Support Vector Machines

Support Vector Machines (SVMs) are supervised machine learning models which are predominantly employed in classification problems. SVMs involve finding the hyperplane that best divides a dataset into its two classes. This enables the categorisation of new samples.

Intuition

Consider the linearly separable dataset in Figure 1. A dataset is said to be linearly-separable if the positive and negative samples can be separated by a hyperplane (a subspace whose dimension is one less than that of its ambient space). For example, if we have a two-dimensional linearly-separable dataset, we can separate the positive and negative samples with a straight line.

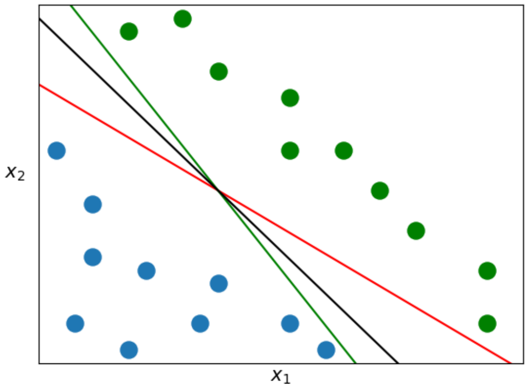

Figure 1: A linearly separable dataset. The green circles represent the positive samples and the blue circles the negative samples. Some possible hyperplanes separating the classes are shown.

There are an infinite number of hyerplanes that separate the positive and negative classes. So how do we go about choosing the optimum hyperplane for separating the dataset in order to make predictions on test samples?

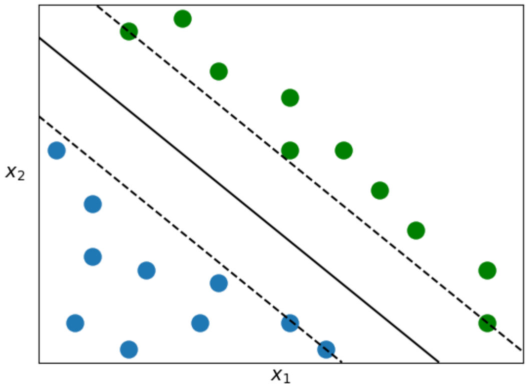

Figure 2: The SVM method. The dotted lines represent the margin and the solid line bisecting the margin represents the hyperplane. The sample (or samples if they are equidistant) of a class nearest to the hyperplane lies on the margin.

The SVM method is to pass the hyperplane through the centre of a margin within which no samples are permitted to lie (Figure 2). The positive sample (or samples if they are equidistant) which is nearest to the hyperplane will lie on the margin on the positive side of the hyperplane, and the negative sample nearest to the hyperplane will lie on the margin on the negative side. These samples which lie precisely on the margins are known as the support vectors, and they fully specify the decision boundary, i.e., the positions of the other samples have no bearing on the position of the hyperplane. The model has the lowest chance of misclassifying the test samples when the margin width is maximised. The more distance we can put between the hyperplane and the nearest sample from each class, the more confidence we can have in the model's ability to distinguish between the classes, and consequently, to successfully classify new samples. Therefore, the maximum margin width coincides with the optimal hyperplane.

Maximising the Margin

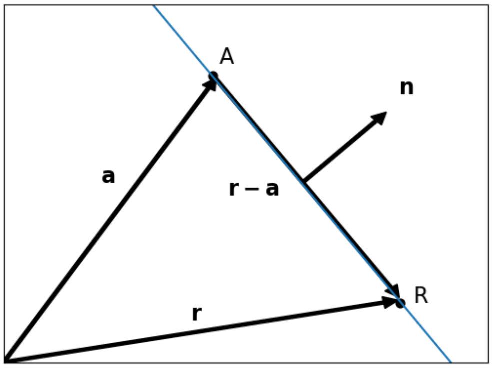

Figure 3: The vector equation of a plane (represented by the blue line) is given by $\mathbf{r} \cdot \mathbf{n} = d$

In Figure 3 we see a plane with two points on it, A and R, with the corresponding position vectors **a** and **r**, respectively. The vector **r-a** must line on the plane, so we may write (**r-a**) $\cdot$ $\mathbf{n} = 0$, where **n** is a vector perpendicular to the plane. This is the equation of the plane and can be expanded to

$$\mathbf{r} \cdot \mathbf{n} = \mathbf{a} \cdot \mathbf{n} = d \tag{1}$$

or

$$\mathbf{r} \cdot \mathbf{n} - d = 0 \tag{2}$$

where the constant d determines the position of the plane.

