# CS5691 - Pattern Recognition and Machine Learning
# Jul – Nov, 2023

## “Practice Worksheet: SVM, Kernels & Logistic Regression”

November 17, 2023

**This worksheet is provided for practice with SVM, Kernels and Logistic Regression only. Students are requested to kindly refer to the Mid Semester and Quiz Worksheets for practice with the other topics included in the End Semester Examination.**

1. Consider a SVM Hard Margin problem where the decision boundary is defined by $z(x) = 0$ where $z(x) := w^T x + b$.
   (i) Derive an expression for the (Euclidean) distance between the margins (margin boundaries).
   (ii) What is the distance of the origin $(0, 0)$ to the decision boundary $z(x) = 0$?

**Solution:**

(i) $\frac{2}{\|w\|}$ (Derivation not shown)

(ii) $\frac{b}{\|w\|}$ (Derivation not shown)

2. Consider a soft margin SVM where C is the penalty parameter. Explain how the behavior of SVM as a classifier will change as C is increased from a very small value to a very high value.

3. Let $u \in \mathbf{R}^d$ be a point. Let $w \in \mathbf{R}^d, b \in \mathbf{R}$ and the hyperplane given by $w, b$ is $\{x \in \mathbf{R}^d : w^T x + b = 0\}$. Consider the following problem of projection of the point $u$ on to a (hyper)plane given by $w, b$.

$$\min_{v \in \mathbf{R}^d} \frac{1}{2} \|v - u\|^2$$

$$\text{s.t. } w^T v + b = 0$$

   Derive the solution to the above problem via solving the Lagrangian dual (which is an unconstrained quadratic problem, and hence can be easily solved; reviewing the minimax theorem and KKT conditions seen in class may help). Then show that the distance of the point u to the hyperplane given by $w, b$ is $\frac{|w^T u + b|}{\|w\|}$.

4. Let $\{(x_1, y_1), ..., (x_n, y_n)\}$ be a linearly separable binary classification dataset. Let $w^*, b^*$ be any solution to the problem below:

$$\max_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{\|w\|}$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1$$

   Show that $\min_{i \in [n]} y_i(w^T x_i + b) = 1$.

5. Consider a soft margin SVM problem with $C$ set to some constant. Let $\alpha^*$ be the dual solution, and let $w^*, b^*$ be the primal solution. Let the dataset be $(x_i, y_i)$ with $i$ ranging from 1 to $n$.
   (i) If $\alpha^* = 0$, what are the possible range of values of $(w^*)^T x_i + b^*$?

(ii) If $0 < \alpha^* < C$, what are the possible range of values of $(w^*)^T x_i + b^*$?

(iii) If $\alpha^* = C$, what are the possible range of values of $(w^*)^T x_i + b^*$?

(Hint: Use KKT complementary slack conditions and $\beta_i^* = C - \alpha_i^*$)

6. Consider the following 1-dimensional classification dataset with 8 points given by:

$$X^T = \begin{bmatrix} 1 & 2 & 4 & 5 & 6 & 7 & 9 & 10 \end{bmatrix}$$

$$y^T = \begin{bmatrix} +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \end{bmatrix}$$

(i) Solve Kernel SVM (Hard Margin) (i.e. give the optimal $\alpha^*$) with $K(u, v) = exp(-\gamma(u - v)^2)$ where $\gamma = 0.1$.
Hint: $\alpha_2^* = \alpha_7^* = \alpha_3^* = \alpha_6^* > 0$ while all other $\alpha_i^* = 0$.
Find such an $\alpha^*$ and then use KKT conditions to show that it is optimal.

(ii) Give $b^*$ for all $\alpha^*$ above.

(iii) Give the decision function $(w^*)^T \phi(x) + b^*$ for $x \in \mathbf{R}$.

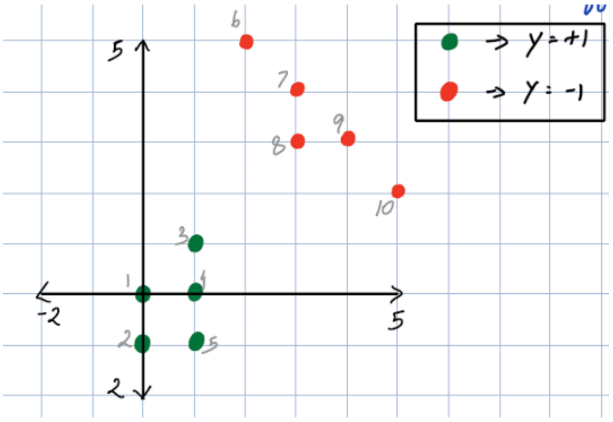
7. Consider the following hard margin SVM problem with both $w$ and $b$. Assume the kernel to be the linear kernel.

