Support Vector Machines

Carmel Gafa

March 4, 2021

Abstract

This paper is the first in a series of articles that illustrate various algorithms that can be used in budget prediction scenarios. The objectives of these articles is to provide the necessary background and implementation knowledge to the persons involved in these challenges. It is the aim of these papers to be as exhaustive as possible; the required mathematical, implementation and practical issues are all discussed in length.

1 Introduction

In this first article, prediction based on support vector machines is investigated. Although support vector machines is widely known as a classification algorithm, as we shall see late on, it can be easily adapted top predict the most likely outcome in a series. TODO: history This paper is divided into XXXXXX sections. First, the required mathematical concepts are discussed so to provide the necessary tools to examine support vector machines. A description of the concepts making up state vector machines can be found in Section 3. TODO: other sections

2 Mathematical Background

The concepts necessary to understand the support vector machines algorithm in detail are discussed here so they can serve as a reference in the latter sections.

2.1 Maximum and minimum values of a function of two variables

A function $z = f(x, y)$ is said to have a maximum value at $P(a, b)$ if $f(a, b)$ is greater than the value at the near-by point $Q(a + \delta x, b + \delta y)$ for all values of δx and δy however small, positive or negative, that is in all directions from P.

Similarly, $z = f(x, y)$ is said to have a minimum value at $P(a, b)$ is $f(a, b)$ is less than the value at the near-by point $Q(a + \delta x, b + \delta y)$ for all values of δx and δy however small, positive or negative, that is in all directions from P.

To determine minimum and maximum values we must therefore investigate the sign of the value of $f(a+\delta x, b+\delta y) - f(a,b)$

If $f(a+\delta x, b+\delta y) - f(a,b) < 0$ then we have a maximum value at $P(a,b)$
If $f(a+\delta x, b+\delta y) - f(a,b) > 0$ then we have a minimum value at $P(a,b)$

Now Taylor's theorem expands a function of two variables $f(x+\delta x, y+\delta y)$ in terms of $f(x,y)$, powers of δx and δy and successive derivatives of $f(x,y)$ and can be stated as

$$f(x+\delta x, y+\delta y) = f(x,y) + \left\{\delta x f'_x(x,y) + \delta y f'_y(x,y)\right\} + \frac{1}{2!}\left\{\delta x^2 f''_{xx}(x,y) + 2\delta x \delta y f''_{yx}(x,y) + \delta y^2 f''_{yy}(x,y)\right\} + \ldots \tag{1}$$

Hence, in the case examined here,

$$f(a+\delta x, b+\delta y) = f(a,b) + \left\{\delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y}\right\} + \frac{1}{2!}\left\{\delta x^2 \frac{\partial^2 f}{\partial x^2} + 2\delta x \delta y \frac{\partial^2 f}{\partial x \partial y} + \delta y^2 \frac{\partial^2 f}{\partial y^2}\right\} + \ldots$$

therefore,

$$f(a+\delta x, b+\delta y) - f(a,b) = \left\{\delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y}\right\} + \frac{1}{2!}\left\{\delta x^2 \frac{\partial^2 f}{\partial x^2} + 2\delta x \delta y \frac{\partial^2 f}{\partial x \partial y} + \delta y^2 \frac{\partial^2 f}{\partial y^2}\right\} + \ldots$$

For very small δx and δy higher order derivatives become negligible and the equation can be approximated by:

$$f(a+\delta x, b+\delta y) - f(a,b) = \left\{\delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y}\right\}$$

For a stationary value of z at (a,b)

$$f(a+\delta x, b+\delta y) - f(a,b) = \left\{\delta x \frac{\partial f}{\partial x} + \delta y \frac{\partial f}{\partial y}\right\} = 0$$

