156 Assignment 4 solutions

Problem 1: Suppose $k(\mathbf{x}, \mathbf{x}')$ and $k'(\mathbf{x}, \mathbf{x}')$ are two kernels. Show that the following are also kernels.

(a) $k_1(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + k'(\mathbf{x}, \mathbf{x}')$.

(b) $k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')k'(\mathbf{x}, \mathbf{x}')$.

Solution. Write ϕ, ϕ' for the feature vectors defining k, k':

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \quad k'(\mathbf{x}, \mathbf{x}') = \phi'(\mathbf{x})^\top \phi'(\mathbf{x}').$$

(a): Let ψ be the feature vector

$$\psi(\mathbf{x}) = (\phi(\mathbf{x}), \phi'(\mathbf{x})).$$

That is, ψ is the concatenation of the two original feature vectors. Then,

$$\psi(\mathbf{x})^\top \psi(\mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}') + \phi'(\mathbf{x})^\top \phi'(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + k'(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}').$$

So k_1 is a kernel.

(b): Write

$$\psi(\mathbf{x}) = \{\phi_i(\mathbf{x})\phi'_j(\mathbf{x})\}_{i,j}.$$

Thus, ψ is the vector whose components are the pairwise products of the entries of $\phi(\mathbf{x})$ and $\phi'(\mathbf{x})$. Then, we compute

$$\begin{aligned}
\psi(\mathbf{x})^\top \psi(\mathbf{x}') &= \sum_{i,j} \psi_{i,j}(\mathbf{x})\psi_{i,j}(\mathbf{x}') \\
&= \sum_{i,j} \phi_i(\mathbf{x})\phi'_j(\mathbf{x})\phi_i(\mathbf{x}')\phi'_j(\mathbf{x}') \\
&= \left(\sum_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}')\right)\left(\sum_j \phi'_j(\mathbf{x})\phi'_j(\mathbf{x}')\right) \\
&= \phi(\mathbf{x})^\top \phi(\mathbf{x}')\phi'(\mathbf{x})^\top \phi'(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')k'(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

Thus the product $k_2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')k'(\mathbf{x}, \mathbf{x}')$ is a kernel, as claimed. □

Remark. It is also possible to solve this problem using the condition that the Gram matrices are positive semidefinite, though this is more difficult in case (b). We provide a solution for part (a), and defer to this Wikipedia page for a proof needed for part (b).

Alternate solution for (a). Take an arbitrary tuple of points $\{\mathbf{x}_n\}_{n=1}^N$; we need to demonstrate that the Gram matrix $\mathbb{K} = \big(k_1(\mathbf{x}_n, \mathbf{x}_m)\big)_{n,m=1}^N$ is positive semidefinite. For each vector $\mathbf{v} \in \mathbb{R}^N$,

$$\mathbf{v}^\top \mathbb{K} \mathbf{v} = \mathbf{v}^\top \big(\mathbb{K}^{(1)} + \mathbb{K}^{(2)}\big)\mathbf{v} = \mathbf{v}^\top \mathbb{K}^{(1)} \mathbf{v} + \mathbf{v}^\top \mathbb{K}^{(2)} \mathbf{v},$$

where $\mathbb{K}^{(1)}$ is the matrix $\big(k(\mathbf{x}_n, \mathbf{x}_n)\big)_{n,m=1}^N$ and $\mathbb{K}^{(2)}$ is the matrix $\big(k'(\mathbf{x}_n, \mathbf{x}_m)\big)_{n,m=1}^N$. Since $\mathbb{K}^{(1)}, \mathbb{K}^{(2)}$ are positive semidefinite by virtue of being Gram matrices of the kernels k, k', respectively, we have

$$\mathbf{v}^\top \mathbb{K}^{(1)} \mathbf{v} \geq 0, \quad \mathbf{v}^\top \mathbb{K}^{(2)} \mathbf{v} \geq 0.$$

Thus, we conclude

$$\mathbf{v}^\top \mathbb{K} \mathbf{v} \geq 0.$$

Since $\mathbf{v} \in \mathbb{R}^N$ was arbitrary, we conclude that $\mathbb{K}$ is positive semidefinite. Since the tuple $\{\mathbf{x}_n\}_{n=1}^N$ was arbitrary, we conclude that k_1 is a kernel, as claimed. $\square$

Problem 2: Show that, if the 1 on the right-hand side of the constraint

$$t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1$$

is replaced by some arbitrary constant $\gamma > 0$, the solution for the maximum margin hyperplane is unchanged.

Solution. We consider the modified problem in question:

$$\begin{aligned} &\underset{\mathbf{w},b}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{w}\|_2^2 \\ &\text{satisfying} \quad t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq \gamma \quad \forall n. \end{aligned} \tag{$\mathcal{P}_\gamma$}$$

Let $\mathbf{w}' = \frac{1}{\gamma}\mathbf{w}$, $b' = \frac{1}{\gamma}b$. Then, from the identities

$$\frac{1}{2}\|\mathbf{w}\|^2 = \gamma^2 \cdot \frac{1}{2}\|\mathbf{w}'\|_2^2,$$

$$t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) = t_n((\gamma\mathbf{w}')^\top \phi(\mathbf{x}_n) + \gamma b') = \gamma t_n((\mathbf{w}')^\top \phi(\mathbf{x}_n) + b'),$$

it follows that in these variables, $(\mathcal{P}_\gamma)$ takes the form

$$\begin{aligned} &\underset{\mathbf{w}',b'}{\operatorname{argmin}} \ \gamma^2 \cdot \frac{1}{2}\|\mathbf{w}'\|_2^2 \\ &\text{satisfying} \quad t_n((\mathbf{w}')^\top \phi(\mathbf{x}_n) + b') \geq 1 \quad \forall n. \end{aligned} \tag{0.1}$$

Since the γ^2 can be pulled outside the argmin in (0.1), we see that the minimization problem in $(\mathbf{w}', b')$ is identical to the minimization problem $(\mathcal{P}_1)$, i.e. with $\gamma = 1$.