This means that we can write the equation of the hyperplane in the SVM problem as

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3}$$

where **x** is a position vector lying on the hyperplane, and **w** is a vector normal to the hyperplane. Notice that we have changed the sign of the constant.

For a given sample (x^i, y^i), the functional margin with respect to the hyperplane is defined as

$$\gamma^i = y^i(\mathbf{w} \cdot \mathbf{x}^i + b) \tag{4}$$

It is important to note that, unlike in logistic regression, when working with SVMs we assign $y = -1$ to the negative class, not $y = 0$. Positive samples ($y = 1$) should be situated 'above' the hyperplane, where we have $\mathbf{w} \cdot \mathbf{x}^i + b > 0$, and negative samples should be situated 'below' the hyperplane, where $\mathbf{w} \cdot \mathbf{x}^i + b < 0$. As a result, the functional margin is positive for correctly classified samples. We are more confident in our predictions when a sample lies far away from the hyperplane, i.e., when $|\mathbf{w} \cdot \mathbf{x}^i + b|$ is large. Hence, a positive and large functional margin represents a correct and confident prediction.

It would appear that to find the optimal hyperplane we just have to maximise the functional margin. Unfortunately, this is not the case. If we replace $\mathbf{w}$ with $3\mathbf{w}$ and b with $3b$ in equation (4) (scaling the input by a factor of 3), we are making the functional margin three times larger. It is possible to arbitrarily increase the functional margin just by scaling the inputs in this manner, and without identifying the optimal position of the hyperplane. This would suggest that we are required to place some constraint on $\mathbf{w}$. Let's turn our attention to the geometric margin.

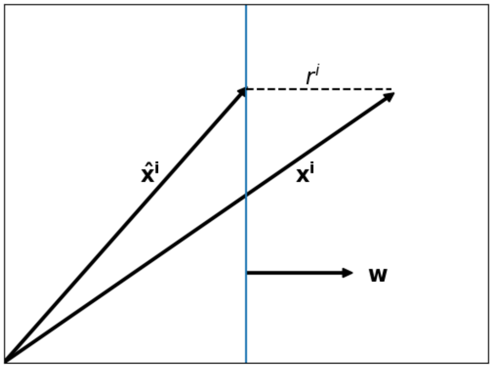

Figure 4: The distance of the sample x^i from the hyperplane is given by r^i.

In Figure 4, the sample x^i has the position vector $\mathbf{x}^i$. The nearest point to x^i on the hyperplane is $\hat{x}^i$ with the position vector $\hat{\mathbf{x}}^{\mathbf{i}}$ ($\hat{x}^i$ is the orthogonal projection of x^i onto the hyperplane). As earlier, $\mathbf{w}$ is a vector perpendicular to the hyperplane. We know that the shortest distance between a point and a plane is along a line perpendicular to the plane. Therefore,

$$\mathbf{x}^i = \hat{\mathbf{x}}^i + r^i\tilde{\mathbf{w}} \tag{5}$$

where r^i is the distance between x^i and $\hat{x}^i$. As $\hat{x}^i$ lies on the hyperplane, we have

$$\mathbf{w} \cdot \hat{\mathbf{x}}^{\mathbf{i}} + \mathbf{b} = \mathbf{0} \tag{6}$$

Substituting equation (5) into equation (6) we obtain

$$\mathbf{w} \cdot (\mathbf{x}^i - r^i\tilde{\mathbf{w}}) + b = 0$$

$$\begin{aligned}\mathbf{w}\cdot\mathbf{x}^i - r^i\frac{\mathbf{w}\cdot\mathbf{w}}{w} + b &= 0\\ \mathbf{w}\cdot\mathbf{x}^i - r^i w + b &= 0\\ r^i &= \frac{\mathbf{w}\cdot\mathbf{x}^i + b}{w} \qquad (7)\end{aligned}$$

where $w = |\mathbf{w}|$. Because the distance r^i must be positive, we can write

$$r^i = \frac{y^i(\mathbf{w}\cdot\mathbf{x}^i + b)}{w} = \frac{\gamma^i}{w} \tag{8}$$

which applies for both positive and negative samples.

