Support Vector Machine: Complete Theory of Support Vectors

Suppose we're given these two samples of blue stars and purple hearts (just for schematic representation and no real data are used here), and our job is to find out a line that separates them best. What do we mean by best here ?

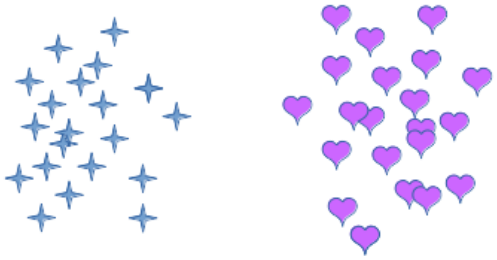

Figure 1: Samples in a 2D plane with some separation between them

Let's see the image below. Could you guess which line would separate the two samples better?

Figure 2: Lines (Hyperplanes?) that could potentially separate the two samples

Yes, the red line on the left is better than the orange line because, we say that the red line creates the '**widest road'** (margin) between the two groups. See the image below

Figure 3: **Widest road** approach for separating two groups

The samples on the edge of the boundary lines (dotted) lines, are known as '**Support Vectors'.** On the left side there are two such samples (blue stars), compared to the one on the right. Few important points about Support vectors are-

1. *Support Vectors are the samples that are most difficult to classify.*
2. *They directly affect the process to find the optimum location of the decision boundaries (dotted lines).*
3. *Only a very small subset of training samples (Support vectors) can fully specify the decision function (We will see this in more detail once we learn the math behind SVM).*
4. ***If the Support Vectors are removed from the data set, it will potentially change the position of the dividing line** (in case of space with higher dimensional space, the line is called Hyperplane).*

We have come far enough to appreciate that this is a **constrained optimization** problem. *Optimization* — because, we are to find the line from which the support vectors are maximally separated and *Constrained* — because, the support vectors should be away from the road and not on the road. We will use **Lagrange Multipliers** to solve this problem, so let's start with a very simple example of using Lagrange multiplier .

Lagrange Multipliers: When and how to use

Suppose we are given a function *f(x,y,z,…)* for which we want to find extrema, subject to the condition *g(x,y,z,…)=k*. The idea used in Lagrange multiplier is that the gradient of the objective function *f*, lines up either in parallel or anti-parallel direction to the gradient of the constraint *g*, at an optimal point. In such case, one the gradients should be some multiple of another. Let's see using an example —

Consider a problem of obtaining extrema of a function $f(x, y) = 8x^2 - 2y$ under the constraint $x^2 + y^2 = 1$.

Using Lagrange multiplier we solve it the following way

$$\nabla f = \lambda \nabla g\,;\;\; g(x,y) = \left(x^2 + y^2\right) \tag{1.1}$$

Equating component wise

$$16x = 2\lambda x;\;\; x(16 - 2\lambda) = 0 \tag{1.2}$$
$$-2 = 2\lambda y;\;\; y = -\frac{1}{\lambda}$$

So from first part we get either $x = 0$, or $\lambda = 8$; Using constraint $x^2 + y^2 = 1$, we get $y = \pm 1$; If $\lambda = 8$; then

$$x^2 + \frac{1}{\lambda^2} = 1;\;\; x = \pm\frac{3\sqrt{7}}{8}; \text{ for } \lambda = 8\ . \tag{1.3}$$

Finally we can write the solutions as $\left(-\frac{3\sqrt{7}}{8}, -\frac{1}{8}\right)$, $\left(\frac{3\sqrt{7}}{8}, -\frac{1}{8}\right)$, $(0,1)$, $(0,-1)$. The function values at these points are $f\left(\pm\frac{3\sqrt{7}}{8}, -\frac{1}{8}\right) = 8.125$, $f(0,1) = -2$, $f(0,-1) = 2$.

So we found the maximum and minimum values of the function and see that *it has a unique minimum, two maxima and a saddle point*. If we plot these functions *f* and *g,* then we will understand the concept of Lagrange multiplier even better.