(i) Argue what points are support vectors.

(ii) Argue what would be the optimal hyperplane and give $w^*, b^*$.

(iii) Also argue what $\alpha^*$ should be.

(You can use software to check intuition, and use KKT conditions to verify if a proposed $\alpha^*$ is actually an optimal solution.)



(iv) Repeat parts (i), (ii), (iii) if point $(x_8, y_8)$ is removed.

(v) Repeat parts (i), (ii), (iii) with point $(x_8, y_8), (x_9, y_9), (x_7, x_7)$ removed.

(Hint: Optimal $\alpha^*$ need not be unique even if $w^*$ and $b^*$ are.)

8. Consider the following 2-dimensional classification dataset with 5 points given by:

$$X^T = \begin{bmatrix} 1 & 1 & 2 & 4 & 5 \\ 1 & 0 & 5 & 4 & 2 \end{bmatrix}$$

$$y^T = \begin{bmatrix} +1 & +1 & -1 & -1 & -1 \end{bmatrix}$$

Consider the hard margin SVM problem with linear kernel $k(u, v) = u^T v$.

(i) Give the support vectors just by looking at the data. Give reasons.

(ii) Give the dual solution $\alpha^*$ using the answer to the above part.

(iii) Check if the entire solution got above is the right answer using KKT conditions. (Thus also checking the first part guessed by "eyeballing".)

(iv) Derive the primal solution $w^*, b^*$ from the dual solution $\alpha^*$ and draw a figure illustrating the final solution.

9. Consider the following 2-dimensional binary classification dataset with 10 points given by

$$X^T = \begin{bmatrix} 1 & 1 & 2 & 2 & 4 & 4 & 5 & 5 & 2.9 & 3.1 \\ 0 & 1 & 0 & 1 & 3 & 4 & 3 & 4 & 6 & 6 \end{bmatrix}$$

$$y^T = \begin{bmatrix} -1 & -1 & -1 & -1 & +1 & +1 & +1 & +1 & -1 & +1 \end{bmatrix}$$

Consider the soft-margin linear SVM problem with $C = 0.1, 1, 10, 100$. For each $C$ evaluate the following $w, b$. By evaluate, we mean you should give the slack variable $\xi$ that make the $w, b, \xi$ feasible, and also give the value of the objective.
(i) $w = (\frac{1}{2}, 0), b = \frac{-3}{2}$ (ii) $w = (1, 0), b = -3$ (iii) $w = (4, 0), b = -12$ (iv) $w = (16, 0), b = -48$
(v) $w = (64, 0), b = -192$ (vi) $w = (\frac{1}{4}, \frac{1}{4}), b = \frac{-5}{4}$ (vii) $w = (\frac{1}{2}, \frac{1}{2}), b = \frac{-5}{2}$ (viii) $w = (1, 1), b = -5$
(ix) $w = (2, 2), b = -10$ (x) $w = (4, 4), b = -20$

**Solution:** To obtain the lowest $\xi_i$ possible, solve for $\xi_i$ in 1 - $y_i(w^T x_i + b)$.

10. (i) Why is logistic regression called regression?

(ii) Consider the following 2-dimensional classification dataset with 8 points given by:

$$X^T = \begin{bmatrix} -2 & -2 & -1 & -1 & 1 & 1 & 2 & 3 \\ -1 & 2 & 1 & 2 & 1 & 3 & 3 & 2 \end{bmatrix}$$

$$y^T = \begin{bmatrix} +1 & +1 & +1 & -1 & -1 & +1 & -1 & -1 \end{bmatrix}$$

Run one iteration of gradient descent with the logistic regression objective by hand. No bias required, only the 2-dimensional weight vector is to be optimised. Choose the step size $\eta = 1$. Initialise at $w = [0, 0]^T$.

