DS M4

techworldthink • February 20, 2022

Support vector machines

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine

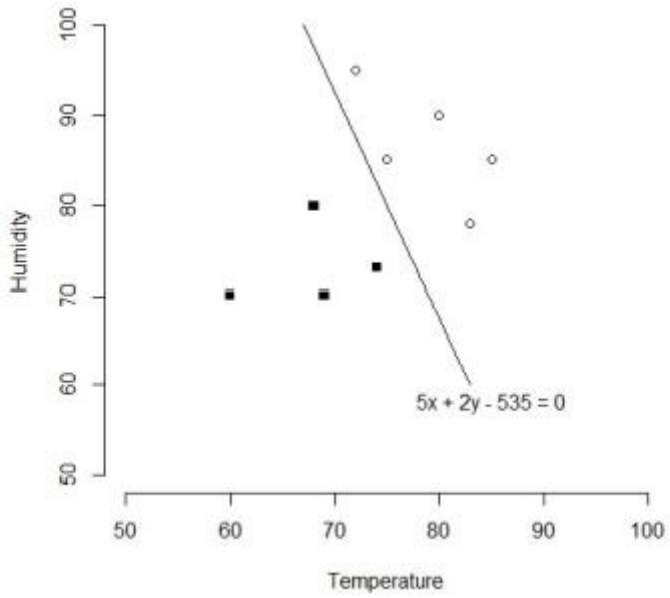
SVM can be of two types:

- Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

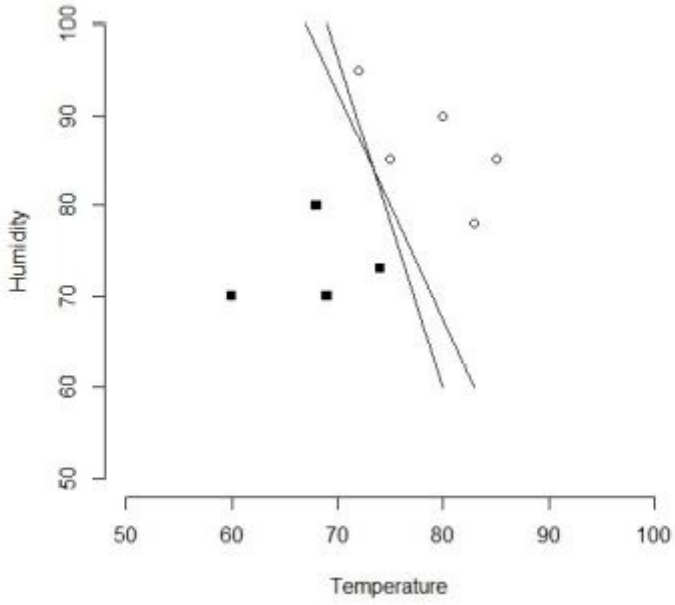
Two-class data set When there are only two class labels the data is said to be a “two-class data set”.

Scatter plot of the data

A separating line



Several separating lines

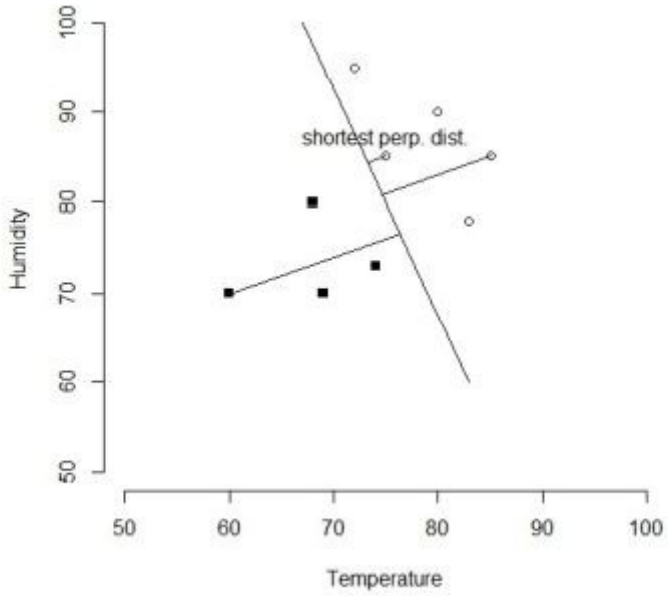


Margin of a separating line

To choose the “best” separating line, we introduce the concept of the margin of a separating line. Given a separating line for the data, we consider the perpendicular distances of the data points from the separating line. The double of the shortest perpendicular distance is called the “margin of the separating line”.

