1 Introduction to Support Vector Machine

This introduction aims to briefly explain the motivation and important development steps of the formulation and solution of an SVM. Implementation and background information is not included but we provide pointers to the original literatures. The nomenclature follows [1].

SVM was first introduced by Boser, Guyon and Vapnik in 1992 [2] to elicit complex patterns from data (e.g. classification, regression). For our interest, we only introduce its usage in classification. Reader may refer to Section 6.2 of [1] for the regression case.

Formally, SVM finds a separation hyperplane of the two (or multiple) classes of data \( \{x_i, y_i\}, i = 1, ..., l \) where \( \{x_i\} \) is a data matrix and \( \{y_i\} \) class labels. A special feature of SVM (like other kernel machines) is that it maps the data \( x_i \) from its input space \( X \), to a feature space \( F \):

\[
\mathbf{x} = (x_1, ..., x_n) \mapsto \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), ..., \phi_N(\mathbf{x})).
\]  

(1)

The mapping is usually nonlinear, which helps to realize a complex classification rule using a linear hyperplane in the feature space. Figure 1 illustrates this advantage of SVM.

1.1 Linear Learning Machine

Before we introduce the concept of kernel, we first deal with a linear case (without mapping) and this will help to realize some critical features of SVM.

Suppose we have a training set \( S = \{x_i, y_i\}, \mathbf{x} \in X \) and \( y_i \in \{-1, +1\} \). If the data is separable in the input space, then we can derive a separation hyperplane as shown in Figure 2.

![Image](image.png)

Figure 1: A feature map can simplify the classification task [1]
Figure 2: A separation hyperplane \((w, b)\) for a two dimensional training set [1]

\[
f(x) = \langle w \cdot x \rangle + b,
\]

and the decision function will be

\[
h(x) = \text{sgn}(f(x)).
\]

This problem is ill-defined since there are infinite number of hyperplanes separating the two classes if possible. It is proved that the hyperplane with the maximal (geometric) margin will minimize the risk of overfitting (Chapter 4 in [1]). Two definitions of margin exists:

- A functional margin: \(\text{func} = \min_i y_i f(x_i)\);
- A geometric margin: \(\text{geom} = \min_i \frac{y_i f(x_i)}{||w||}\).

If we set the functional margin to 1, i.e. for the data points on the boundaries (larger O and X in Figure 3) we have

\[
y_i < w \cdot x_i > + b = 1,
\]

then the geometric margin can be calculated as

\[
\gamma = \frac{1}{2} \left( \frac{w}{||w||_2} \cdot x^+ \right) - \frac{1}{||w||_2} \left( \frac{w}{||w||_2} \cdot x^- \right)
\]

\[
= \frac{1}{2 ||w||_2} (\langle w \cdot x^+ \rangle - \langle w \cdot x^- \rangle)
\]

\[
= \frac{1}{||w||_2}.
\]

Therefore, maximizing the margin is equivalent to minimizing \(\langle w \cdot w \rangle\).
1.2 Primal and Dual Formulation

The primal problem is thus the following:

$$\begin{align*}
\min_{w,b} & \quad \langle w \cdot w \rangle, \\
\text{s.t.} & \quad y_i (\langle w \cdot x_i \rangle + b) \geq 1, \quad i = 1, \ldots, l.
\end{align*}$$

(4)

To solve this, we write out its Lagrangian:

$$\frac{1}{2} \langle w \cdot w \rangle - \sum_{i=1}^{l} y_i \alpha_i [\langle w \cdot x_i \rangle + b] - 1].$$

(5)

The KKT conditions require the derivatives of the Lagrangian to be zero, and therefore:

$$w = \sum_{i=1}^{l} y_i \alpha_i x_i,$$

(6)

$$0 = \sum_{i=1}^{l} y_i \alpha_i.$$

Plug (6) into (5) and include the other KKT conditions, we derive its dual form:

$$\begin{align*}
\max_{\alpha} & \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle, \\
\text{s.t.} & \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \\
& \quad \alpha_i \geq 0, \quad i = 1, \ldots, l.
\end{align*}$$

(7)

It is proved (See Chapter 5 of [1]) that this problem has strong duality and therefore solving the dual is equivalent to solving the primal.
From KKT we also know that $\alpha_i$ is nonzero only when $y_i(<w \cdot x_i> + b) = 1$. This means that only those points on the boundaries will have nonzero $\alpha$s, and these points are called the “support vector”s. This property of $\alpha$ is known as the sparseness of SVM.