11. Recall that the Empirical Logistic Loss Minimisation is given by

$$\hat{R}(w) = \sum_{i=1}^{n} log(1 + exp(-y_i w^T x_i)) = \sum_{i=1}^{n} \Psi_L(y_i w^T x_i)$$

Prove that $\Psi_L : \mathbf{R} \to \mathbf{R}$ is convex. Specifically prove that $\Psi_L(u) = log(1 + exp(-u))$ is convex in $u$. Subsequently argue that $\hat{R}(w)$ is convex in $w$.

12. Let the data instance $X$ be a d-dimensional vector. A function $K : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$ is valid kernel function if there exists $\phi : \mathbf{R}^d \to \mathbf{R}^{d'}$ such that $K(u, v) = \phi(u)^T \phi(v)$. Such a $\phi$ is called a feature map for the kernel $K$.

(i) Let $d = 2, k = 2$. Prove that $K(u, v) = (u^T v)^k$ is a valid kernel. Give the feature map corresponding to this kernel.

(ii) Repeat the above for $d = 3, k = 2$.

(iii) Repeat the above for $d = 2, k = 3$.

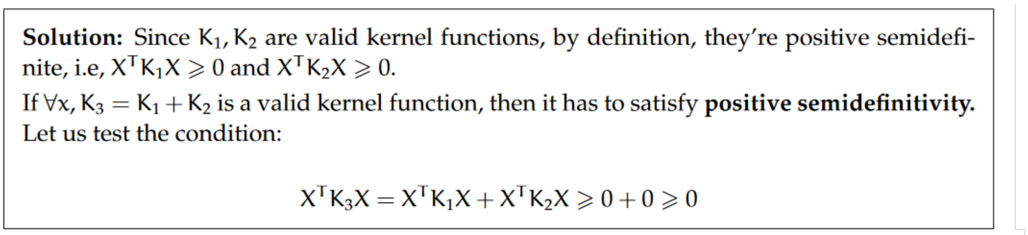
(iv) Infer the general form of the feature map $\phi$ of the kernel $K : (u, v) \to (u^T v)^k$ for any $d, k$.

(v) Repeat the four items above for the kernel $K(u, v) = (1 + u^T v)^k$.

13. Let $K_1$ and $K_2$ be a valid kernel functions, with feature mapping $\varphi_1 : \mathbb{R}^d \to \mathbb{R}^{d_1}$ and $\varphi_2 : \mathbb{R}^d \to \mathbb{R}^{d_2}$.
*Solutions by Siddharth D P (EE18B072) from the earlier offering of PRML.*

(i) Show that $K_3 = K_1 + K_2$ is also a valid kernel. Give the feature mapping $\varphi_3$ corresponding to $K_3$ in terms of $\varphi_1$ and $\varphi_2$.

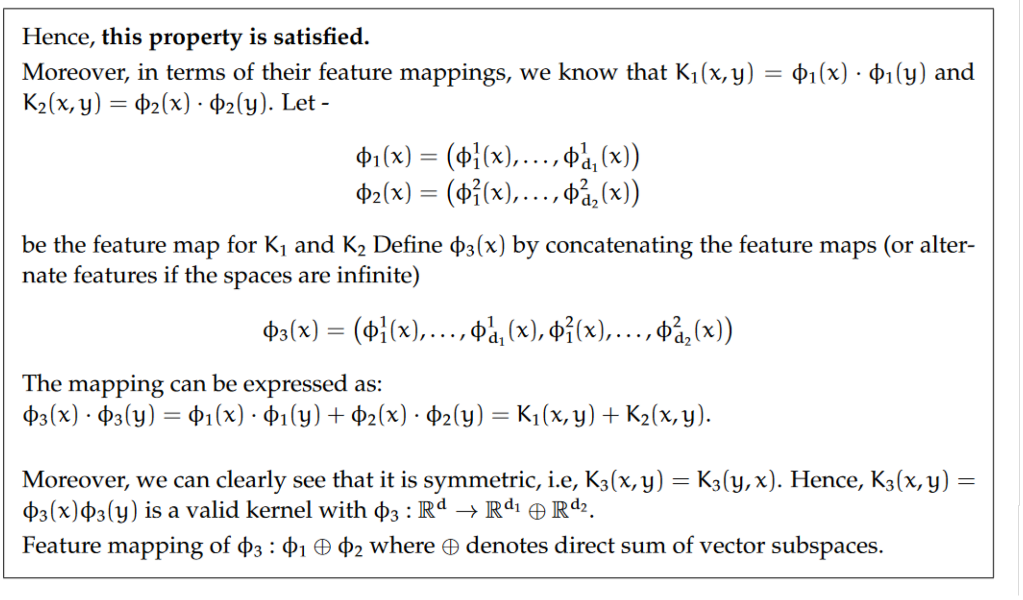
**Solution:**



**Solution:** Since $K_1, K_2$ are valid kernel functions, by definition, they're positive semidefinite, i.e, $X^T K_1 X \geqslant 0$ and $X^T K_2 X \geqslant 0$.

