CS 337, Fall 2023
Support Vector Machines and Kernels

Scribes: Sri Lakshmi Teja, Joshitha Chowdhary, Ananth Krishna Kidambi,
Divakar Sai Savaram, Kanakala Dittu Akanksh, Shikhar Parmar
(***Equal contribution by all scribes***)
Edited by: Govind Saju

Sept 11, 2023

Recap

We are currently studying about the support vector machines. In this, the goal is to find the best decision boundary. The aim is to maximize the margin while classifying the points. Here margin refers to the distance of the closest points across both classes. The following formula represents the margin.

$$r(w,b) = \min_{i=1ton} \frac{(w^T x_i + b)}{\|w\|}$$

Now the SVM optimization problem to solve is:

$$\max_{w \in R^n} 2r(w,b)$$

$$s.t \forall i,\ y_i(w^T x_i + b) \geq 0$$

the above optimization problem can be further reduced to the following problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t \forall i,\ y_i(w^T x_i + b) \geq 1$$

The given problem represents Hard Margin SVM optimization. In cases where margin constraints are violated, we aim to minimize the hinge loss while maximizing the margin, resulting in a modified optimization problem which is as follows:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C.\Sigma_i \max\{0, 1 - y_i(w^T x_i + b)\}$$

The above problem indicates the soft margin SVM optimization. Here hyper parameter C indicates the penalty or regularization factor.

1 Dual Form of Optimization Problems

The generic constrained optimization problem is as follows:

$$\begin{aligned} p^* = \min_w f(w) \quad & (1) \\ \text{such that:} \quad g_1(w) &\leq 0 \\ g_2(w) &\leq 0 \\ \cdots & \\ g_n(w) &\leq 0 \end{aligned}$$

This is also known as the **primal** problem. This problem can be reformulated using Lagrange multipliers as follows-

$$\begin{aligned} \mathcal{L}(w, \alpha) &= f(w) + \sum_i \alpha_i g_i(w) \\ p^* &= \min_w \max_{\alpha_i \geq 0} \mathcal{L}(w, \alpha) \quad (2) \end{aligned}$$

where $\{\alpha_i\}$ are the Lagrange multipliers and $\mathcal{L}(w, \alpha)$ is the Lagrangian. Note that this problem is unconstrained and hence easier to analyze.

Equivalence of (1) and (2)

For any w, if, for some i, $g_i(w) > 0$, then the term $\alpha_i g_i(w)$ is unbounded, and hence,

$$p^* = \begin{cases} f(w) & \forall i g_i(w) \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

Now, consider the **dual** problem:

$$d^* = \max_{\alpha_i \geq 0} \min_w L(w, \alpha)$$

In general, $d^* \leq p^*$, since

$$\begin{aligned} & \forall_{\alpha_i \geq 0, w} L(w, \alpha) \geq \min_w L(w, \alpha) \\ \implies & \forall_{\alpha_i \geq 0, w} \max_{\alpha_i \geq 0} L(w, \alpha) \geq \min_w L(w, \alpha) \\ \implies & \forall_{\alpha_i \geq 0} \min_w \max_{\alpha_i \geq 0} L(w, \alpha) \geq \min_w L(w, \alpha) \\ \implies & \min_w \max_{\alpha_i \geq 0} L(w, \alpha) \geq \max_{\alpha_i \geq 0} \min_w L(w, \alpha) \\ \implies & p^* \geq d^* \end{aligned}$$

However, in the case of SVMs, it can be shown that $d^* = p^*$, that is, the optimal values of the dual and the primal problems are the same.

2 The SVM Optimization Problem

2.1 Dual form for the SVM

Hard-Margin SVM: The dual form of the hard margin SVM is given by-

$$\max_{\alpha_i \geq 0} \left(\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \right) \quad (3)$$

such that $\forall i, \alpha_i \geq 0$, $\sum_i \alpha_i x_i = 0$, and the optimal value of w is given by

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

Soft-Margin SVM: The dual form of the soft margin SVM is similar to that of the hard margin SVM, with the additional constraint that $\forall i, \alpha_i \leq C$. That is-

$$\max_{\alpha_i \geq 0} \left(\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \right) \tag{4}$$

such that $\forall i, 0 \leq \alpha_i \leq C$, $\sum_i \alpha_i x_i = 0$, and the optimal value of w is the same as the previous case.

