4033/5033 Assignment: Bonus-Point Problem

Sasank Sribhashyam

Task 1. Let k_1 and k_2 be two valid kernel functions. Prove $\alpha_1 k_1 + \alpha_2 k_2$ is also a valid kernel, where α_1, α_2 are two constants. (Tip: you may first prove $\alpha_1 k_1$ and $\alpha_2 k_2$ are each valid then their sum is valid.)

To prove that $\alpha_1 k_1 + \alpha_2 k_2$ is also a valid kernel function when α_1 and α_2 are constants and k_1 and k_2 are valid kernel functions, we can use Mercer's theorem. Mercer's theorem provides a necessary and sufficient condition for a function to be a valid kernel.

Mercer's theorem states that a function $K(x, y)$ is a valid kernel if and only if it satisfies the following two conditions:

1. Symmetry: $K(x, y) = K(y, x)$ for all x and y.
2. Positive semi-definiteness: For any finite set of points $\{x_1, x_2, \ldots, x_n\}$ and any set of real numbers $\{c_1, c_2, \ldots, c_n\}$, the kernel matrix $K_{ij} = K(x_i, x_j)$ is positive semi-definite. In other words, for any set of n points and any set of n real numbers, the matrix $\left[c_1, c_2, \ldots, c_n\right] K \left[c_1, c_2, \ldots, c_n\right]^T$ (the outer product of the coefficient vector and the kernel matrix) should be non-negative.

Now, let's prove that $\alpha_1 k_1$ and $\alpha_2 k_2$ are each valid kernels:

1. $\alpha_1 k_1(x, y)$ is a valid kernel:
- Symmetry: $\alpha_1 k_1(x, y) = \alpha_1 k_1(y, x)$ since multiplication by a constant does not affect symmetry. - Positive semi-definiteness: Let $\{x_1, x_2, \ldots, x_n\}$ be a finite set of points and $\{c_1, c_2, \ldots, c_n\}$ be a set of real numbers. The kernel matrix for $\alpha_1 k_1$ is given by $K_{ij} = \alpha_1 k_1(x_i, x_j)$. Now, consider the matrix $\left[c_1, c_2, \ldots, c_n\right] K \left[c_1, c_2, \ldots, c_n\right]^T$:

$$\left[c_1, c_2, \ldots, c_n\right] \begin{bmatrix} \alpha_1 k_1(x_1, x_1), \alpha_1 k_1(x_1, x_2), \ldots, \alpha_1 k_1(x_1, x_n) \\ \alpha_1 k_1(x_2, x_1), \alpha_1 k_1(x_2, x_2), \ldots, \alpha_1 k_1(x_2, x_n) \\ \vdots \\ \alpha_1 k_1(x_n, x_1), \alpha_1 k_1(x_n, x_2), \ldots, \alpha_1 k_1(x_n, x_n) \end{bmatrix}$$

$$\left[c_1, c_2, \ldots, c_n\right] \begin{bmatrix} \alpha_1 c_1 k_1(x_1, x_1), \alpha_1 c_1 k_1(x_1, x_2), \ldots, \alpha_1 c_1 k_1(x_1, x_n) \\ \alpha_1 c_2 k_1(x_2, x_1), \alpha_1 c_2 k_1(x_2, x_2), \ldots, \alpha_1 c_2 k_1(x_2, x_n) \\ \vdots \\ \alpha_1 c_n k_1(x_n, x_1), \alpha_1 c_n k_1(x_n, x_2), \ldots, \alpha_1 c_n k_1(x_n, x_n) \end{bmatrix}$$

Since k_1 is a valid kernel, we know that the resulting matrix is positive semi-definite because it is the product of a valid kernel matrix and a positive semi-definite diagonal matrix (each element on the diagonal is the square of a real number). Therefore, $\alpha_1 k_1(x, y)$ is a valid kernel.

2. $\alpha_2 k_2(x, y)$ is a valid kernel:
- Symmetry: $\alpha_2 k_2(x, y) = \alpha_2 k_2(y, x)$ since multiplication by a constant does not affect symmetry. - Positive semi-definiteness: Similarly to the previous case, $\alpha_2 k_2(x, y)$ is also a valid kernel by the same reasoning.

Now, we can prove that $\alpha_1 k_1 + \alpha_2 k_2$ is a valid kernel:

1. Symmetry: Since both $\alpha_1 k_1$ and $\alpha_2 k_2$ are symmetric kernels, their sum $\alpha_1 k_1 + \alpha_2 k_2$ is also symmetric.