If $\forall x, K_3 = K_1 + K_2$ is a valid kernel function, then it has to satisfy **positive semidefinitivity.** Let us test the condition:

$$X^T K_3 X = X^T K_1 X + X^T K_2 X \geqslant 0 + 0 \geqslant 0$$



Hence, **this property is satisfied.**

Moreover, in terms of their feature mappings, we know that $K_1(x, y) = \phi_1(x) \cdot \phi_1(y)$ and $K_2(x, y) = \phi_2(x) \cdot \phi_2(y)$. Let -

$$\phi_1(x) = \left(\phi_1^1(x), \ldots, \phi_{d_1}^1(x)\right)$$
$$\phi_2(x) = \left(\phi_1^2(x), \ldots, \phi_{d_2}^2(x)\right)$$

be the feature map for $K_1$ and $K_2$ Define $\phi_3(x)$ by concatenating the feature maps (or alternate features if the spaces are infinite)

$$\phi_3(x) = \left(\phi_1^1(x), \ldots, \phi_{d_1}^1(x), \phi_1^2(x), \ldots, \phi_{d_2}^2(x)\right)$$

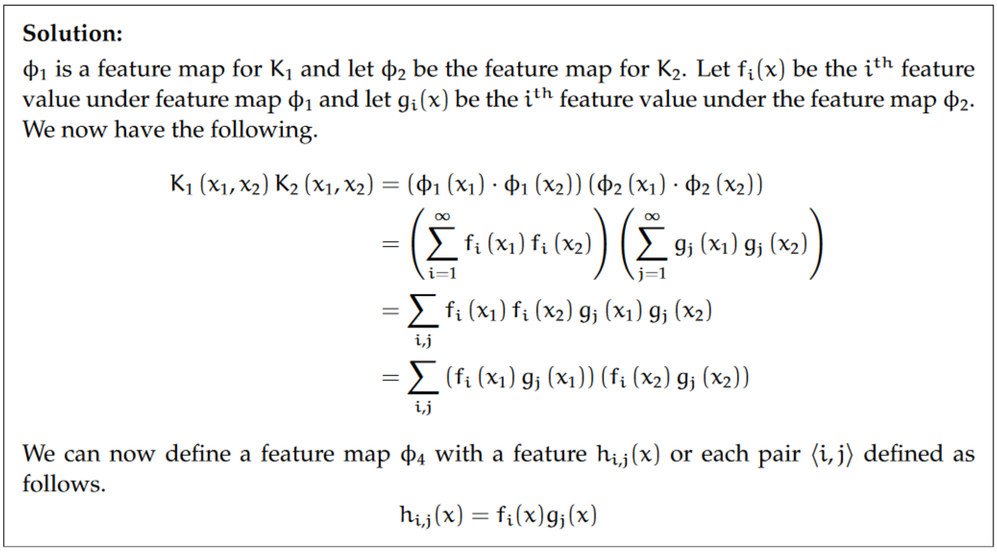
The mapping can be expressed as:
$\phi_3(x) \cdot \phi_3(y) = \phi_1(x) \cdot \phi_1(y) + \phi_2(x) \cdot \phi_2(y) = K_1(x, y) + K_2(x, y)$.

Moreover, we can clearly see that it is symmetric, i.e, $K_3(x, y) = K_3(y, x)$. Hence, $K_3(x, y) = \phi_3(x)\phi_3(y)$ is a valid kernel with $\phi_3 : \mathbb{R}^d \to \mathbb{R}^{d_1} \oplus \mathbb{R}^{d_2}$.

Feature mapping of $\phi_3 : \phi_1 \oplus \phi_2$ where $\oplus$ denotes direct sum of vector subspaces.

(ii) Show that $K_4 = K_1 \cdot K_2$ is also a valid kernel. Give the feature mapping $\varphi_4$ corresponding to $K_4$ in terms of $\varphi_1$ and $\varphi_2$.