Since δx and δy are small independent increments,

$$\delta x \frac{\partial f}{\partial x} = 0$$

and

$$\delta y \frac{\partial f}{\partial y} = 0$$

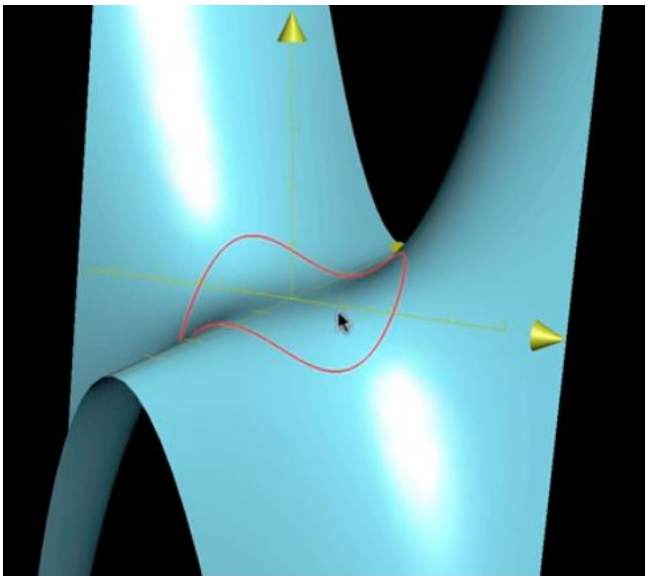

Figure 1: A multivariate function subject to a constraint

2.2 Lagrange Multipliers

In many practical situations it is often required to determine the points at which stationary values occur, with the added condition that the points lie on a pre-described curve. This is referred to as a constrained optimization problem. In essence, the objective of this technique is to maximize (or minimize) some multivariate function that is however subject to a certain constraint.

Figure 1 shows, as an example the function $f(x, y) = x^2y$ subject yo a unit circle constraint $x^2 + y^2 = 1$. We notice that the projection of the circle on the function has a number of peaks (and troughs) and this techniques is used to determine the maximum peak. An alternative way to visualize this function is to consider the x-y plane as in Figure 2 and visualize contour lines for the $f(x, y)$, whilst clearly visualizing the constraint as the unit circle. Different values of $f(x, y)$ will produce different contours, some will intersect the constraint, others will lie outside. Of interest here is the contour that is tangent to the constraint as that is the maximum value of $f(x, y)$ given the stated constraint.

So,if we, as an example, have to determine stationary points of the function

$$f(x, y)$$

with variables x and y constrained by the relation

$$\phi(x, y) = 0 \tag{2}$$

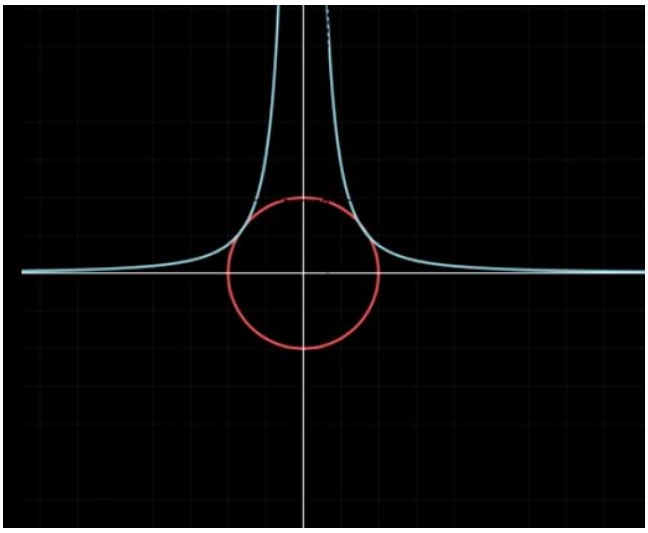

Figure 2: X-Y plane projection of the system in Figure 1

As we saw in the previous section the total differential reduced to zero at the stationary points,

$$\frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y = 0 \tag{3}$$

Also, since $\phi(x, y) = 0$

$$\frac{\partial \phi}{\partial x}\delta x + \frac{\partial \phi}{\partial y}\delta y = 0 \tag{4}$$

if we multiply each term in (4) be a multiplier $-\lambda$ and then add (4) and (3) we get

$$\left(\frac{\partial f}{\partial x} - \lambda\frac{\partial \phi}{\partial x}\right)\delta x + \left(\frac{\partial f}{\partial y} - \lambda\frac{\partial \phi}{\partial y}\right)\delta y = 0$$

since δx and δy are independent increments.

