Support Vector Machines for Binary Classification

> "Support Vector Machines ... are a very sophisticated idea that has a simple implementation, which should be in the tool bag of every civilized person." - Patrick Winston[1]

Introduction

Support vector machine (SVM) learning is a supervised learning technique which may be used as a binary classification system or to find regression formulae. It uses the conceptually simple idea that there is a decision boundary that can be determined from the multi-dimensional data that bisects the data (in the case of binary classification) into two portions of space using a hyperplane. Alternatively, when investigating SVM-regression the hyperplane becomes the regression plane that straddles the function which is being evaluated.

The first mention of a SVM-like system is by Vapnik and Lerner where the two described an implementation of a non-linear generalization called a Generalized Portrait algorithm.[2] As research has 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 very powerful 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 (sometimes denoted "vanilla"),
2. Curvilinear or polynomial hyperplane and,
3. A Radial basis function hyperplane.

Four common SVM kernel formulae investigated are:

1. Linear: $K(x_i,\ y_j) \ = \ < x,\ y >$
   - The linear kernel does not transform the data at all.
2. Polynomial: $K(x_i, x_j) \ = \ (\gamma\ x_i^T\ x_j \ + \ r)^d,\ \gamma > 0$
   - The polynomial kernel has a simple non-linear transform of the data.
   - Such that $\gamma,\ r$ and d are kernel parameters.
3. Radial Basis Function (RBF): $K(x_i, x_j) \ = \ exp(-\gamma \parallel x_i^T - x_j \parallel^2), \gamma > 0$
   - The Gaussian RBF kernel which performs well on many data and is a good default
4. Sigmoidal: $K(x_i, x_j) \ = \ \tanh(\gamma\ x_i^T\ x_j \ + \ r),\ \gamma > 0$
   - The sigmoid kernel produces a SVM analogous to the activation function similar to a perceptron with a sigmoid activation function.[3]

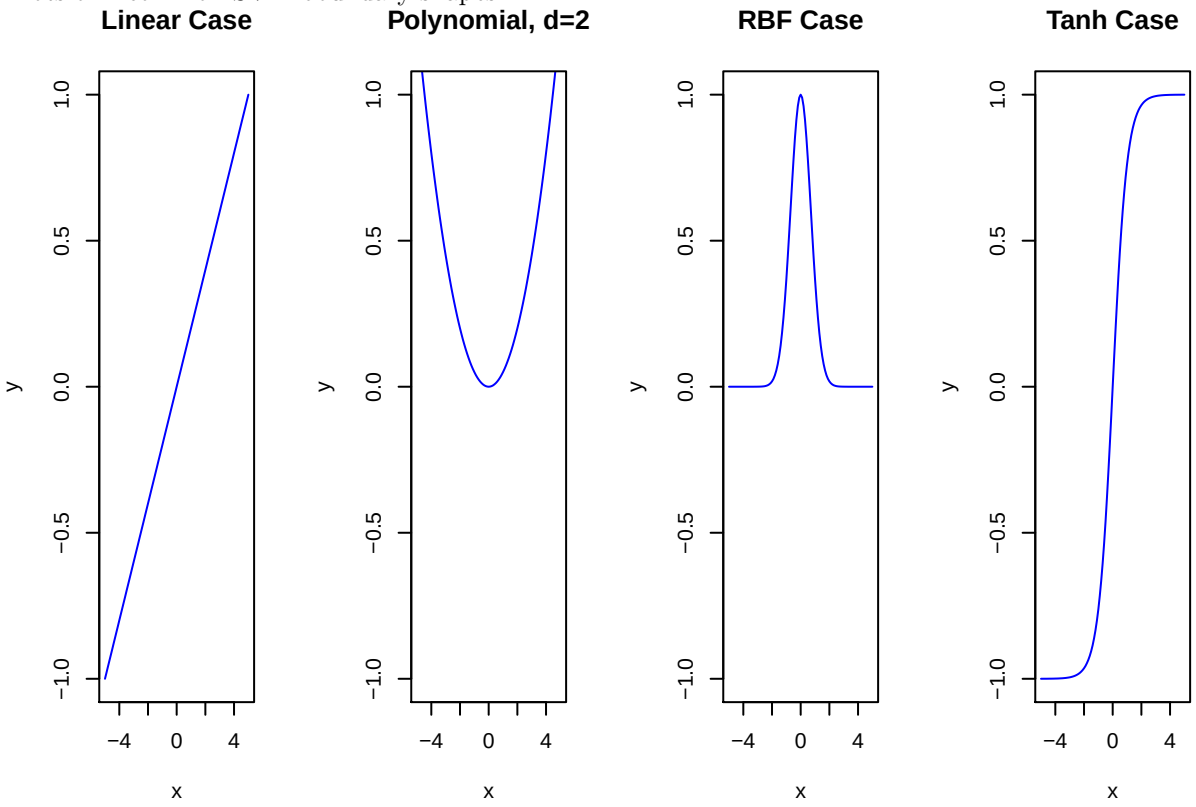
It is important to note, at this time, there are no reliable rules for which kernel, i.e. boundary shape, to use with any given data set.