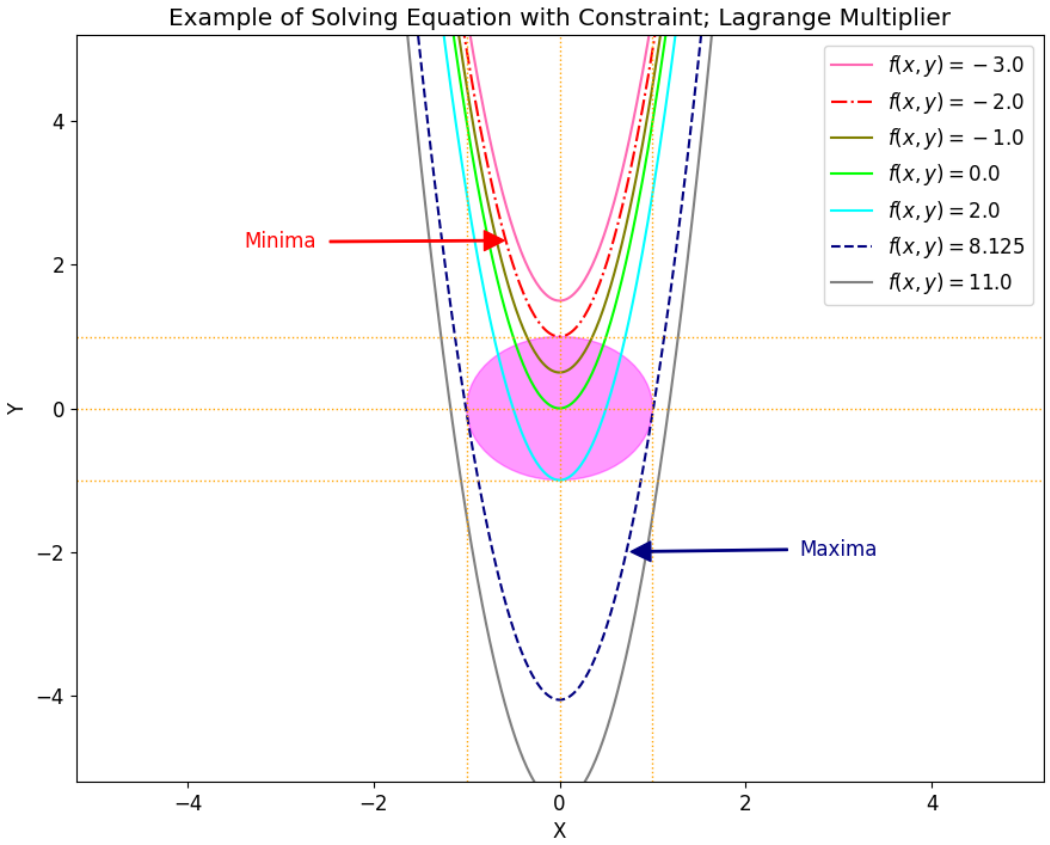

Figure 4: Visualizing Lagrange Multiplier Method

From the figure above we can clearly appreciate that ***the extrema of constrained function f, lie on the surface of the constraint g,*** which is a circle of unit radius. It is a necessary condition. Also the tangent vectors of the function and the constraint are either parallel or anti-parallel at each extremum. Code used for this plot can be found in my github.

We are now ready to go deep into the mathematics behind SVM and successfully apply this technique.

Mathematics of Support Vector Machine:

If you have forgotten the problem statement, let me remind you once again. In figure 1, we are to find a line that best separates two samples. We consider a vector (W) perpendicular to the median line (red line) and, an unknown sample which can be represented by vector *x.*

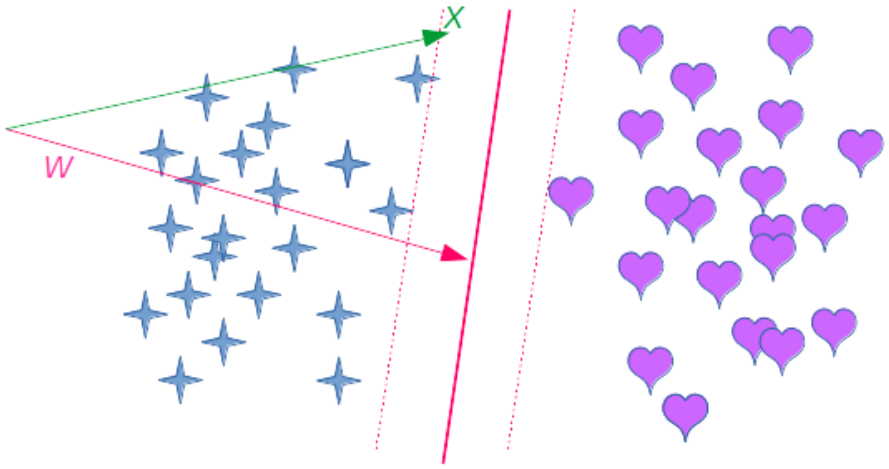

Figure 5: Determine on which side of the road a new sample X lies

To determine on which side of the median line the unknown sample X lies, we first take a projection of X along the perpendicular to the median line, *w*. If this projection is greater than some number (called as bias), then we say that it is on the right side of the line. It is usually called as **decision rule**. Let's put that into an equation