2. Positive semi-definiteness: Let $\{x_1, x_2, \ldots, x_n\}$ be a finite set of points, and $\{c_1, c_2, \ldots, c_n\}$ be a set of real numbers. The kernel matrix for $\alpha_1 k_1 + \alpha_2 k_2$ is given by $K_{ij} = (\alpha_1 k_1(x_i, x_j) + \alpha_2 k_2(x_i, x_j))$. Now, consider the matrix $\left[c_1, c_2, \ldots, c_n\right] K \left[c_1, c_2, \ldots, c_n\right]^T$:

$$\left[c_1, c_2, \ldots, c_n\right] \begin{bmatrix} (\alpha_1 k_1(x_1,x_1)+\alpha_2 k_2(x_1,x_1)), (\alpha_1 k_1(x_1,x_2)+\alpha_2 k_2(x_1,x_2)), \ldots, (\alpha_1 k_1(x_1,x_n)+\alpha_2 k_2(x_1,x_n)) \\ (\alpha_1 k_1(x_2,x_1)+\alpha_2 k_2(x_2,x_1)), (\alpha_1 k_1(x_2,x_2)+\alpha_2 k_2(x_2,x_2)), \ldots, (\alpha_1 k_1(x_2,x_n)+\alpha_2 k_2(x_2,x_n)) \\ \vdots \\ (\alpha_1 k_1(x_n,x_1)+\alpha_2 k_2(x_n,x_1)), (\alpha_1 k_1(x_n,x_2)+\alpha_2 k_2(x_n,x_2)), \ldots, (\alpha_1 k_1(x_n,x_n)+\alpha_2 k_2(x_n,x_n)) \end{bmatrix}$$

$$\left[c_1, c_2, \ldots, c_n\right] \left[\begin{matrix} (\alpha_1 c_1 k_1(x_1,x_1)+\alpha_2 c_1 k_2(x_1,x_1)), (\alpha_1 c_1 k_1(x_1,x_2)+\alpha_2 c_1 k_2(x_1,x_2)), \ldots, (\alpha_1 c_1 k_1(x_1,x_n)+\alpha_2 c_1 k_2(x_1,x_n)) \\ (\alpha_1 c_2 k_1(x_2,x_1)+\alpha_2 c_2 k_2(x_2,x_1)), (\alpha_1 c_2 k_1(x_2,x_2)+\alpha_2 c_2 k_2(x_2,x_2)), \ldots, (\alpha_1 c_2 k_1(x_2,x_n)+\alpha_2 c_2 k_2(x_2,x_n)) \\ \vdots \\ (\alpha_1 c_n k_1(x_n,x_1)+\alpha_2 c_n k_2(x_n,x_1)), (\alpha_1 c_n k_1(x_n,x_2)+\alpha_2 c_n k_2(x_n,x_2)), \ldots, (\alpha_1 c_n k_1(x_n,x_n)+\alpha_2 c_n k_2(x_n,x_n)) \end{matrix}\right.$$

Now, since both $\alpha_1 k_1$ and $\alpha_2 k_2$ are valid kernels (by our earlier proofs), we know that their respective matrices are positive semi-definite. Therefore, the resulting matrix is also positive semi-definite because it is a linear combination of two positive semi-definite matrices (each scaled by a positive constant α_1 or α_2).

Therefore, $\alpha_1 k_1 + \alpha_2 k_2$ is a valid kernel function.

Task 2. Derive the bias-variance trade-off based on the following setting. Let x be a fixed instance (a constant) and y be its random label generated through $y = f_*(x) + \varepsilon$, where f_* is our target function and ε is a random noise with $\mathbb{E}_\varepsilon(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Let f be a function we learned from a random training set (and thus is random by itself). Define its expected prediction error on the fixed instance x (with random label y) as

$$\bar{er}(f) = \mathbb{E}_{f,\varepsilon}(f(x) - y)^2, \tag{1}$$

where the expectation is taken over the randomness of both f and ε. Please show $\bar{er}(f)$ can be decomposed into the sum of three terms. The first term should depend on (but not necessarily exactly the same as) the bias of f, defined as

$$\mathrm{bias}(f) = |\mathbb{E}_f f(x) - f_*(x)|. \tag{2}$$

The second term should depend on the variance of f, defined as

$$\mathrm{Var}(f) = \mathbb{E}_f(f(x) - \mathbb{E}_f f(x))^2. \tag{3}$$

The last term should depend on data noise i.e., σ^2.