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

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

[3] https://rpubs.com/mzc/mlwr_svm_concrete

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 are more easily described and can even be shown with a simple 2 dimensional graphic.

SUPPORT-VECTOR NETWORKS

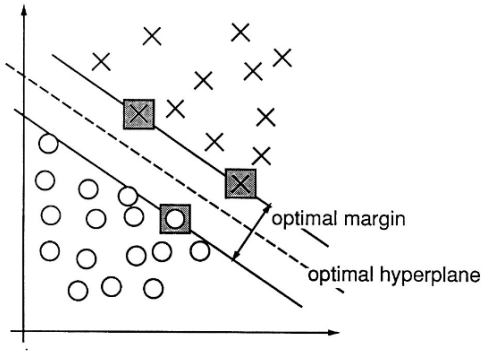


Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

Given a set of labeled pairs of data:

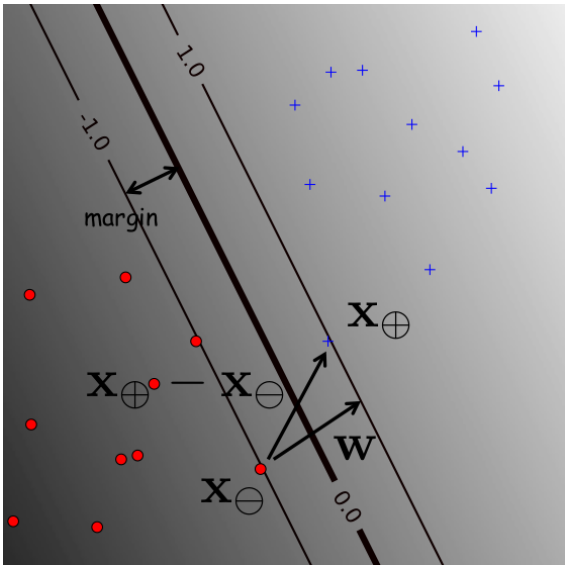
$$\{(x_1,\ y_1),\ ...,\ (x_m,\ y_m)\}, \quad y_i \in \{1,\ -1\}$$

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

Therefore, we may write

$$f(x_i) = \begin{cases} \geq 0; \ y_i = 1 \\ < 0; \ y_i = -1 \end{cases}$$

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 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 with differing amounts of hyperparameters to modify. The intuition for the SVM-poly and SVM-rbf are 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 mathematical 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.[4]

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

SUPPORT-VECTOR NETWORKS

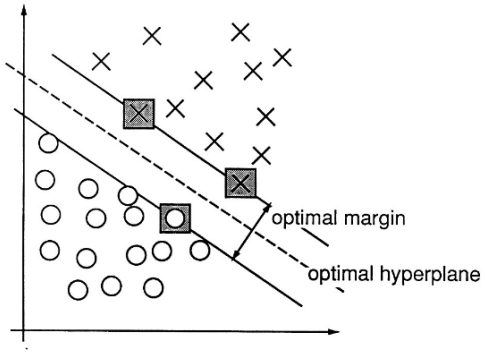


Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

As the usage of SVM grew different issues presented problems for defining and coding the decision boundary were found. In the simplest case, the datapoints 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.[5]

SUPPORT VECTOR MACHINES

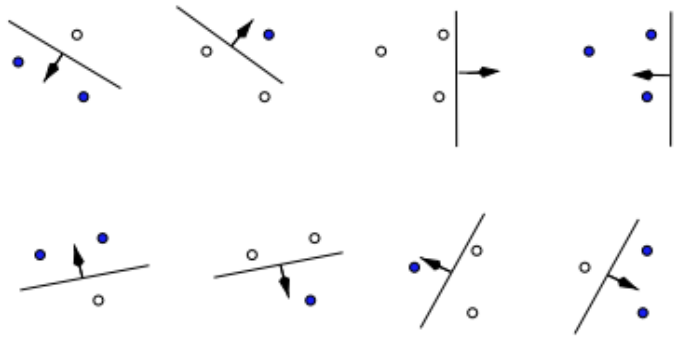


Figure 1. Three points in $\mathbf{R}^2$, shattered by oriented lines.

