Book Exercise 6.7: Verify the results 6.17 and 6.18 for constructing valid kernels.

(6.17) Given 2 valid kernels $k_1(x,x')$, $k_2(x,x')$ then $K(x,x') = k_1(x,x') + k_2(x,x')$ is a valid kernel.

↳ let $k_1(x,x') = \phi_1(x)^T\phi_1(x')$ and $k_2(x,x') = \phi_2(x)^T\phi_2(x')$

We first show symmetry;

$$K(x,x') = k_1(x,x') + k_2(x,x') = k_1(x',x) + k_2(x',x) = K(x',x) \checkmark$$

Need ψ such that $K(x,x') = \psi(x)^T\psi(x') = \phi_1(x)^T\phi_1(x') + \phi_2(x)^T\phi_2(x')$

So, we can take $\psi(x) = \begin{bmatrix}\phi_1(x)\\ \phi_2(x)\end{bmatrix}$ and we see:

$$\psi(x)^T\psi(x') = [\phi_1(x)\ \phi_2(x)]\begin{bmatrix}\phi_1(x')\\ \phi_2(x')\end{bmatrix} = \phi_1(x)^T\phi_1(x') + \phi_2(x)^T\phi_2(x') = K(x,x') \checkmark$$

(6.18) $K(x,x') = k_1(x,x')k_2(x,x')$ is a valid kernel. Show it.

↳ let $k_1(x,x') = \phi(x)^T\phi(x')$, $k_2(x,x') = \phi'(x)^T\phi'(x')$. Then,

$$K(x,x') = [\phi(x)^T\phi(x')][\phi'(x)^T\phi'(x')] \Rightarrow \sum_{i,j}[\phi_i(x)\phi_j(x')][\phi_i'(x)\phi'(x')]$$

so we let $\psi(x) = (\phi_i(x)\phi_j'(x))_{i,j}$ and we see

$$\psi(x)^T\psi(x') = k_1(x,x')k_2(x,x') = k(x,x') \checkmark$$

7.2:

Show that if the 1 on the right hand side of constraint 7.5 is replaced with $\gamma > 0$, the solution for the maximum margin hyperplane is unchanged.

⇒ Then, the constraint would be: $t_n(w^T\phi(x_n)+b) \geq \gamma \quad n=1,\dots,N$

If we look at the case where equality holds, there will always be one point such that the constraint is active. Once the margin is maximized, there will be at least two points with the active constraint.

The maximization problem is then:

$$\underset{w,b}{\operatorname{argmax}}\left\{\frac{1}{\|w\|}\min_n\left(t_n(w^T\phi(x_n)+b)\right)\right\} \text{ subject to } t_n(w^T\phi(x_n)+b) \geq \gamma$$

We can see then that the maximization problem boils down to maximizing $\|w\|^{-1}$, which is equivalent to minimizing $\|w\|^2$.

So we can formulate the problem in LaGrangian terms as:

$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \{ t_n (w^T \phi(x_n) + b) - \gamma \}$ where the $\frac{1}{2}$ is for convenience and a_n are the Lagr. coeff.

① Take derivative wrt w:

$\hookrightarrow \|w\| - \sum_{n=1}^{N} a_n t_n \phi(x_n) = 0 \Rightarrow w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$ ✓

② Take derivative wrt b:

$\frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n t_n w^T \phi(x_n) + a_n t_n b - a_n \gamma \Rightarrow -\sum_{n=1}^{N} a_n t_n = 0$

$\Rightarrow \sum_{n=1}^{N} a_n t_n = 0$ ✓

Thus, we see the same result holds with $\gamma > 0$ instead of 1. $\blacksquare$

3. Given a dataset $D = \{(\mathbf{x}_n, t_n)\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^D$ and $t_n \in \{-1, 1\}$ for all n. The following is a formulation of soft-margin L_2-SVM, a variant of the standard SVM obtained by squaring the hinge loss:

$x_n, t_n \quad n = 1, \ldots, N \quad t_n \in \{-1, 1\}$

$$\underset{\mathbf{w}, b, \xi}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{n=1}^{N} \xi_n^2$$

$$\text{subj. to} \quad t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n \quad \forall n$$

$$\xi_n \geq 0 \quad \forall n$$

(a) Show that removing the last set of constraints $\{\xi_n \geq 0 \quad \forall n\}$ does not change the optimal solution to the problem above. Provide a complete proof.

(b) Describe the role of the hyperparameter $C \geq 0$. pg. 352

a) Pf: The last constraint $\xi_n \geq 0$ does not change the solution because of the definition/formulation of ξ_n.

$$\xi_n = \begin{cases} 0 & \text{if } x_n \text{ on correct side of margin} \\ |t_n - y(x_n)| & \text{otherwise} \end{cases}$$

Thus, we see that $\xi_n \geq 0 \ \forall n$ implicitly, because:

$\xi_n = 0 \geq 0$

OR $\xi_n = |t_n - y(x_n)| \geq 0$

and these two cases cover all $n \in [1, N]$.

Therefore, the optimization problem

$\operatorname{argmin} \frac{1}{2}\|w\|_2^2 + C \sum_{n=1}^{N} \xi_n^2$ subject to $t_n y(x_n) \geq 1 - \xi_n \ \forall n$ and $\xi_n \geq 0 \ \forall n$

is redundant in terms of the last explicit constraint, and so it can equivalently be rewritten as

$$\arg\min \frac{1}{2}\|w\|_2^2 + C\sum_{n=1}^{N} \varepsilon_n^2 \quad \text{subject to } t_n y(x_n) \geq 1-\varepsilon_n \ \forall n$$

and it will have the same result, since $\varepsilon_n \geq 0$ implicitly. ✓

b) The hyperparameter $C \geq 0$ controls the trade-off between the slack variable penalty and the margin. The hyperparameter C is similar to a regularization parameter in that it controls the model complexity and overfitting, but it works inversely to a regularization parameter. It is inverse in the fact that a higher C value correlates to a less overfitted model. When $C = 0$ we have hard margin SVM, and when $C \to \infty$ we recover the linearly separable case of SVM.