We define the functional margin of a dataset, γ, as the smallest of the γ^i of the individual training samples in the dataset. Likewise, the geometric margin, r, is defined as the smallest of the r^i of the individual training samples in the dataset, and measures the distance between the hyperplane and the nearest sample, i.e., the distance between the hyperplane and the margin. Hence, we can write

$$r = \frac{\gamma}{w} = \frac{y^j(\mathbf{w}\cdot\mathbf{x}^j + b)}{w} \tag{9}$$

where the sample x^j, with position vector $\mathbf{x}^j$ (a support vector), is nearest to the hyperplane.

From equation (9) we can see that in contrast to the functional margin, the geometric margin is invariant to the scaling of $\mathbf{w}$ and b. Also, when $w = 1$, the geometric and functional margins are equivalent.

When we defined the geometric margin above, we demanded that

$$r^i \geq r, \qquad i = 1, 2, ..., m \tag{10}$$

where we have m samples. This requirement ensures that no training samples lie within the margin. It can be rewritten as

$$\frac{y^i(\mathbf{w}\cdot\mathbf{x}^i + b)}{w} \geq \frac{\gamma}{w}, \qquad i = 1, 2, ..., m \tag{11}$$

or

$$y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq \gamma, \qquad i = 1, 2, ..., m \tag{12}$$

Therefore, our maximum margin objective can be expressed as

$$\max_{\mathbf{w},b} \frac{\gamma}{w} \quad \textit{subject to} \quad y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq \gamma, \qquad i = 1, 2, ..., m \tag{13}$$

Unfortunately, $\frac{\gamma}{w}$ is non-convex and we cannot solve this optimisation problem. Earlier, we mentioned that the geometric margin is invariant to the scaling of $\mathbf{w}$ and b, and we can now use this property to our advantage. Mathematically, the problem takes its simplest form if we choose to scale the geometric margin $\frac{\gamma}{w}$ such that

$$\gamma = 1 \tag{14}$$

From equation (9) it follows that

$$|\mathbf{w} \cdot \mathbf{x}^j + b| = 1 \tag{15}$$

where again $\mathbf{x}^j$ (a support vector) is the position vector of the sample, x^j, located nearest to the hyperplane, i.e., on the margin.

Consequently, the margins are defined by the planes

$$\mathbf{w} \cdot \mathbf{x} + b = \pm 1 \tag{16}$$

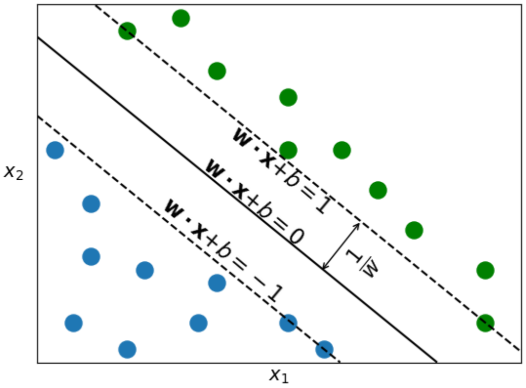
$\mathbf{w} \cdot \mathbf{x} + b = 1$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = -1$
$\frac{1}{w}$
x_2
x_1

Figure 5: The geometry of the SVM problem after scaling of the geometric margin such that $\gamma = 1$.