We shall also notice that in the dual problem (Equation (8)), the original data only appears in the inner product form. Taking into consideration the strong dual property of this problem, we conclude (happily) that the decision is built upon the inner product of the original data, i.e. the Gram matrix \{<x_i \cdot x_j>\}_{i,j}, rather than the data itself. This is the key property leads to the usage of kernels.

1.3 Kernel

Consider the mapping in Equation (1). If we perform a linear separation in the feature space, the dual problem will be similar to Problem (8):

$$
\begin{align*}
\max_{\alpha} & \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j <\phi(x_i) \cdot \phi(x_j)> , \\
\text{s.t.} & \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \\
& \quad \alpha_i \geq 0, i = 1, ..., l.
\end{align*}
$$

(8)

Since the mapping $\phi()$ only exists in the inner product form, we introduce “kernel” as a function on $x$: $$K(x, z) = <\phi(x) \cdot \phi(z)>.$$ (9)

By constructing a kernel, it is not necessary to know the exact form and dimensionality of $\phi$. However, the kernel must acquire certain properties:

- Symmetric: $K(x, z) = K(z, x)$;
- Cauchy-Shwarz inequality: $K(x, z)^2 \leq K(x, x)K(z, z)$
- Mercer’s Theorem: the Gram matrix \{<x_i \cdot x_j>\}_{i,j} must be positive semi-definite so that from the decomposition: $$K(x, z) = \sum_{i} \lambda_i \phi_i(x)\phi_i(z),$$

$\sqrt{\lambda_i}\phi_i(x)$ can be extracted as features. This property also ensures that the dual problem (Problem (9)) is convex (or concave if regarded as a maximization) and has no local minima (maxima), which is an advantage of SVM compared with neural networks.

Some example of kernels are:

- Polynomial(homogeneous): $K(x, z) = <x \cdot z>^d$;
- Polynomial(inhomogeneous): $K(x, z) = (<x \cdot z> + 1)^d$;
- Radial basis function: $K(x, z) = exp(-\gamma||x - z||^2)$, for $\gamma > 0$;
• Gaussian radial basis function: $K(x, z) = \exp\left( -\frac{||x-z||^2}{2\sigma^2} \right)$;

• Hyperbolic tangent: $K(x, z) = \tanh(\kappa < x \cdot z > + c)$, for $\kappa > 0$ and $c < 0$.

The usage of these kernels is not studied. But it is recommended that RBF kernel is always the first choice [3]. Figure 4 shows a classification result using a Gaussian kernel.

### 1.4 Non-separable situation

So far we assume that there exists a hyperplane that perfectly separates the two classes in the feature space. In case this is not true, the primal problem will have no solution and the dual objective will be unbounded. To avoid this, slack variables are introduced to relax the constraints. The modified primal problem looks like the following:

$$
\min_{w, b} \quad \frac{1}{2} < w \cdot w > + C \sum_{i} \xi_i, \\
\text{s.t.} \quad y_i(< w \cdot x_i > + b) \geq 1 - \xi_i, \quad i = 1, ..., l;
$$

or:

$$
\min_{w, b} \quad \frac{1}{2} < w \cdot w > + C \sum_{i} \xi_i, \\
\text{s.t.} \quad y_i(< w \cdot x_i > + b) \geq 1 - \xi_i, \quad i = 1, ..., l;
$$
\[
\min_{w,b} \frac{1}{2} \langle w \cdot w \rangle + C \sum_{i} \xi_i^2,
\]
\[
s.t. \quad y_i (\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, ..., l.
\]

These two problem formulations are called 1-norm and 2-norm soft-margin classifiers accordingly. Either can be used to deal with the non-separable situation. The parameter \(C\) is an arbitrary value and shall be tuned in order to find the best classifier.

1.5 Summary

This section reviewed the development and essential properties of an SVM. We started by modeling the two-class classification problem as a maximum margin problem and found out that with strong duality, this problem can be solved with the gram matrix rather than the data matrix. Such an observation led to the introduction of “kernel”s and thus complex patterns can be recognized by being mapped to unknown feature spaces and classified with hyperplanes.

References