Thus, if $(\mathbf{w}, b)$ solve $(\mathcal{P}_\gamma)$, then $(\mathbf{w}', b')$ solve $(\mathcal{P}_1)$, and vice versa. In particular, the solution to $(\mathcal{P}_\gamma)$ comes from solving the standard SVM problem $(\mathcal{P}_1)$ for $(\mathbf{w}', b')$ and setting $\mathbf{w} = \gamma\mathbf{w}'$, $b = \gamma b'$.

Finally, fix the two solutions above. Observe that the maximum-margin hyperplane corresponding to $(\mathcal{P}_\gamma)$ is the set

$$\mathbf{y}: \quad \mathbf{w}^\top \mathbf{y} + b = 0,$$

while the maximum-margin hyperplane corresponding to $(\mathcal{P}_1)$ is the set

$$\mathbf{y}: \quad (\mathbf{w}')^\top \mathbf{y} + b' = 0.$$

From the relations $\mathbf{w} = \gamma \mathbf{w}'$, $b = \gamma b'$, it follows that the two problems define the same maximum-margin hyperplane, as claimed.

□

Problem 3: Take a dataset $D = \big\{(\mathbf{x}_n, t_n)\big\}_{n=1}^N$ 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:

$$\begin{aligned}
\underset{\mathbf{w}, b, \xi}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{n=1}^N \xi_n^2 \\
\text{subj. to} \quad & t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n \quad \forall n \\
& \xi_n \geq 0 \quad \forall n
\end{aligned}$$

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

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

Solution. (a): Suppose that we have a choice of parameters $(\mathbf{w}, b, \xi)$ satisfying the constraints

$$t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n \quad \forall n,$$

but not necessarily the second set of constraints. We define a new set of parameters:

$$\begin{cases}
\mathbf{w}' = \mathbf{w} \\
b' = b \\
\xi_n' = \max(\xi_n, 0).
\end{cases}$$

We claim that $(\mathbf{w}', b', \xi')$ satisfy the full set of original constraints, and that

$$E(\mathbf{w}', b', \xi') \leq E(\mathbf{w}, b, \xi),$$

where E is the loss function in question:

$$E(\mathbf{w}, b, \xi) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{n=1}^N \xi_n^2.$$

We first check that $(\mathbf{w}', b', \xi')$ satisfy the constraints. For each n, note that

$$\xi_n' = \max(\xi_n, 0) \geq 0,$$

so this constraint is fulfilled. Also, note that

$$\xi_n' = \max(\xi_n, 0) \geq \xi_n,$$

so that

$$1 - \xi_n \geq 1 - \xi_n',$$

and hence

$$t_n((\mathbf{w}')^\top \phi(\mathbf{x}_n) + b') = t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n \geq 1 - \xi_n'.$$

Here we have used the fact that we assumed $t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n$. Thus, $(\mathbf{w}', b', \xi')$ satisfies all the original constraints.

Next, we check the inequality in the cost function. We claim that, for each $1 \leq n \leq N$,

$$(\xi_n')^2 \leq \xi_n^2 \tag{0.2}$$

We demonstrate this now. Fix some particular n. If $\xi_n \geq 0$, then $\xi_n' = \xi_n$, and thus (0.2) holds. If instead $\xi_n < 0$, then $\xi_n' = 0$ and $\xi_n^2 > 0$, so certainly $(\xi_n')^2 \leq \xi_n^2$. Thus, (0.2) holds for each n, and in particular

$$\sum_{n=1}^{N} (\xi_n')^2 \leq \sum_{n=1}^{N} \xi_n^2.$$

From this, it immediately follows that

$$E(\mathbf{w}', b', \xi') \leq E(\mathbf{w}, b, \xi).$$

Thus, we have demonstrated each of our claims. In particular, if we have an optimal solution $(\mathbf{w}, b, \xi)$ to the problem without the constraints $\{\xi_n \geq 0\}_n$, then the new tuple $(\mathbf{w}', b', \xi')$ is a better solution that also satisfies the constraints $\{\xi_n \geq 0\}_n$. Thus, removing those constraints does not alter the solution to the minimization problem.

(b): C governs the trade-off between minimizing the parameters ξ_n and the parameter $\|\mathbf{w}\|_2$. If C is very large, then a typical choice of $\mathbf{w}, b, \xi$ will have the property that

$$\frac{1}{2}\|\mathbf{w}\|_2^2 \ll C \sum_{n=1}^{N} \xi_n^2.$$

Thus, in this setting, it is more impactful to decrease the ξ_n's than to decrease the w_n's. Conversely, if $C \sim 0$, then for typical choice of $\mathbf{w}, b, \xi$ we have

$$\frac{1}{2}\|\mathbf{w}\|_2^2 \gg C \sum_{n=1}^{N} \xi_n^2.$$

Thus, it will make more sense to choose the parameter $\mathbf{w}$ to have small norm than to choose ξ to have small norm. Overall, C governs the relative size of $\mathbf{w}, \xi$ in the optimal solution to the soft-margin SVM. $\square$