To derive the bias-variance trade-off for the expected prediction error $\bar{er}(f)$, we'll start with the definition of $\bar{er}(f)$:

$$\bar{er}(f) = \mathbb{E}_{f,\epsilon}[(f(x) - y)^2] \tag{4}$$

Now, we want to decompose this into three terms: bias, variance, and data noise. Let's do it step by step, keeping the expectations clear:

1. Expand the square inside the expectation:

$$\bar{er}(f) = \mathbb{E}_{f,\epsilon}[f^2(x) - 2f(x)y + y^2] \tag{5}$$

2. Now, split this expectation into three terms:

$$\bar{er}(f) = \mathbb{E}_{f,\epsilon}[f^2(x)] - 2\mathbb{E}_{f,\epsilon}[f(x)y] + \mathbb{E}_{f,\epsilon}[y^2] \tag{6}$$

a. First Term: Bias The bias term should be calculated as:

$$\text{Bias}(f) = \mathbb{E}_f \left[(f(x) - f^*(x))^2 \right] \tag{7}$$

Expanding this term inside the expectation:

$$\begin{align} \mathbb{E}_{f,\epsilon} \left[f^2(x) \right] &= \mathbb{E}_f \left[(f(x) - f^*(x) + f^*(x))^2 \right] \tag{8} \\ &= \mathbb{E}_f \left[(f(x) - f^*(x))^2 + 2(f(x) - f^*(x))f^*(x) + f^*(x)^2 \right] \tag{9} \end{align}$$

Now, take the expectation $\mathbb{E}_{f,\epsilon}$ of this expression:

$$\mathbb{E}_{f,\epsilon} \left[f^2(x) \right] = \mathbb{E}_f \left[(f(x) - f^*(x))^2 \right] + 2\mathbb{E}_f \left[(f(x) - f^*(x))f^*(x) \right] + \mathbb{E}_{f,\epsilon} \left[f^*(x)^2 \right] \tag{10}$$

The first term on the right-hand side represents the squared bias, which we denote as $\text{Bias}^2(f)$.

b. Second Term: Variance For the variance term:

$$\text{Var}(f) = \mathbb{E}_f \left[(f(x) - \mathbb{E}_f[f(x)])^2 \right] \tag{11}$$

Expanding this term inside the expectation:

$$\begin{align} \mathbb{E}_{f,\epsilon} \left[f(x)y \right] &= \mathbb{E}_f \left[f(x)(f^*(x) + \epsilon) \right] \tag{12} \\ &= \mathbb{E}_f \left[f(x)f^*(x) \right] + \mathbb{E}_f \left[f(x)\epsilon \right] \tag{13} \end{align}$$

Now, take the expectation $\mathbb{E}_{f,\epsilon}$ of this expression:

$$\mathbb{E}_{f,\epsilon} \left[f(x)y \right] = \mathbb{E}_f \left[f(x)f^*(x) \right] + \mathbb{E}_f \left[f(x)\epsilon \right] \tag{14}$$

The first term on the right-hand side represents the expected value of the product of $f(x)$ and the true function $f^*(x)$, which we denote as $\mathbb{E}_f[f(x)f^*(x)]$.

c. Third Term: Data Noise For the data noise term:

$$\mathbb{E}_{f,\epsilon} \left[y^2 \right] = \mathbb{E}_f \left[(f^*(x) + \epsilon)^2 \right] = \mathbb{E}_f \left[f^*(x)^2 + 2f^*(x)\epsilon + \epsilon^2 \right] \tag{15}$$

Now, take the expectation $\mathbb{E}_{f,\epsilon}$ of this expression:

$$\mathbb{E}_{f,\epsilon} \left[y^2 \right] = \mathbb{E}_f \left[f^*(x)^2 \right] + 2\mathbb{E}_f \left[f^*(x)\epsilon \right] + \mathbb{E}_{f,\epsilon} \left[\epsilon^2 \right] \tag{16}$$

The last term on the right-hand side represents the variance of the data noise, which is σ^2.

Now, combining these three terms:

$$\overline{er}(f) = \text{Bias}^2(f) + \text{Var}(f) + \sigma^2 \tag{17}$$

So, we have successfully decomposed the expected prediction error $\bar{er}(f)$ into the sum of three terms: the squared bias, the variance, and the data noise. This is the bias-variance trade-off, which shows that as you try to reduce bias (by making your model more complex), you often increase variance, and vice versa. The goal is to find the right balance to minimize the expected prediction error.