Condition for a sample to be on the right hand side of the median line

$$\vec{w} \cdot \vec{x} \geq b; \;\; \vec{w} \cdot \vec{x} + b \geq 0; \tag{1.4}$$

We don't know specifically anything about the number *b* at this point. Neither we know anything about *w*, except that it is a perpendicular to the median line. To determine *w* and *b*, we now consider known samples and insist on conditions as below

$$\vec{w} \cdot \vec{x_r} + b \geq 1; \;\; \vec{w} \cdot \vec{x_l} + b \leq 1; \tag{1.5}$$

where x_r is a known sample on the right of the *boundary line*, and x_l is a known sample on the left of the boundary line.

So, for known samples on the right (left) we insist that the decision rule (eq. 1.4) to be greater (less) than or equal to 1. *For samples on the boundary line (support vectors) the equality sign holds.* Now our aim is to have a single equation instead of two and to do that let's introduce a variable which is positive for the samples on the right and negative for samples on the left.

$$y_i = \begin{cases} +1; & \text{for } x_r \\ -1; & \text{for } x_l \end{cases} \tag{1.6}$$

With the introduction of this new variable, we will multiply L.H.S. of both conditional equations (eq. 1.5) and we get

$$\begin{aligned} y_i \left(\vec{w} \cdot \vec{x_i} + b\right) &\geq 1; \\ y_i \left(\vec{w} \cdot \vec{x_i} + b\right) - 1 &\geq 0; \end{aligned} \tag{1.7}$$

Another condition we will impose here (rather intuitive) is that for samples on the gutter ('Support vectors') L.H.S would be exactly zero.

$$y_{SV} \left(\vec{w} \cdot \vec{x_{SV}} + b\right) - 1 = 0; \text{ Only for Support Vectors} \tag{1.8}$$

Since our main aim is to find the widest road between the samples, we will now proceed to define the distance between the two parallel lines (or the width of the road).

To do this we first select any two support vectors on either side and calculate the difference vector

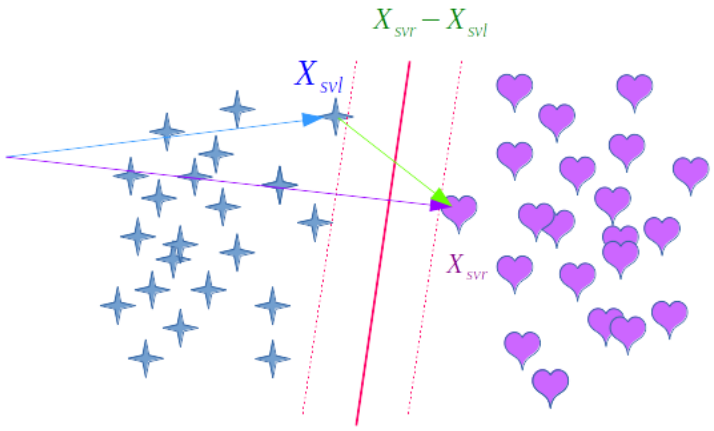

Figure 6: Difference of two support vectors

From the figure above, you can understand that *the width of the road will be the dot product of this difference vector with an unit vector perpendicular to the road.*

But wait ! we have already defined a vector perpendicular to the best line, i.e. *w.* So we are ready to formulate the distance of the road as

Width of the road

$$\text{Width} = (\overrightarrow{x_{svr}} - \overrightarrow{x_{svl}}) \cdot \frac{\overrightarrow{w}}{|w|} \tag{1.9}$$

Can we simplify this equation ? Let's look back to eq. 1.8 and see whether you can come up with something .

Using eq. 1.8 and using the fact that *y* is positive for samples on the right and negative for samples on the left, we can simplify equation 1.9 as

$$\text{Width} = \frac{1-b}{|w|} - \frac{-b-1}{|w|} = \frac{2}{|w|} \tag{1.10}$$

Since we want to maximize the width of the road, our whole problem right now boils down to a very important info —

$$\text{Minimize } \frac{1}{2}|w|^2$$

keeping in mind the constraints that we have before i.e. eq. 1.8.

