# Support Vector Machines

13

Learning Objectives

* Understand the concept of Support Vector Machines
* Learn the algorithm for SVM
* Know about the Kernel method that makes SVM possible
* Know the many advantages and disadvantages of SVMs

### INTRODUCTION

Support Vector Machine (SVM) is a mathematically rigorous, machine learn- ing technique to build a linear binary classifier. It creates a hyperplane in a high-dimensional space that can accurately slice a dataset into two segments according to the desired objective. The algorithms for developing the classifier can be mathematically challenging though. SVMs are popular since they are state-of-the-art for many practical problems, such as identifying spam emails and other text mining applications.

#### Caselet: Prostate Cancer Detection Using Protein Biomarkers

*Mass spectrometry technology is being used to generate a fast and efficient profile of the protein content of complex samples. It can generate data on up to 40,000 variables that constitute the profile for a protein sample. Data analytics can be used for narrowing the search space for protein biomarker candidates. These spectra can be used to identify peaks on certain biomarkers whose intensities correlate with a particular outcome variable, e.g., prostate cancer. Thorough cross-validation studies and randomization tests are performed on a prostate cancer dataset with over 300 patients, obtained at the Eastern Virginia Medical School using SELDI-TOF mass spectrometry. A two-stage linear SVM-based procedure led to average classification accuracies of 87 percent on a four-group classification problem, using only 13 variables (or peaks).*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| # of peaks used | | | | | | | |
| 10 | 15 | 20 | 25 | 30 | 35 | 50 | 70 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Quadr. | 74.7 | 74.7 | 74.1 | 74.7 78.2 (6.8) 77.8 (7.3) 78.7 (6.6) 76.8 (7.1) |
| Discr. | (7.4) | (9.6) | (8.4) | (7.1) |
| Nonpar | 76.7 | 77.4 | 77.7 | 78.6 80.0 (6.3) 79.9 (7.3) 78.1 (6.5) 76.1 (7.6) |
| (Kernel) | (7.1) | (8.4) | (6.9) | (6.6) |

(*Contd.*)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| kNN | 73.4 | 76.4 | 76.9 | 76.6 | 75.8 (6.7) 77.2 (6.9) 73.9 (7.5) 69.8 (6.7) |
|  | (7.4) | (6.9) | (6.0) | (6.1) |  |
| Fisher | 72.4 | 77.3 | 80.8 | 80.1 | 81.8 (6.0) 84.6 (5.2) 85.5 (6.1) 84.3 (5.1) |
| Linear | (7.3) | (6.9) | (6.5) | (5.8) |  |
| Linear | 75.4 | 79.3 | 81.7 | 81.3 | 83.7 (6.8) 83.1 (6.6) 83.5 (6.1) 84.0 (6.2) |
| SVM | (6.4) | (7.4) | (7.2) | (5.7) |  |

*(Source: *http://bmcbioinformatics.biomedcentral.com*)*

1. *What can be said about the accuracy of various techniques? Does the accuracy increase by increasing the number of variables used in the analysis?*
2. *As the predictive models get more accurate, what will be the role of doctors? What other skills would they need to learn?*

### SVM MODEL

An SVM is a classifier function in a high-dimensional space that defines the decision boundary between two classes. The support vectors are the data points that define the ‘gutters’ or the boundary condition on either side of the hyperplane, for each of the two classes. The SVM model is thus conceptually easy to understand.

Suppose there is a labeled set of points classified into two classes. The goal is to find the best classifier between the points of the two types.

X2



X1

FIGURE 13.1 Data Points for Classification

SVM takes the widest street (a vector) approach to demarcate the two classes and thus finds the hyperplane that has the widest margin, i.e., largest distance to the nearest training data points of either class (Figure 13.2).

In Figure 13.2, the hard-line is the optimal hyperplane. The dotted lines are the gutters on the sides of the two classes. The gap between the gutters is the maximum or widest margin. The classifier (hyperplane) is defined by only those points that fall on the gutters on both sides. These points are called the support vectors (shown in their bold shape). The rest of the data points in their class are irrelevant for defining the classifier (shown unfilled).

Abstractly, suppose that the training data of *n* points is

where represents the *p*-value vector for point and is its binary class value

of 1 or –1. Thus there are two classes represented as 1 and –1.