2.2 Some Observations Based on the Dual Form

1. By constructing the dual, we find $w = \sum_{i=1} \alpha_i y_i x_i$
2. If the solution to the primal problem is w^*, and the solution to the dual is α^*, the KKT conditions on w^*, α^* hold for the SVM problems. One of these, the dual complementarity constraint, is of particular interest. It states that-

   $$\forall i, \alpha_i g_i(w^*) = 0$$

   where $g_i(w) = y_i(w^{*T} x_i + b) - 1$. This means that $\alpha_i > 0$ corresponds to those training examples where $g_i(w) = 0$, which is the set of support vectors $\{x_i | y_i(w^{*T} x_i + b) = 1\}$.
3. The dual form operates on the inner products of training instance pairs

2.3 Test time with SVM Classifiers

We classify new points based on the side of the hyperplane on which it lies. That is, we classify a new point x as follows-

$$\begin{aligned} \text{Prediction} &= \text{sign}(w^* T x + b) \\ &= \text{sign} \left(\sum_i \alpha_i^* y_i x_i^T x + b^* \right) \end{aligned}$$

Derivation of the Dual Problems of SVM

Recall that the Lagrangian of the hard margin SVM problem has the form-

$$\mathcal{L}(w,b) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i(w^T x_i + b) - 1\right)$$

and the optimization problem is given by-

$$\min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w,b)$$

However, for the SVM problem, this is the same as the optimization problem-

$$\max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w,b)$$

Now, we can solve for the optimal w and b by taking the derivatives of $\mathcal{L}(w,b)$ with respect to w and b and setting them to zero. This gives us-

$$\begin{aligned}
&\frac{\partial \mathcal{L}(w,b)}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \\
\implies &w = \sum_i \alpha_i y_i x_i \\
&\frac{\partial \mathcal{L}(w,b)}{\partial b} = \sum_i \alpha_i y_i = 0 \\
\implies &\sum_i \alpha_i y_i = 0
\end{aligned}$$

We can now use this to simplify the Lagrangian as follows-

$$\begin{aligned}
\min_{w,b} \mathcal{L}(w,b) &= \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i(w^T x_i + b) - 1\right) \\
&= \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_i \alpha_i y_i ((\sum_j \alpha_j y_j x_j^T) x_i + b) + \sum_i \alpha_i \\
&= \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j
\end{aligned}$$

with the constraint that $\sum_i \alpha_i y_i = 0$, which is the dual form of the hard margin SVM.

The Lagrangian for the soft margin SVM is given by-

$$\mathcal{L}(w,b,\xi_i) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left(y_i(w^T x_i + b) - 1 + \xi_i\right)$$

Note that here $\frac{\partial \mathcal{L}(w,b)}{\partial w}$ and $\frac{\partial \mathcal{L}(w,b)}{\partial b}$ are the same as the previous case, and since $\xi_i \geq 0$, we need $\frac{\partial \mathcal{L}(w,b)}{\partial \xi_i} \leq 0$, which gives the additional constraint $\forall i, \alpha_i \leq C$.

However, for datasets that are not linearly separable, we need a non-linear decision boundary, which can be obtained by transforming the input vectors using a non-linear function ϕ. This gives us-

$$\text{Prediction} = \text{sign}\left(\sum_i \alpha_i^* y_i \phi(x_i)^T \phi(x) + b^*\right)$$

However, this leads to 2 problems:

1. How do we find $\phi(x)$?
2. It is computationally very expensive to enumerate $\phi(x)$ and compute the dot product $\phi(x_i)^T\phi(x)$ when the dimensionality of the transformed space is very high.

In order to solve these problems, we use the **kernel trick**.

3 Kernels to the Rescue

Definition of Kernels

Definition. *A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a **kernel** if there exists a function $\phi : \mathbb{R}^d \to \mathbb{R}^D$ such that $K(x, y) = \phi(x)^T\phi(y)$.*

So, why are Kernels useful? Well, note that the dual form of the SVM depends only on the inner products of the training instance pairs. So, if we can compute the inner products (that is, the kernel function) of the transformed vectors without explicitly computing the transformed vectors, we can use the dual form of the SVM without explicitly computing the transformed vectors. This is known as the **kernel trick**.

For example, consider the following example-

$$\phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}$$

The kernel function for this transformation is given by-

$$\begin{aligned} K(x, y) &= \phi(x)^T\phi(y) \\ &= x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 \\ &= (x_1y_1 + x_2y_2)^2 \\ &= (x^Ty)^2 \end{aligned}$$

Note that this can be computed without explicitly computing $\phi(x)$ and $\phi(y)$. For larger examples of $\phi(x)$, it may be infeasible (impossible when $\phi(x)$ has infinite dimensions) to explicitly find the transformed vectors and compute the dot product. However, we can still compute the kernel function $K(x, y)$.

Now a question arises are all functions kernels? The answer is no. We can use the following criteria to check if a function is a valid kernel:

1. A kernel K(X, Y) is a valid kernel if it corresponds to an underlying $\phi(X)$[Projecting X to a high, possibly infinite dimensional space]. That is, K(X, Y) = $\phi(X)^T\phi(Y)$.

   This criterion has limited use as it is difficult to check if there exists an underlying $\phi(X)$.

