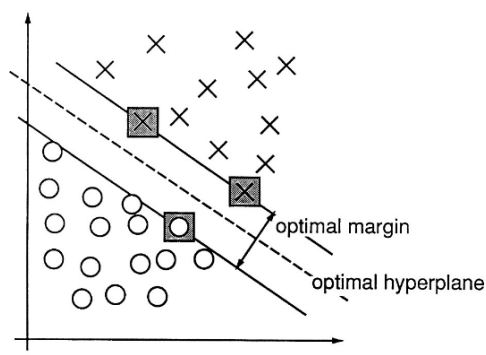


Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

Given a set of labeled pairs of data:

¹³Vladimir Vapnik & Corinna Cortes, Machine Learning, 20, 273-297, 1995

$$\{(X_1, y), \dots, (X_m, y)\}, \quad y \in \{1, -1\}, \quad \text{where } X^{m \times n} \in \mathbb{R} \quad (7)$$

For mathematical convenience, the labels are a set of values 1 or -1.

Therefore, we may write.

$$f(x_i) = \begin{cases} x \geq 0; y = 1 \\ x < 0; y = -1 \end{cases} \quad (8)$$

This is no different than is currently done in beginner level algebra. As is shown in the example below, the same is true for higher-dimensional problems.

Will be described and calculated in more detail in this report. However, there are alternative implementations of SVM.

In this experiment, three implementations of SVM have been used. The three are denoted as SVM-Linear (svm-lin), SVM-Polynomial (svm-poly), and SVM-Radial Basis Function (svm-rbf).

The switches in the R/caret software are easy such that one can use a number of kernels by changing the name of method..

with differing amounts of hyperparameters to modify. The intuition for the svm-poly and svm-rbf is also fairly straightforward. Instead of using a linear hyperplane to bisect the hi-dimensional space, which describes the decision boundary, the mathematics for a polynomial curvilinear function or a radial basis function may be utilized.

Yet another measurable difference that was investigated in this experiment was the use of a kernel transformation. It is conceivable to envision a hyperplane with no transformations utilized. Alternatively, the kernel transformations of original data can be used to increase the ability of the function to differentiate between positively and negatively labeled samples. A mathematical treatment can be found by Christopher Burges.¹⁴

As the usage of SVM grew, different issues presented problems for defining and coding the decision boundary were found. In the simplest case, the data points that sit along the support vector are nicely and neatly on the positive or the negative side. This is known as a hard margin which delineates the decision boundary. In reality, the decision boundary may include positive or negative datapoints that sporadically cross the boundary. In the circumstance where the decision boundary has similar points on either side, a penalty may be enlisted to deter the mathematics from choosing a boundary that includes too many misfit datapoints. In 1995, Support Vector Machines were described by Vladimir Vapnik and Corinna Cortes while at Bell Labs dealt with the soft-margin that occurs in the above situation.¹⁵

SVM is a non-parametric approach to regression and classification models.

What is Non-parametric?

For that matter, what is parametric learning and models. Just as we have learned that machine learning models can be supervised, unsupervised, or even semi-supervised another characteristic between machine learning models is whether they are parametric or not.

In Webster's dictionary ¹⁶ states a *parameter* is

- a. Estimation of values which enter into the equation representing the chosen relation
- b. [An] independent variable through functions of which other functions may be expressed -
Frank Yates, a 20th-century statistician

Another excellent explanation of this idea includes;

¹⁴Christopher Burges, Tutorial on Support Vector Machines for Pattern Recognition, D.M. & Knowl. Dis., 2, 121-167, 1998

¹⁵Vladimir Vapnik & Corinna Cortes, Machine Learning, 20, 273-297, 1995

¹⁶Webster's third new international dictionary, ISBN 0-87779-201-1, 1986

Does the model have a fixed number of parameters, or does the number of parameters grow with the amount of training data? The former is called a parametric model, and the latter is called a non-parametric model. Parametric models have the advantage of often being faster to use, but the disadvantage of making stronger assumptions about the nature of the data distributions. Non-parametric models are more flexible, but often computationally intractable for large datasets.¹⁷

Since Support Vector Machines are best described as a system where increasing the amount of training data, the numbers of parameters may grow as well. Therefore SVM is a non-parametric technique. Considering this idea in more detail, the estimation of the decision boundary does not entirely rely on the estimation of independent values (i.e., the values of the parameters). SVM is fascinating because the decision boundary may only rely on a small number of data points, otherwise known as support vectors.

In short, one guiding idea of SVM is a geometric one. In a binary-class learning system, the metric for the concept of the “best” classification function can be realized geometrically¹⁸ by using a line or a plane (more precisely called a hyperplane when discussing multi-dimensional datasets) to separate the two labeled groups. The hyperplane that separates the labeled sets is also known as a decision boundary.

This decision boundary can be described as having a hard or soft margin. As one might suspect, there are instances where the delineation between the labels is pronounced when this occurs decision boundary produces a hard margin. Alternatively, when the demarcation between the labeled groups is not so well defined by a straight and rigid line, the decision boundary provided is a soft margin. In either case, researchers have built up mathematics to deal with hard and soft margins. As an aside, the use of penalization is one method for coping with data points that impinge on the boundary hyperplane.

By introducing a “soft margin” instead of a hard boundary, we can add a slack variable ξ to account for the amount of a violation by the classifier, which later can be minimized.

In short, one guiding idea of SVM is a geometric one. In a binary-class learning system, the metric for the concept of the “best” classification function can be realized geometrically¹⁹ by using a line or a plane (more precisely called a hyperplane when discussing multi-dimensional datasets) to separate the two labeled groups. The hyperplane that separates the labeled sets is also known as a decision boundary.

Incidentally,

Big O notation ²⁰

Algorithm	Training	Prediction
SVM (Kernel)	$O(n^2p + n^3)$	$O(n_s v_p)$

Where p is the number of features, $n_s v_p$ is the number of support vectors

There are three properties that make SVMs attractive for data scientists:²¹

1. SVMs construct a maximum margin separator—a decision boundary with the largest possible distance to example points. This helps them generalize well.
2. SVMs create a linear separating hyperplane, but they have the ability to embed the data into a higher-dimensional space, using the so-called kernel trick. Often, data that are not linearly separable in the original input space are easily separable in the higher- dimensional space. The high-dimensional linear separator is actually nonlinear in the original space.

¹⁷Kevin P. Murphy, Machine Learning, A Probabilistic Perspective, MIT Press, ISBN 978-0-262-01802-9, 2012

¹⁸Xindong Wu, et al., Top 10 algorithms in data mining, Knowl Inf Syst, 14:1–37, DOI:10.1007/s10115-007-0114-2, 2008

¹⁹Xindong Wu, et al., Top 10 algorithms in data mining, Knowl Inf Syst, 14:1–37, DOI:10.1007/s10115-007-0114-2, 2008

²⁰Alexandros Karatzoglou, David Meyer, Kurt Hornik, ‘Support Vector Machines in R’, Journal of Statistical Software, April 2006, Volume 15, Issue 9.

²¹Stuart Russell and Peter Norvig, Artificial Intelligence, A Modern Approach, Third Edition, Pearson, ISBN-13: 978-0-13-604259-4, 2010

This means the hypothesis space is greatly expanded over methods that use strictly linear representations.

3. SVMs are a nonparametric method—they retain training examples and potentially need to store them all. On the other hand, in practice they often end up retaining only a small fraction of the number of examples—sometimes as few as a small constant times the number of dimensions. Thus SVMs combine the advantages of nonparametric and parametric models: they have the flexibility to represent complex functions, but they are resistant to overfitting.