$$\frac{\partial f}{\partial x} - \lambda\frac{\partial \phi}{\partial x} = 0 \tag{5}$$

$$\frac{\partial f}{\partial y} - \lambda\frac{\partial \phi}{\partial y} = 0 \tag{6}$$

or

$$\frac{\partial f}{\partial x} = \lambda\frac{\partial \phi}{\partial x} \tag{7}$$

$$\frac{\partial f}{\partial y} = \lambda \frac{\partial \phi}{\partial y} \tag{8}$$

or in matrix form

$$\begin{bmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \end{bmatrix} = \lambda \begin{bmatrix} \frac{\partial \phi}{\partial x} \\ \\ \frac{\partial \phi}{\partial y} \end{bmatrix}$$

These two equations together with the constraint equation (2) that can be used to determine the values of x and y at the stationary points and if required, the value of λ.

Graphically this means, that at the maximum point, the gradient of $f(x_m, y_m)$ is proportional to the gradient of $\phi(x_m, y_m)$ and the constant of proportionality is the Lagrange multiplier λ, hence,

$$\bigtriangledown f(x_m, y_m) = \lambda \bigtriangledown \phi(x_m, y_m) \tag{9}$$

In the contour plot, the gradient at any point is perpendicular to the function, hence one can visualize equation (9) in Figure 3

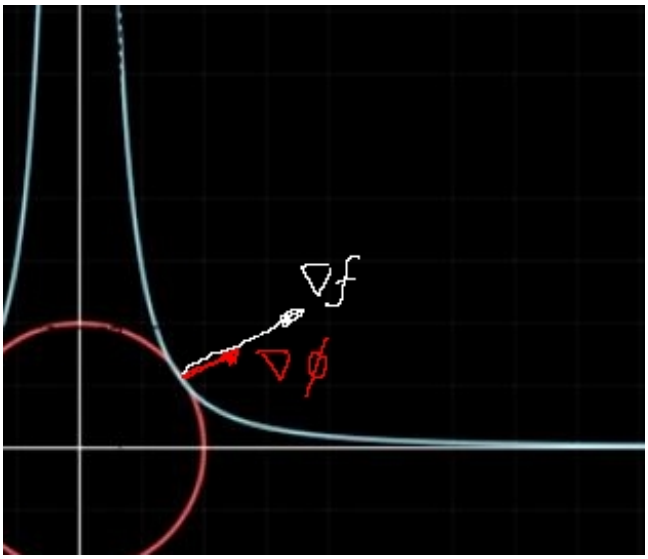

Figure 3: Graphical representation of the Lagrange multiplier

2.3 The Lagrangian

As we have seen previously, given a function to be maximized $f(x, y)$ constrained by $\phi(x, y) = b$, the maximum will be achieved when the contour of f is just tangent to that of ϕ. Therefore the gradient of f, $\bigtriangledown f$ is proportional to that of

ϕ, $\bigtriangledown\phi$. The two gradients can be equated by the introduction of the Lagrange multiplier,

$$\bigtriangledown f(x,y) = \lambda \bigtriangledown \phi(x,y) \tag{10}$$

The Lagrangian is defined as:

$$\mathcal{L}(x,y,\lambda) = f(x,y) - \lambda\left(\phi(x,y) - b\right) \tag{11}$$

where b is a constant.
Of particular importance is the value of $\bigtriangledown\mathcal{L} = \mathbf{0}$ as

$$\begin{bmatrix} \frac{\delta\mathcal{L}}{\delta x} \\ \frac{\delta\mathcal{L}}{\delta y} \\ \frac{\delta\mathcal{L}}{\delta\lambda} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