**Solution:**



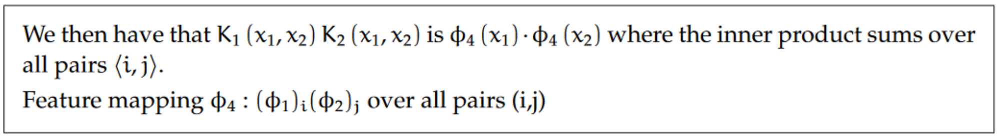
**Solution:**

$\phi_1$ is a feature map for $K_1$ and let $\phi_2$ be the feature map for $K_2$. Let $f_i(x)$ be the $i^{th}$ feature value under feature map $\phi_1$ and let $g_i(x)$ be the $i^{th}$ feature value under the feature map $\phi_2$. We now have the following.

$$\begin{aligned}K_1(x_1,x_2)K_2(x_1,x_2) &= (\phi_1(x_1)\cdot\phi_1(x_2))(\phi_2(x_1)\cdot\phi_2(x_2))\\ &= \left(\sum_{i=1}^{\infty} f_i(x_1)f_i(x_2)\right)\left(\sum_{j=1}^{\infty} g_j(x_1)g_j(x_2)\right)\\ &= \sum_{i,j} f_i(x_1)f_i(x_2)g_j(x_1)g_j(x_2)\\ &= \sum_{i,j}(f_i(x_1)g_j(x_1))(f_i(x_2)g_j(x_2))\end{aligned}$$

We can now define a feature map $\phi_4$ with a feature $h_{i,j}(x)$ or each pair $\langle i,j\rangle$ defined as follows.

$$h_{i,j}(x) = f_i(x)g_j(x)$$

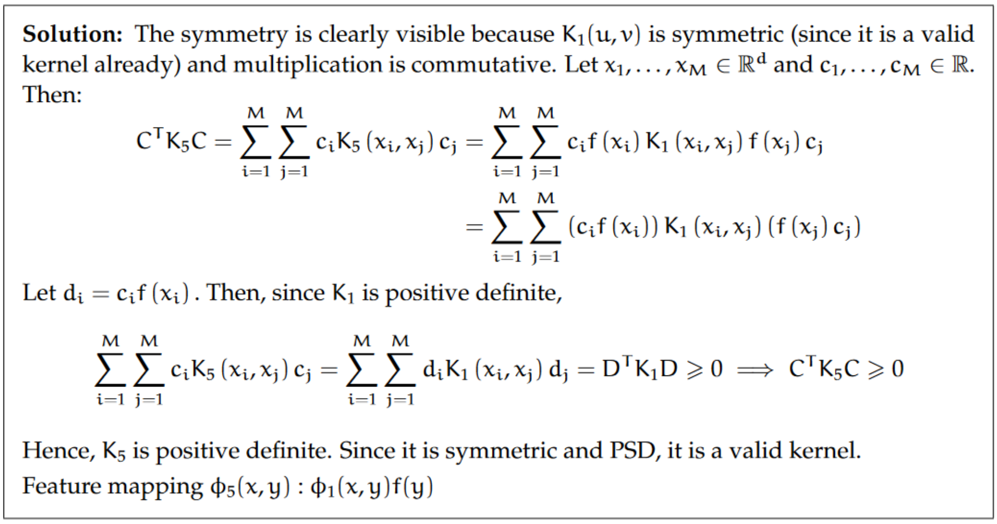


We then have that $K_1(x_1,x_2)K_2(x_1,x_2)$ is $\phi_4(x_1)\cdot\phi_4(x_2)$ where the inner product sums over all pairs $\langle i,j\rangle$.

Feature mapping $\phi_4 : (\phi_1)_i(\phi_2)_j$ over all pairs (i,j)

(iii) Show that $K_5 = f(u)K_1(u,v)f(v)$ is also a valid kernel for any function $f:\mathbb{R}^d\to\mathbb{R}$. Give the feature mapping $\varphi_5$ corresponding to $K_5$ in terms of $\varphi_1$ and $f$.