The maximum margin objective is now

$$\max_{\mathbf{w},b} \frac{1}{w} \quad \textit{subject to} \quad y^i(\mathbf{w} \cdot \mathbf{x}^i + b) \geq 1, \qquad i = 1, 2, ..., m \tag{17}$$

In practice, instead of maximising $\frac{1}{w}$ we choose to minimise w, or for greater convenience, $\frac{w^2}{2}$. So, the optimisation problem becomes

$$\min_{\mathbf{w},b} \frac{w^2}{2} \quad \textit{subject to} \quad y^i(\mathbf{w} \cdot \mathbf{x}^i + b) \geq 1, \qquad i = 1, 2, ..., m \tag{18}$$

This is a quadratic programming problem, which consists of a quadratic objective function subject to linear constraints. This optimisation problem is known as hard-margin SVM, and is presented in its primal form. It does not allow for any violations of the margin condition.

Soft-Margin SVM

In practice, linearly-separable datasets are rare. It is much more common to encounter non-linearly separable datasets, such as that shown in Figure 6.

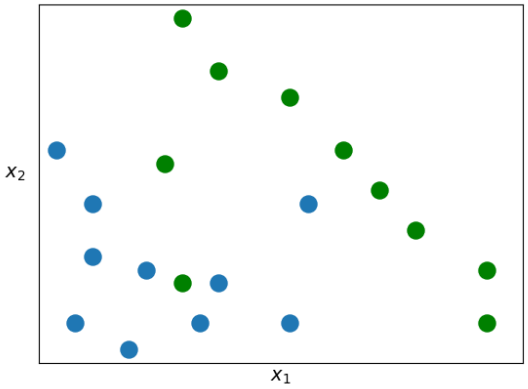

Figure 6: A non-linearly separable dataset. The green circles are positive samples and the blue are negative. It is not possible to separate the positive and negative samples with a straight line.

It should be clear that hard-margin SVM, which requires the positive and negative samples to be separated by a hyperplane, plus a strictly observed margin, cannot be employed with a non-linearly separable dataset.

Soft-margin SVM can operate with non-linearly separable data. The introduction of a slack variable, ε^i, corresponding to each sample, x^i, permits a sample to lie within the margin or even on the wrong side of the hyperplane. ε^i is measured from the margin corresponding to the label (y^i) of the sample: for a correctly classified sample $\varepsilon^i = 0$, for a sample within the margin but on the right side of the hyperplane $0 < \varepsilon^i < 1$, and for a misclassified sample $\varepsilon^i > 1$.

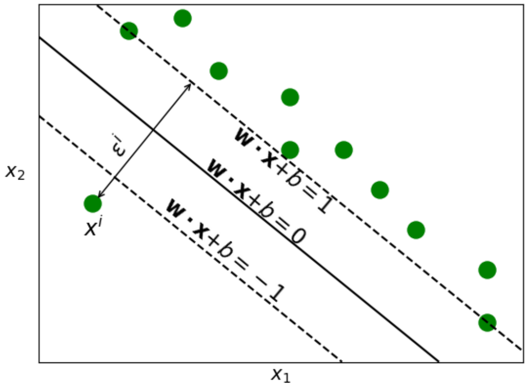

Figure 7: The positive sample x^i is situated on the negative side of the hyperplane. The slack variable ε^i measures the distance between x^i and the positive margin $\mathbf{w} \cdot \mathbf{x} + b = 1$.

Therefore, we formulate our new constraint as

$$y^i(\mathbf{w} \cdot \mathbf{x}^i + b) \geq 1 - \varepsilon^i \tag{19}$$

The issue now is that ε^i is completely unrestricted, which means we are enabling the sample to lie anywhere in the space, even far over the wrong side of the hyperplane if ε^i is large enough. This makes the entire optimisation problem useless, unless we find a way to keep the ε^i as small as possible. Therefore, when we add the slack variables to the optimisation problem we make sure to regularise them:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \frac{w^2}{2} + C\sum_{i=1}^{m}\varepsilon^i \quad \textit{subject to} \quad y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq 1 - \varepsilon^i \\ &\varepsilon^i \geq 0 \qquad i = 1, 2, ..., m \end{aligned} \tag{20}$$

This is the soft-margin SVM optimisation objective in its primal form. We have seen the regularisation parameter C before with linear and logistic regression, and here it performs a similar role, penalising large values of ε^i. If C is large, the slack variables are heavily penalised and any misclassification is unacceptable. If $C = 0$, the slack variables are not penalised and large misclassifications are acceptable.