If we consider

$$\frac{\delta\mathcal{L}}{\delta x} = 0$$

This can be evaluated to

$$\frac{\delta f}{\delta x} - \lambda\left(\frac{\delta\phi}{\delta x}\right) = 0$$

or

$$\frac{\delta f}{\delta x} = \lambda\left(\frac{\delta\phi}{\delta x}\right)$$

which is equation (7). Similarly

$$\frac{\delta\mathcal{L}}{\delta y} = 0$$

can be evaluated to

$$\frac{\delta f}{\delta y} - \lambda\left(\frac{\delta\phi}{\delta y}\right) = 0$$

or

$$\frac{\delta f}{\delta y} = \lambda\left(\frac{\delta\phi}{\delta y}\right)$$

which is equation (8).
Finally,

$$\frac{\delta\mathcal{L}}{\delta\lambda} = 0$$

can be evaluated to

$$-(\phi(x,y)-b)=0$$

or

$$\phi(x,y)=b$$

which is the constraint equation.

3 Support Vector Machines

Decision boundary techniques separate data points into regions, or classes to which they belong.

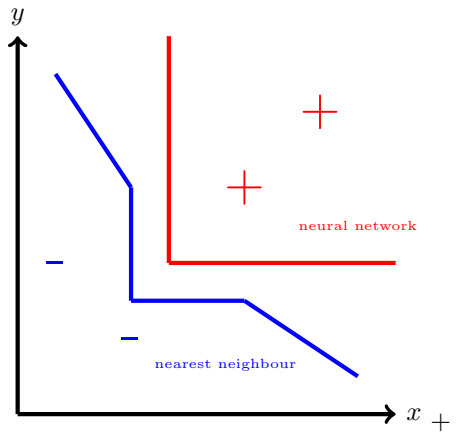

Figure 4: Different decision boundary techniques

The claim of support vector machines is that given a space with negative samples and positive samples a divisor can be constructed with the view towards putting in the widest boundary that separates the positive samples from the negative ones; thus maximizing the separation between the classes.

This analysis is started with a vector $\vec{w}$ of unknown length constrained to be perpendicular to the divisor. An unknown sample u is somewhere on the plane with a vector $\vec{u}$ to it. Of interest is whether u is on the left or right of the divisor. To determine this, $\vec{u}$ is projected on $\vec{w}$ as the distance of the projected vector will determine whether the divisor has been crossed. This can be determined if we verify that this projection is larger or equal to some constant c, thus to imply a positive sample,

$$\vec{w}.\vec{u} \geq c$$

if we define $b=-c$, we can say, without loss of generality that

$$\vec{w}.\vec{u}+b \geq 0 \tag{12}$$

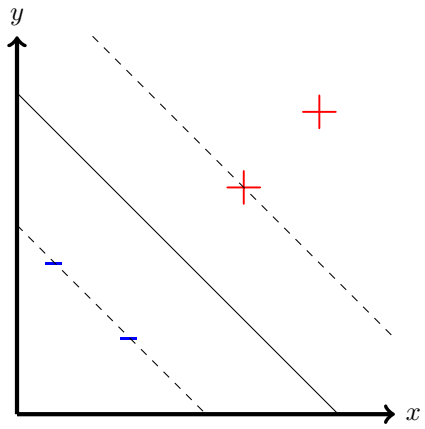

Figure 5: SVM decision boundary.

This is the decision rule. We reiterate that b and $\vec{w}$ are not known, and the only information available is that $\vec{w}$ is perpendicular to the divisor. Equation(12) defines the divisor hyperplane.

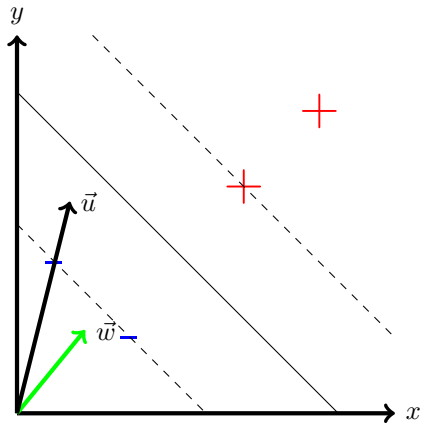

Figure 6: Unknown sample vector.