Maximal margin separating line

The “best” separating line is the one with the maximum margin



The separating line with the maximum margin is called the “maximum margin line” or the “optimal separating line”. This line is also called the “support vector machine”.

Support vectors

The data points which are closest to the maximum margin line are called the “support vectors” and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

The required criterion

As per theory of support vector machines, the equation of the maximum margin line is used to devise a criterion for taking a decision.

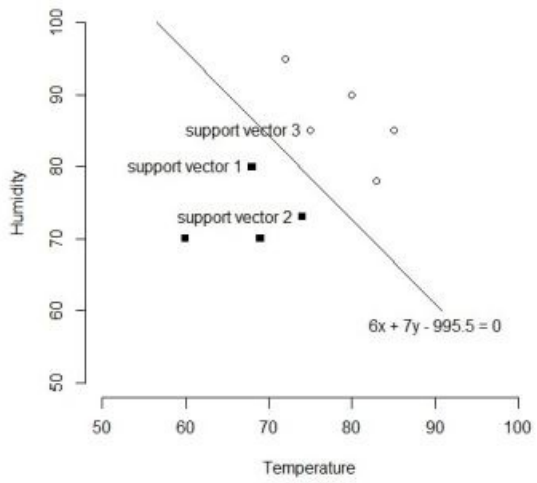


Figure 10.6: Support vectors for data in Table 10.1

of temperature and humidity on a given day. Then the decision as to whether play tennis on that day is "yes" if

$$7x + 6y - 995.5 < 0$$

and "no" if

$$7x + 6y - 995.5 > 0.$$

Hyperplanes

Hyperplanes are certain subsets of finite dimensional vector spaces which are similar to straight lines in planes and planes in three-dimensional spaces.

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Consider the n-dimensional vector space R n .

The set of all vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ in R^n whose components satisfy an equation of the form

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

where $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n$ are scalars, is called a hyperplane in the vector space R^n.

Let $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, then using the notation of inner product, Eq.(10.5) can be written in the following form:

$$\alpha_0 + \vec{\alpha} \cdot \vec{x} = 0$$

Two-class data sets

In a machine learning problem, the variable being predicted is called the output variable, the target variable, the dependent variable or the response. A two-class data set is a data set in which the target variable takes only one of two possible values only. If the target variable takes more than two possible values, the data set is called a multi-class dataset.

Linearly separable data

data set is linearly separable if we can find a hyperplane in the n-dimensional vector space R n.

ata set with two class labels is linearly separable, then, in general, there will be several separating hyperplanes for the data set

Given a two-class data set, there is no simple method for determining whether the data set is linearly separable. One of the efficient ways for doing this is to apply the methods of linear programming.

Maximal margin hyperplanes

Consider a linearly separable data set having two class labels “−1” and “+1”. Consider a separating hyperplane H for the data set

1. Consider the perpendicular distances from the training instances to the separating hyperplane H and consider the smallest such perpendicular distance. The double of this smallest distance is called the margin of the separating hyperplane H.

2. The hyperplane for which the margin is the largest is called the maximal margin hyperplane (also called maximum margin hyperplane) or the optimal separating hyperplane.