An alternate derivation of the soft-margin SVM optimisation problem involves the hinge-loss function, $L(t)$ which is defined as

$$L(t) = max\,\{0, 1-t\} \tag{21}$$

where

$$t = y(\mathbf{w}\cdot\mathbf{x} + b) \tag{22}$$

and

$$\begin{aligned} max\,\{0,1-t\} &= \begin{cases} 0 & if \quad 1-t \leq 0 \\ 1-t & if \quad 1-t > 0 \end{cases} \\ &= \begin{cases} 0 & if \quad t \geq 1 \\ 1-t & if \quad t < 1 \end{cases} \end{aligned} \tag{23}$$

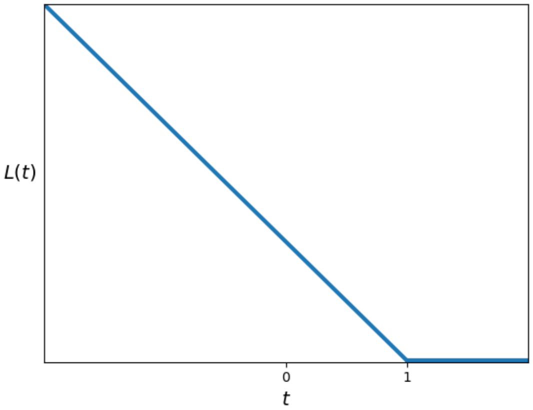

Figure 8: The hinge-loss function, $L(t)$.

For a sample, x^i, lying on the correct side of the hyperplane, and on or outside of the margin ($|\mathbf{w} \cdot \mathbf{x}^i + b| \geq 1$), equation (22) gives $t \geq 1$ and the hinge-loss returns a value of zero. If the sample is lying on the correct side of the hyperplane but within the margin ($0 < |\mathbf{w} \cdot \mathbf{x}^i + b| < 1$), we have $0 < t < 1$. In this case the hinge-loss returns the positive value $1 - t$. Finally, if the sample is lying on the wrong side of the hyperplane, then $t < 0$ and the hinge-loss return value of $1 - t$ increases the further the sample lies over the wrong side of the hyperplane.

Using the hinge-loss function, the soft-margin SVM problem is formulated as the following unconstrained optimisation problem:

$$\min_{\mathbf{w},b} \quad \frac{w^2}{2} + C \sum_{i=1}^{m} \, max\left\{0, 1 - y^i(\mathbf{w} \cdot \mathbf{x}^i + b)\right\} \tag{24}$$

It can be shown that (24) is equivalent to (20):

Using equation (22), the constraint (19) for a single sample can be written as

$$\varepsilon \geq 1 - t \tag{25}$$

As $1-t$ is maximised, the upper bound is reached, so $1-t = \varepsilon$. Therefore, we can write

$$max\,\{0, 1-t\} = max\,\{0, \varepsilon\} = \begin{cases} 0 & if \quad \varepsilon \leq 0 \\ \varepsilon & if \quad \varepsilon > 0 \end{cases}$$

By employing the constraint $\varepsilon \geq 0$, we have

$$max\,\{0, 1-t\} = \begin{cases} 0 & if \quad \varepsilon = 0 \\ \varepsilon & if \quad \varepsilon > 0 \end{cases}$$

So, we can say that $max\,\{0, 1-t\}$ is equivalent to ε. Hence,

$$L(t) = max\left\{0, 1 - y^i(\mathbf{w} \cdot \mathbf{x}^i + b)\right\}$$

is equivalent to

$$\varepsilon^i \quad subject\ to \quad y^i(\mathbf{w} \cdot \mathbf{x}^i + b) \geq 1 - \varepsilon^i, \qquad \varepsilon^i \geq 0$$

The Dual Problem

In optimisation, duality refers to the principle that an optimisation problem in one set of variables (the primal variables) may be converted into another optimisation problem in another set of variables (the dual variables). One of the main approaches to duality is Lagrangian duality, wherein the Lagrange multiplier method is utilised to convert the optimisation problem into one in which the non-negative Lagrange multipliers are the dual variables. The conversion from the primal problem to the Lagrange dual problem is often desirable as the dual can be easier to solve and the solution may prove more advantageous to work with.