If we take $\vec{w}$ and project a positive sample onto it

$$\vec{w}.\vec{x}_+ + b \geq 1 \tag{13}$$

likewise

$$\vec{w}.\vec{x}_- + b \leq -1 \tag{14}$$

Introducing variable y_i such that

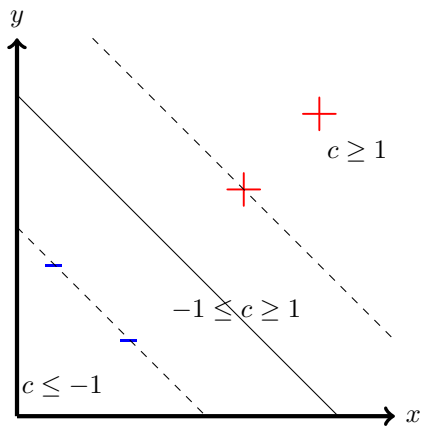

Figure 7: Values of c on the plane.

y_i +1 for positive samples
y_i -1 for negative samples

so equation (13) becomes

$$\vec{w}.\vec{x}_i + b \geq -1$$
$$(\vec{w}.\vec{x}_i + b) - 1 \geq 0$$
$$y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 0$$

for positive samples, and equation (14) becomes

$$\vec{w}.\vec{x}_i + b \leq -1$$
$$(\vec{w}.\vec{x}_i + b) + 1 \leq 0$$
$$y_i(\vec{w}.\vec{x}_i + b) + 1 \leq 0$$
$$y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 0$$

for negative samples.
Therefore we can conclude that

$$y_i(\vec{w}.\vec{x}_i + b) - 1 \geq 0 \tag{15}$$

we also define an additional constraint that

$$y_i(\vec{w}.\vec{x}_i + b) - 1 = 0 \tag{16}$$

for the values exactly on the boundary lines. Vectors to these values are termed as support vectors.

Equations (15) and ((16)) define the constraints for this system.

It is of particular interest to determine the distance between the boundaries. Consider a positive sample x_+ on the positive boundary with a vector $\vec{x}_+$ to it. Similarly, consider a negative sample x_- on the negative boundary with a vector $\vec{x}_-$ to it. The width of the boundary can be found if we find the dot product of $x_+ - x_-$ with a unit vector normal to the boundaries ($\frac{w}{||w||}$). Therefore the width, which is the scalar distance can be found by:

$$width = (\vec{x}_+ - \vec{x}_-).\frac{\vec{w}}{|| \vec{w} ||} \tag{17}$$

but samples on the boundary are constrained by

$$y_i(\vec{w}.\vec{x}_i + b) - 1 = 0$$

where $y_i = 1$ on the boundary. In this case, for the positive sample

$$\vec{w}.\vec{x}_+ = 1 - b$$

and for the negative sample

$$-1(-\vec{w}.\vec{x}_- + b) = -1$$
$$-\vec{w}.\vec{x}_- - b - 1 = 0$$
$$\vec{w}.\vec{x}_- = 1 - b$$

substituting in equation(17)

$$width = \frac{(1-b) - (1+b)}{|| \vec{w} ||} = \frac{2}{|| \vec{w} ||} \tag{18}$$

The objective is to maximize the width, or maximize $\frac{2}{||\vec{w}||}$ or maximize $\frac{1}{||\vec{w}||}$ or minimize $|| \ w \ ||$ or minimize $\frac{1}{2} \ || \ \vec{w} \ ||^2$ under the constraints stated in equation(16) and (15).

If we use Lagrange multipliers,

$$\mathcal{L}(w, b) = \frac{1}{2} \ || \ \vec{w} \ ||^2 - \sum_i \lambda_i y_i \left[(\vec{w}_i \vec{x}_i + b) - 1 \right] \tag{19}$$

In this case, most constraint multipliers, λ_i will be zero unless they are connected with vectors that lie on the boundary lines. Since we are dealing with $|| \ \vec{w} \ ||^2$ this type of optimization problem is referred to as a quadratic optimization one. One should note that the objective of this exercise is to minimize w and maximize b, keeping in mind that the equation of the divisor hyperplane is $(\vec{w}.\vec{x}) + b$.