So rather than intuitively as we have done before, now we can mathematically appreciate that to find the best line is indeed a **constrained optimization** problem.

We are ready to apply our understanding of Lagrange's method now.

First of all here there will be several constraints depending upon the number of support vectors, so the constraint term containing the multiplier will be

$$\alpha_i \left(y_{SVi} \left(\overrightarrow{w} \cdot \overrightarrow{x_{SVi}} + b\right) - 1\right) \ , \tag{1.11}$$

and the Lagrangian that we want to minimize, can be written as

$$L = \frac{1}{2}|w|^2 - \sum_i \alpha_i \ \left(y_{SVi} \left(\overrightarrow{w} \cdot \overrightarrow{x_{SVi}} + b\right) - 1\right) \tag{1.12}$$

Expanding the expression,

$$L = \frac{1}{2}|w|^2 - \sum_i \alpha_i \, y_{SVi} \left(\overrightarrow{w} \cdot \overrightarrow{x_{SVi}} + b\right) + \alpha_i \tag{1.13}$$

Now before starting the minimization, we should identify the variables w.r.t. to which we will differentiate the Lagrangian and set it to zero. If you remember the problem statement and go back to eq. 1.4., we see that vector *w* and bias *b* are the independent variables. So

$$\frac{\partial L}{\partial \overrightarrow{w}} = 0; \text{ differentiate w.r.t a vector} \tag{1.14}$$
$$\frac{\partial L}{\partial b} = 0; \text{ differentiate w.r.t a scalar}$$

Scalar differentiation is as usual but the vector differentiation is performed on scalars here, first the magnitude of the perpendicular vector, and second the dot product of the support vector with the normal vector. The later term for vector differentiation is easy to solve and intuitive for the first term, I would like to give you a hint

$$\frac{\partial}{\partial \overrightarrow{w}} |w|^2 = \frac{\partial}{\partial \overrightarrow{w}} (\overrightarrow{w} \cdot \overrightarrow{w}) \tag{1.15}$$

Using this we can finally write the results as below

$$\frac{\partial L}{\partial \overrightarrow{w}} = \overrightarrow{w} - \sum_i \alpha_i \, y_i \, \overrightarrow{x_{SVi}} = 0 \tag{1.16}$$
$$\frac{\partial L}{\partial b} = \sum_i \alpha_i \, y_i \; = \; 0$$

The above results are most important in this post so far, where we have figured out that **the normal vector (*w*) are linear combination of the support vectors.** We will use this conditions to get our decision rules but let us appreciate that we have reduced another problem in the formulation and that is — because of the conditions from eq. 1.16., *now we can get rid of w and b and instead replace them by the Lagrange multipliers only.* So we use these conditions from eq. 1.16 and try to simplify eq. 1.13

$$L = \frac{1}{2} \left(\sum_i \alpha_i \, y_i \, \overrightarrow{x_{SVi}} \right) \left(\sum_j \alpha_j \, y_j \, \overrightarrow{x_{SVj}} \right) \tag{1.17}$$
$$- \sum_i \alpha_i \, y_i \, \overrightarrow{x_{SVi}} \cdot \left(\sum_j \alpha_j \, y_j \, \overrightarrow{x_{SVj}} \right) - \underbrace{\sum_i \alpha_i \, y_i \, b}_{\sum_i \alpha_i \, y_i = 0} + \sum_i \alpha_i$$

Let's reduce the expression a bit more

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \, \overrightarrow{x_{SVi}} \cdot \overrightarrow{x_{SVj}} \tag{1.18}$$

Finally, we have reached to the end of the post to find out that the **maximization will depend only on the dot product of pairs of support vector.** How awesome is that!!!!

One more interesting result can be reached once we put back the w we get from eq. 1.16. to the decision rule i.e. eq. 1.4 we get

$$\sum_i \alpha_i \, y_i \, \overrightarrow{x_{SV}i} \cdot \overrightarrow{u} \; + \; b \geq 0; \qquad (1.19)$$

So whether a new sample will be on the right of the road depends on the dot product of the support vectors and the unknown sample.

Essentially **everything related to the SVM depends on the simple dot products** of the samples and this for me is mind-blowing and it will hopefully provide some food for thoughts to you too.

Let's wrap up the 2 most important points we have learned

5. We understood that SVM problem is *constrained minimization problem*, both through simple intuition and rigorous mathematics.
6. We learned simple Lagrange's method to solve constrained optimization problem and successfully applied to develop SVM algorithm.