We now briefly review the Lagrange multiplier method (see Appendix for further discussion). Consider a function $f(x, y)$ which we want to optimise subject to the constraint $g(x, y) = 0$. We proceed by constructing the Lagrange function

$$\mathcal{L}(x, y) = f(x, y) - \lambda g(x, y) \tag{26}$$

where $\lambda \geq 0$ is called a Lagrange multiplier. The next step is to find the stationary points of $\mathcal{L}(x, y)$, i.e., the points where $\boldsymbol{\nabla}\mathcal{L} = 0$. Therefore, we need to solve for

$$\frac{\partial \mathcal{L}}{\partial x} = 0, \quad \frac{\partial \mathcal{L}}{\partial y} = 0 \tag{27}$$

Our soft-margin SVM optimisation problem in its primal form, (20), has two constraints, so we need to introduce a Lagrange multiplier for each of them. These will be $\alpha^i \geq 0$ and $\gamma^i \geq 0$. The Lagrange function we construct is

$$\mathcal{L}(w, b, \varepsilon, \alpha, \gamma) = \frac{w^2}{2} + C\sum_{i=1}^{m} \varepsilon^i - \sum_{i=1}^{m} \alpha^i[y^i(\mathbf{w} \cdot \mathbf{x}^i + b) + \varepsilon^i - 1] - \sum_{i=1}^{m} \gamma^i \varepsilon^i \tag{28}$$

where we are summing over all the constraints, i.e. there is an α^i and γ^i for every sample, x^i. Now we need to find the partial derivatives of this Lagrange function with respect to the three primal variables w, b, and ε:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m} \alpha^i y^i \mathbf{x}^i \tag{29}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m} \alpha^i y^i \tag{30}$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon^i} = C - \alpha^i - \gamma^i \tag{31}$$

Setting these partial derivatives equal to zero we obtain

$$\mathbf{w} = \sum_{i=1}^{m} \alpha^i y^i \mathbf{x}^i \tag{32}$$

$$\sum_{i=1}^{m} \alpha^i y^i = 0 \tag{33}$$

$$C = \alpha^i + \gamma^i \tag{34}$$

Next, we substitute equations (32) and (34) into the Lagrange function (28) to find the dual $\mathcal{D}$:

$$\begin{aligned}\mathcal{D}(\varepsilon,\alpha,\gamma) = &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^i\alpha^j y^i y^j \mathbf{x}^i\cdot\mathbf{x}^j + \sum_{i=1}^{m}(\alpha^i+\gamma^i)\varepsilon^i \\ &-\sum_{i=1}^{m}\alpha^i\left\{y^i\left[\left(\sum_{j=1}^{m}\alpha^j y^j \mathbf{x}^j\right)\cdot\mathbf{x}^i + b\right]+\varepsilon^i - 1\right\} - \sum_{i=1}^{m}\gamma^i\varepsilon^i\end{aligned} \tag{35}$$

$$\begin{aligned}\mathcal{D}(\varepsilon,\alpha,\gamma) = &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^i\alpha^j y^i y^j \mathbf{x}^i\cdot\mathbf{x}^j - \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^i\alpha^j y^i y^j \mathbf{x}^j\cdot\mathbf{x}^i + \sum_{i=1}^{m}(\alpha^i+\gamma^i)\varepsilon^i \\ &- b\sum_{i=1}^{m}\alpha^i y^i - \sum_{i=1}^{m}\alpha^i\varepsilon^i + \sum_{i=1}^{m}\alpha^i - \sum_{i=1}^{m}\gamma^i\varepsilon^i\end{aligned} \tag{36}$$

Using equation (33), the fact that $\mathbf{x}^i\cdot\mathbf{x}^j = \mathbf{x}^j\cdot\mathbf{x}^i$, and cancelling terms, we are left with

$$\mathcal{D}(\varepsilon,\alpha,\gamma) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^i\alpha^j y^i y^j \mathbf{x}^i\cdot\mathbf{x}^j + \sum_{i=1}^{m}\alpha^i \tag{37}$$