3. The maximal margin hyperplane is also called the support vector machine for the data set.

4. The data points that lie closest to the maximal margin hyperplane are called the support vectors.

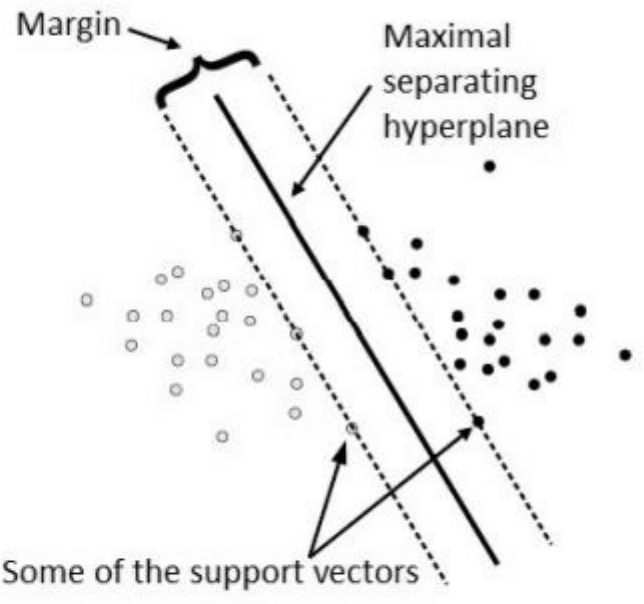


Figure 10.12: Maximal separating hyperplane, margin and support vectors

Kernel functions

In the context of SVM's, a kernel function is a function of the form $K(x, \vec{y})$, where $\vec{x}$ and $\vec{y}$ are n-dimensional vectors, having a special property. These functions are used to obtain SVM-like classifiers for two-class datasets which are not linearly separable.

Let $\vec{x}$ and $\vec{y}$ be arbitrary vectors in the n-dimensional vector space R^n. Let φ be a mapping from R^n to some vector space. A function $K(x, \vec{y})$ is called a kernel

function if there is a function φ such that

$K(\vec{x}, \vec{y}) = \varphi(\vec{x}) \cdot \varphi(\vec{y})$.

Example 1

Let

$$\vec{x} = (x_1, x_2) \in \mathbb{R}^2$$
$$\vec{y} = (y_1, y_2) \in \mathbb{R}^2$$

We define

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y})^2.$$

We show that this is a kernel function. To do this, we note that

$$\begin{aligned} K(\vec{x}, \vec{y}) &= (\vec{x} \cdot \vec{y})^2 \\ &= (x_1 y_1 + x_2 y_2)^2 \\ &= x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2 \end{aligned}$$

Now we define

$$\phi(\vec{x}) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \in \mathbb{R}^3$$
$$\phi(\vec{y}) = (y_1^2, \sqrt{2} y_1 y_2, y_2^2) \in \mathbb{R}^3$$

Then we have

$$\begin{aligned} \phi(\vec{x}) \cdot \phi(\vec{y}) &= x_1^2 y_1^2 + (\sqrt{2} x_1 x_2)(\sqrt{2} y_1 y_2) + x_2^2 y_2^2 \\ &= x_1^2 y_1^2 + 2 x_1 x_2 y_1 y_2 + x_2^2 y_2^2 \\ &= K(\vec{x}, \vec{y}) \end{aligned}$$

This shows that $K(\vec{x}, \vec{y})$ is indeed a kernel function.

10.10.3 Some important kernel functions

In the following we assume that $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$.

1. **Homogeneous polynomial kernel**

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y})^d$$

where d is some positive integer.

2. **Non-homogeneous polynomial kernel**

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + \theta)^d$$

where d is some positive integer and θ is a real constant.

3. **Radial basis function (RBF) kernel**

$$K(\vec{x}, \vec{y}) = e^{-\|\vec{x}-\vec{y}\|^2/2\sigma^2}$$

This is also called the Gaussian radial function kernel.[1]

4. **Laplacian kernel function**

$$K(\vec{x}, \vec{y}) = e^{-\|\vec{x}-\vec{y}\|/\sigma}$$

5. **Hyperbolic tangent kernel function (Sigmoid kernel function)**

$$K(\vec{x}, \vec{y}) = \tanh(\alpha(\vec{x} \cdot \vec{y}) + c)$$

10.11 The kernel method (kernel trick)

10.11.1 Outline

1. Choose an appropriate kernel function $K(\vec{x}, \vec{y})$.
2. Formulate and solve the optimization problem obtained by replacing each inner product $\vec{x} \cdot \vec{y}$ by $K(\vec{x}, \vec{y})$ in the SVM optimization problem.
3. In the formulation of the classifier function for the SVM problem using the inner products of unclassified data $\vec{z}$ and input vectors $\vec{x}_i$, replace each inner product $\vec{z} \cdot \vec{x}_i$ with $K(\vec{z}, \vec{x}_i)$ to obtain the new classifier function.