2. **Mercer's Theorem** : For $K : \mathbb{R}^d \to \mathbb{R}$ to be a valid kernel, for any finite set $\{x_1, ..., x_n\}$, the kernel matrix(which is an $n \times n$ matrix where the $(i, j)^{th}$ entry is K(x_i, x_j)) is symmetric and positive semi-definite (i.e. $v^TKv \geq 0\ \forall v \in R^n$). This condition is necessary and sufficient.

3. We can also use certain transformations of known kernels to show that a function is a valid kernel. If $K_1(x, y)$ and $K_2(x, y)$ are two valid kernel functions then

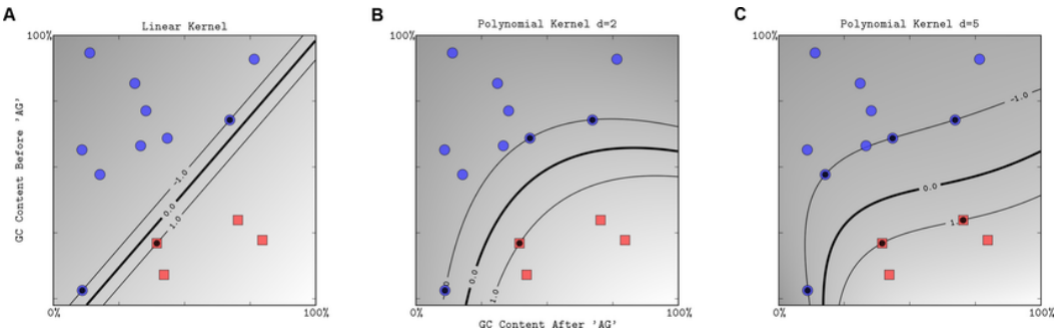
$$K(x, y) = K_1(x, y) + K_2(x, y)$$
$$K(x, y) = K_1(x, y) * K_2(x, y)$$
$$K(x, y) = c * K_1(x, y) \;\; where \;\; c > 0$$

are also valid Kernel functions.

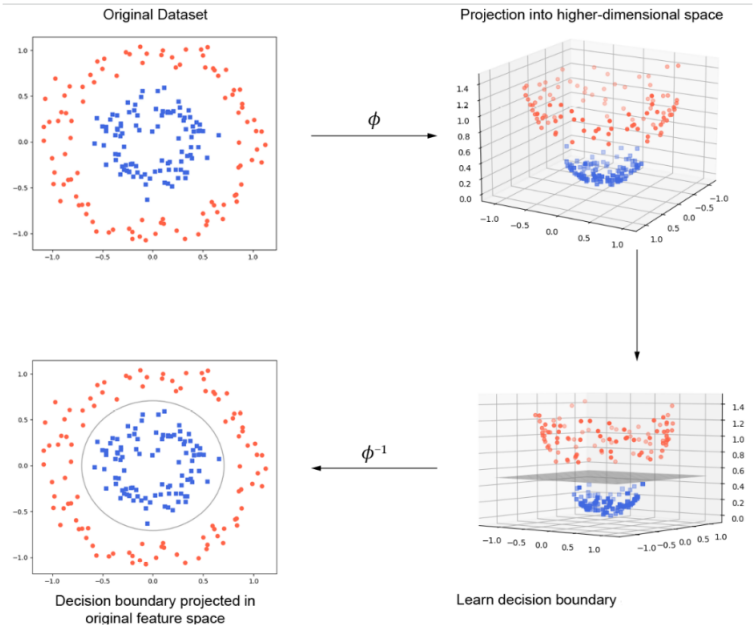
Some examples of valid kernels are-

1. Polynomial Kernel: $K(X, Y) = (X^T Y)^d$ and $K(X, Y) = (c + X^T Y)^d$, where $c > 0$ and d is a positive integer.

   This can be shown using the product of kernels property. Clearly, $K_0(X, Y) = X^T Y$ is a kernel, and hence, $K(X, Y) = (K_0(X, Y))^d = (X^T Y)^d$ is also a kernel. Also, $K_1(X, Y) = c$ using $\phi(X) = [\sqrt{c}]$, and hence using the sum of kernels property, $c + X^T Y$ is also a kernel.



2. Radial Basis Function(RBF) Kernel or Gaussian Kernel: $K(X, Y) = \exp\left(-\frac{\|X-Y\|^2}{2\sigma^2}\right)$



Practice Problem

Is the following function a valid kernel?

$$K(X, Y) = ||X + Y||^2$$

Answer: No. Suppose it is a valid kernel. Then, there exists a $\phi(X)$ such that $K(X, Y) = \phi(X)^T \phi(Y)$. Now, for $X = Y = 0$, we have-

$$\begin{aligned} & K(0, 0) = ||0 + 0||^2 = 0 \\ \implies & \phi(0)^T \phi(0) = 0 \\ \implies & \phi(0) = 0 \\ \implies & K(X, 0) = \phi(X)^T \phi(0) = 0 \quad \forall X \end{aligned}$$

which is a contradiction, since $K(X, 0) = ||X||^2$.