In Lagrangian duality, the dual problem is maximised. Equivalently, we can minimise its negative:

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^i\alpha^j y^i y^j \mathbf{x}^i\cdot\mathbf{x}^j - \sum_{i=1}^{m}\alpha^i \\ &\textit{subject to}\ \sum_{i=1}^{m}\alpha^i y^i = 0 \\ &0\le\alpha^i\le C,\quad i=1,2,...,m\end{aligned} \tag{38}$$

This is the dual form of the soft-margin SVM optimisation problem. Notice how the problem has been expressed exclusively in terms of the dual variables (the Lagrange multipliers), α. It is simple to deduce the constraint $0 \le \alpha^i \le C$ by rearranging equation (34) to get $\alpha^i = C - \gamma^i$.

Equation (32) states that the weight vector, $\mathbf{w}$, solution to the soft-margin SVM problem is a linear combination of the training set position vectors, $\mathbf{x}^i$. A vector, $\mathbf{x}^i$ appears in the expansion only if $\alpha^i \ne 0$. These vectors are the support vectors, which we introduced earlier on. To see why this is true we need to study the constraints used in the Lagrange function, (28). These were

$$\alpha^i[y^i(\mathbf{w}\cdot\mathbf{x}^i+b)+\varepsilon^i-1]=0\quad \textit{and}\quad \gamma^i\varepsilon^i=0$$

If $\alpha^i \ne 0$, we must have $y^i(\mathbf{w}\cdot\mathbf{x}^i+b)+\varepsilon^i-1=0$, which is $y^i(\mathbf{w}\cdot\mathbf{x}^i+b)=1-\varepsilon^i$. If $\varepsilon^i=0$, we obtain $y^i(\mathbf{w}\cdot\mathbf{x}^i+b)=1$, meaning $\mathbf{x}^i$ is situated on the margin. For a sample on the wrong side of the margin, $\varepsilon\ne 0$, which means $\gamma=0$. From equation (34) we get $\alpha^i=C$. Therefore, in soft-margin SVM the support vectors are not only the position vectors, $\mathbf{x}^i$, that lie on the margin, but also those that lie on the wrong side of the margin.

References

Deisenroth, M. S., Faldo, A. A., Ong, C. S. (2020). *Mathematics for Machine Learning.* Cambridge University Press.

Mohri, M., Rostamizadeh, A., Talwalker, A. (2012). *Foundations of Machine Learning.* The MIT Press, Cambridge, Massachusetts.

Ng, A. *Part V - Support Vector Machines.* CS229 Lecture Notes.

The Open University. (2013). *MST224 Mathematical Methods Book 3: Scalar and vector fields.* The Open University, Milton Keynes.

Appendix

The Method of Lagrange Multipliers

Consider the functions $f(x, y) = xy + 1$, and $g(x, y) = x^2 + y^2$, which are plotted in Figure 9.