From Lagrange Multiplier theory, the extremes can be found by using $\bigtriangledown \mathcal{L} = 0$.

$$\frac{\partial \mathcal{L}}{\partial \vec{w}} = \vec{w} - \sum_i \lambda_i y_i \vec{x}_i = 0$$

hence

$$\vec{w} = \sum_i \lambda_i y_i \vec{x}_i \tag{20}$$

It can be concluded from this equation that vector $\vec{w}$ is the linear sum of some of samples in the set, since for the other samples the value of λ_i is zero, as stated before.

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_i \lambda_i y_i = 0$$

that is

$$\sum_i \lambda_i y_i = 0 \tag{21}$$

Substituting equation(20) in (19) we get

$$\mathcal{L} = \frac{1}{2} \sum_i \lambda_i y_i \vec{x}_i \sum_j \lambda_j y_j \vec{x}_j - \sum_i \lambda_i y_i \vec{x}_i . \sum_j \lambda_j y_j \vec{x}_j - \sum_i \lambda_i y_i b + \sum_i \lambda_i$$

Consider the term $\sum_i \lambda_i y_i b$. This can be rewritten as $b \sum_i \lambda_i y_i$ since b is a constant. Form equation (21), this term is 0.

Therefore,

$$\mathcal{L} = \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \vec{x}_i . \vec{x}_j \tag{22}$$

The maximum of this equation depends only on the dot product of pairs of samples. The decision rule, equation (12) becomes;

$$\sum_i \lambda_i y_i \vec{x}_i . \vec{u} + b \geq 0 \implies positive \tag{23}$$

The decision rule depends on the dot product of the unknown and the other samples.

If the samples are not linearly separable, a different perspective is taken by applying a transformation, $\phi(x)$

As was discussed previously, the maximization depends on the dot product of the samples. Hence, by applying the mentioned transformation, maximization now depends on

$$\phi(\vec{x}_i) . \phi(\vec{x}_j)$$

Using the same reasoning the decision becomes dependent on

$$\phi(\vec{x}_i) . \phi(\vec{u})$$

Hence the function will be necessary,

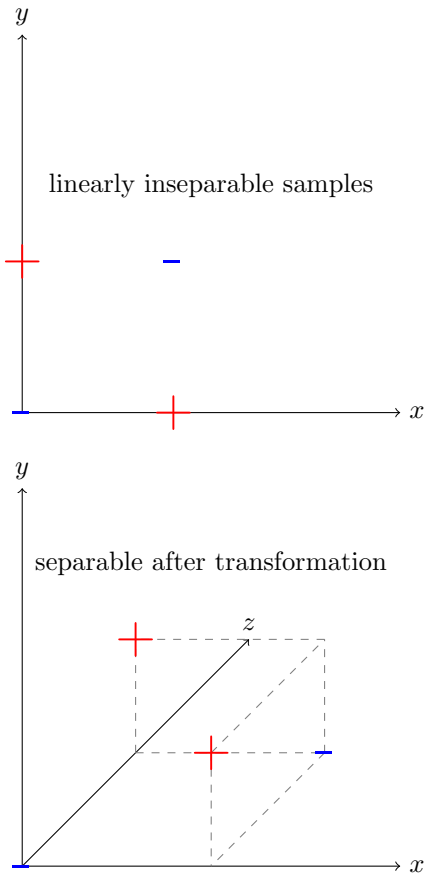

Figure 8: Resolution of linearly inseparable samples.

$$K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i).\phi(\vec{x}_j) \tag{24}$$

Function K is called the kernel and provides the dot product of the two vectors in another space.

3.0.1 Linear Kernel

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i.\vec{x}_j + 1)^n \tag{25}$$

3.0.2 Radial Basis

$$K(\vec{x}_i, \vec{x}_j) = e^{(\frac{||\vec{x}_i - \vec{x}_j||}{\sigma})} \tag{26}$$

If σ is very small over-fitting will take place