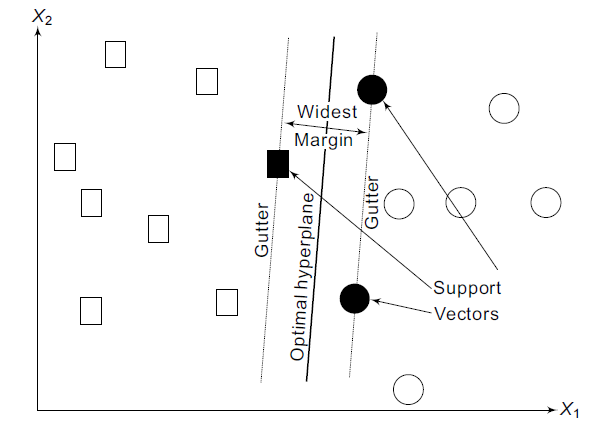


FIGURE 13.2 Support Vector Machine Classifier

Assuming that the data is indeed linearly separable, the classifier hyperplane is defined as a set of points (which is a subset of the training data) that satisfy the equation

where *W* is the normal vector to the hyperplane.

The hard margins can be defined by the following hyperplanes

and The width of the hard margin is ().

For all points not on the hyperplane, they will be safe in their class. Thus, the *y* values will have to be either greater than 1 (for a point in class 1) or less than –1 (for points in class –1).

The SVM algorithm finds the weights vector (*W*) for the features, such that there is the widest margin between the two categories.

Computing an SVM using these equations is a hill-climbing process problem in a convex space. However, by working with points nearest to the classifying boundary only, it reduces sharply the number of data instances to work with. This approach reduces its memory requirements for computation. This is possible because of using kernel methods.

### THE KERNEL METHOD

The heart of an SVM algorithm is the kernel method. Most kernel algorithms are based on optimization in a convex space and are statistically well-founded.

Kernel stands for the core or the germ in a fruit. Kernel methods operate using what is called the ‘kernel trick’. This trick involves computing and working with the inner products of only the relevant pairs of data in the feature space; they do not need to compute all the data in a high-dimensional feature space. The kernel trick makes the algorithm much less demanding in computational and memory resources.

Kernel methods achieve this by learning from instances. They do not apply some standard computational logic to all the features of each input. Instead, they re-member each training example and associate a weight representing its relevance to the achievement of the objective. This could be called instance-based learning. There are several types of support vector models including linear, polynomial, RBF, and sigmoid.

It is somewhat akin to how human beings learn, especially about domains with broad sets of features, such as ethics. We learn lessons from tough situations and discard the ‘normal’ instances. In kernel parlance, we assign high weights to abnormal situations and very low weight to normal situations. The normal situations are almost completely forgotten.

SVMs have evolved to be more flexible and be able to tolerate some amount of misclassification. The margin of separation between the categories is thus a ‘soft margin’ as against a hard margin.

### ADVANTAGES AND DISADVANTAGES OF SVMs

The main strength of SVMs is that they work well even when the number of features is much larger than the number of instances. It can work on datasets with huge feature space, such is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.

Another advantage of SVMs is that even when the optimal decision boundary is a nonlinear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data.

SVMs are conceptually easy to understand. They create an easy-to-understand linear classifier. By working on only a subset of relevant data, they are computationally efficient. SVMs are now available with almost all data analytics toolsets.

The SVM technique has two major constraints – (a) It works well only with real numbers,i.e., all the data points in all the dimensions must be defined by numeric values only, (b) It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.

Training the SVMs is an inefficient and time-consuming process when the data is large. It does not work well when there is much noise in the data, and thus has to compute soft margins. The SVMs will also not provide a probability estimate of classification, i.e., the confidence level for classifying an instance.

Conclusion

Support Vector Machine is a machine learning technique for classifying high-dimensional data into two classes. It creates a hyperplane with the largest amount of separation between the two classes. The classifier is made linear by transforming the original features of input data into new features. SVMs use kernel methods to learn from specific instances that are close to the decision boundary. SVMs are used for text mining, such as spam filtering and outlier detection.

## Questions

1. What is a Support Vector Machine?
2. What are the support vectors?
3. Explain the kernel method.
4. Name two advantages and two limitations of SVMs.

## True/False

1. SVM is a statistical learning technique.
2. SVMs are popular for classifying text documents.
3. SVMs use the entire dataset to develop the classifier.
4. Kernel methods use a limited amount of data to facilitate faster processing.
5. SVMs can create a linear or curvilinear classifier.
6. SVMs can handle partial misclassifications.