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 the Webster's dictionary[6] states a *parameter* is

a. Estimation of values which enter into the equation representing the chosen relation

[5]Vladimir Vapnik & Corinna Cortes, Machine Learning, 20, 273-297, 1995
[6]Webster's third new international dictionary, ISBN 0-87779-201-1, 1986

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

> 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.[7]

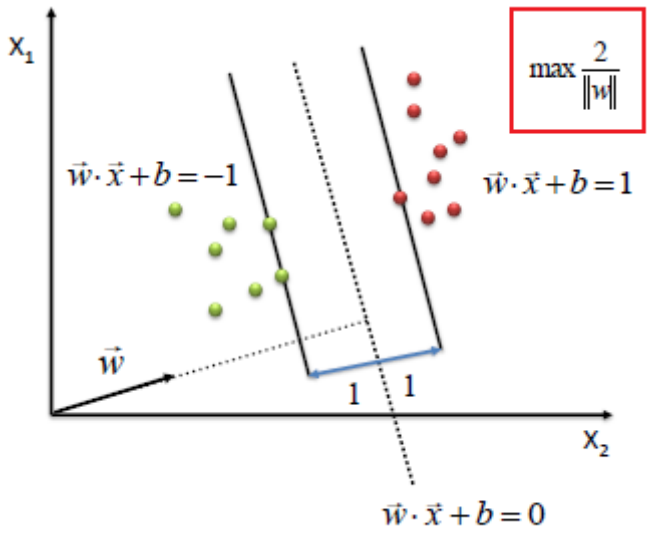
Since Support Vector Machines are best described as a system where increasing the amount of training data, the numbers of parameters may grow as well. Therefore SVM is a non-parametric technique. Considering this idea in more detail, the estimation of the decision boundary does not completely rely on the estimation of independent values (i.e. the values of the parameters). SVM is fascinating because the decision boundary may only rely a small number of datapoints 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 geometrically[8] 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 produced is a soft margin. In either case, researchers have built up the mathematics to deal with hard and soft margins. As an aside, the use of penalization is one method for dealing with datapoints that impinge on the boundary hyperplane.

Patrick Winston calls it the 'wide highway approach'. - See: Patrick Winston

By introducing a "soft margin" instead of a hard boundary we can introduce a slack variable xi to account for the amount of a violation by the classifier which later can be minimized.



If we were trying to find:

[7] Machine Learning, A Probabilistic Perspective, Kevin P. Murphy, MIT Press, ISBN 978-0-262-01802-9, 2012
[8] Xindong Wu, et al, Top 10 algorithms in data mining, Knowl Inf Syst, 14:1–37, DOI:10.1007/s10115-007-0114-2, 2008

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

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

$$W^T X_{\oplus} + b = a$$

$$W^T X_{\ominus} + b = -a$$

Subtracting the two equations:

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

Divide by the norm of w:

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

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 geometrically[ˆ6x] 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

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

The history of SVMs Large margin linear classifiers Vapnik, V., and A. Lerner. Pattern recognition using generalized portrait method. Automation and Remote Control, 24, 774–780, 1963.

Large margin non-linear classifiers B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Fifth Annual Workshop on Computational Learning Theory, pages 144—152, 1992

SVMs for non-separable data C. Cortes and V. N. Vapnik, Support vector networks. Machine Learning, vol. 20, no. 3, pp. 273-29

SVM {caret} / https://rpubs.com/PranovMishra/476455

- Tuning parameter C = cost for optimized model
- grid = expand.grid(C = seq(0.5, 10, 0.5)) USE: 2ˆ-10 to 2ˆ15 ????? SEE PAPER

for rbf: - Create grid control: sigma, C - tune.Grid = data.frame(expand.grid(sigma = seq(1, 10, 1)), - C = seq(1, 10, 1)) —

There are three properties that make SVMs attractive for data scientists:[9]

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

1. SVMs construct a maximum margin separator 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 examples25.