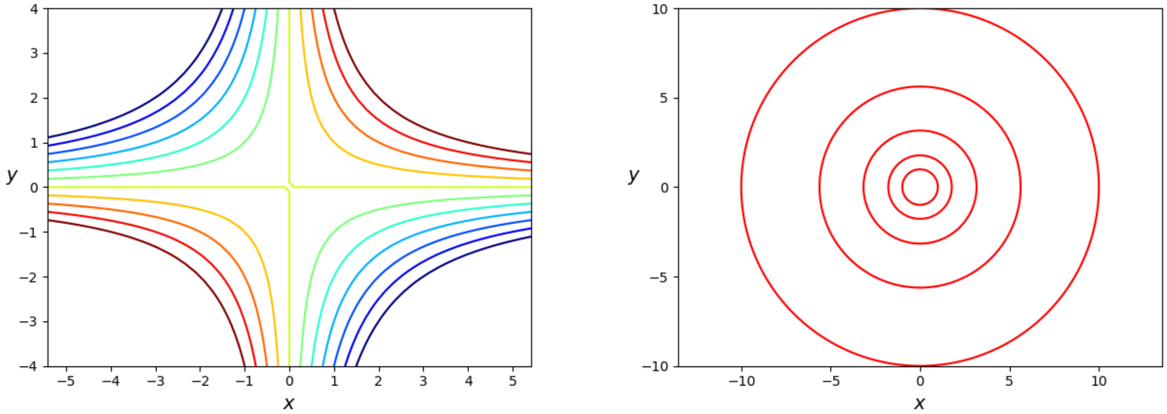

Figure 9: (left) A contour plot of the function $f(x, y) = xy + 1$. The maroon contour line is that with the largest value of $f(x, y)$. In descending order of values of $f(x, y)$ the contour lines are coloured: maroon, red, orange, green, aqua, blue, and navy. (right) A contour plot of the function $g(x, y) = x^2 + y^2$.

Gradient vectors always point in the direction perpendicular to contour lines. This means that the gradient vector of $g(x, y)$ will point radially outwards (as viewed in Figure 9) everywhere.

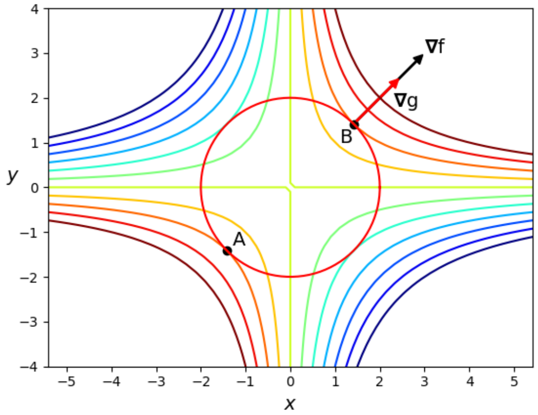

Figure 10: The two maxima of $f(x, y)$ subject to $g(x, y) = 2$ occur at the points A and B. The gradients $\boldsymbol{\nabla} f$ and $\boldsymbol{\nabla} g$ are parallel at these points. The minima of $f(x, y)$ subject to $g(x, y) = 2$ occur where the constraint curve just kisses the aqua blue contour line of $f(x, y)$.

Let's say we are constrained to $g(x, y) = 2$ on the surface of $f(x, y)$. This means we are constrained to a contour line of $g(x, y)$. If we plotted all the contour lines of $f(x, y)$ on Figure 10, we would find that some would intersect $g(x, y) = 2$ at two points, some would not intersect it at all, and one of the contour lines would just kiss the constraint curve, as happens at points A and B. These points are the maxima of $f(x, y)$ subject to $g(x, y) = 2$, because $f(x, y)$ will take on a lower value anywhere else we move to on the constraint. If we start at B, for example, and move clockwise along the constraint, we will be descending continuously until we reach one of the minima of $f(x, y)$ subject to the constraint, which occur where the constraint curve just kisses the aqua blue contour line of $f(x, y)$ in Figure 10. Then we ascend continuously until reaching A. At points A and B it is clear that the tangents to the contour line of $f(x, y)$ and $g(x, y) = 2$ are parallel. Therefore, the gradients $\boldsymbol{\nabla} f$ and $\boldsymbol{\nabla} g$ must also be parallel at these points (the gradients could be anti-parallel for a different surface and constraint). As a consequence, we can write

$$\boldsymbol{\nabla} f = \lambda \boldsymbol{\nabla} g \tag{39}$$

where λ is a non-zero constant. We can write equation (39) in component form as

$$\frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x} = 0, \quad \frac{\partial f}{\partial y} - \lambda \frac{\partial g}{\partial y} = 0 \tag{40}$$

Introducing the Lagrange function $\mathcal{L}(x, y) = f(x, y) - \lambda g(x, y)$, we can see that the equations (40) can be written as

$$\frac{\partial \mathcal{L}}{\partial x} = 0, \quad \frac{\partial \mathcal{L}}{\partial y} = 0$$

Together with the constraint $g(x, y) = 2$, we have three equations for three unknowns and can eliminate λ to find values for x and y at the stationary points of $f(x, y)$ subject to the constraint.