**Solution:**



**Solution:** The symmetry is clearly visible because $K_1(u,v)$ is symmetric (since it is a valid kernel already) and multiplication is commutative. Let $x_1,\ldots,x_M\in\mathbb{R}^d$ and $c_1,\ldots,c_M\in\mathbb{R}$. Then:

$$\begin{aligned}C^TK_5C = \sum_{i=1}^{M}\sum_{j=1}^{M} c_iK_5(x_i,x_j)c_j &= \sum_{i=1}^{M}\sum_{j=1}^{M} c_if(x_i)K_1(x_i,x_j)f(x_j)c_j\\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}(c_if(x_i))K_1(x_i,x_j)(f(x_j)c_j)\end{aligned}$$

Let $d_i = c_if(x_i)$. Then, since $K_1$ is positive definite,

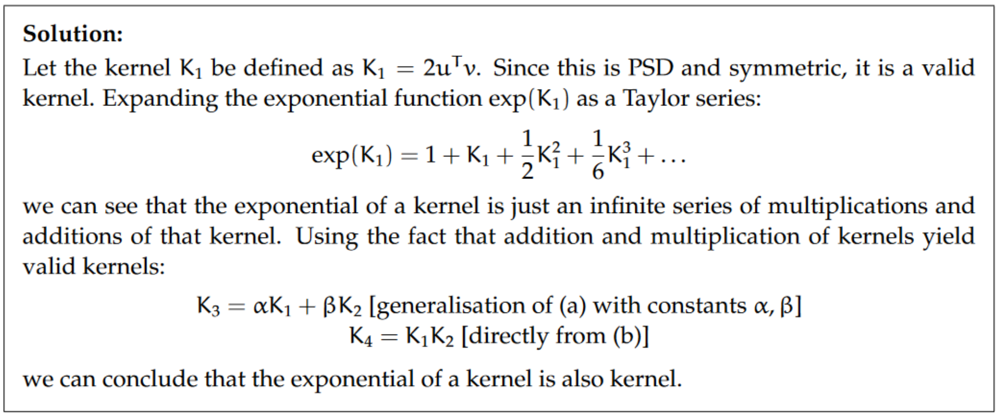
$$\sum_{i=1}^{M}\sum_{j=1}^{M} c_iK_5(x_i,x_j)c_j = \sum_{i=1}^{M}\sum_{j=1}^{M} d_iK_1(x_i,x_j)d_j = D^TK_1D \geqslant 0 \implies C^TK_5C \geqslant 0$$

Hence, $K_5$ is positive definite. Since it is symmetric and PSD, it is a valid kernel.

Feature mapping $\phi_5(x,y) : \phi_1(x,y)f(y)$

(iv) Show that a Kernel given by $K(u,v) = \exp(2u^Tv)$ is a valid kernel. [Hint: Use the results above on a polynomial expansion of $exp(t)$.]

**Solution:**



**Solution:**

Let the kernel $K_1$ be defined as $K_1 = 2u^T v$. Since this is PSD and symmetric, it is a valid kernel. Expanding the exponential function $\exp(K_1)$ as a Taylor series:

$$\exp(K_1) = 1 + K_1 + \frac{1}{2}K_1^2 + \frac{1}{6}K_1^3 + \ldots$$

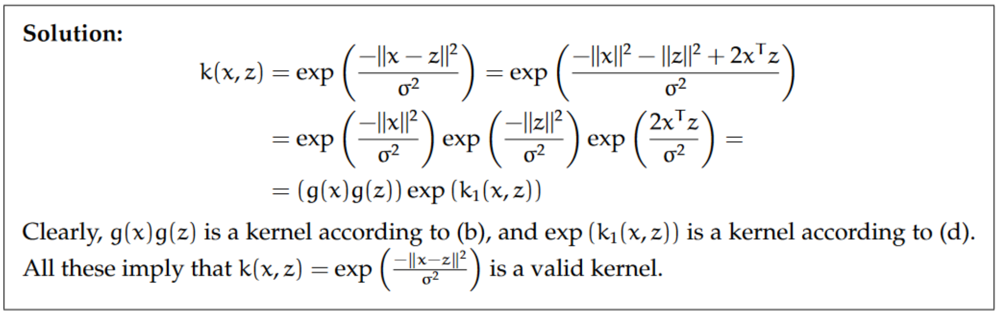
we can see that the exponential of a kernel is just an infinite series of multiplications and additions of that kernel. Using the fact that addition and multiplication of kernels yield valid kernels:

$$K_3 = \alpha K_1 + \beta K_2 \text{ [generalisation of (a) with constants } \alpha, \beta]$$
$$K_4 = K_1 K_2 \text{ [directly from (b)]}$$

we can conclude that the exponential of a kernel is also kernel.

(v) (Optional) Show that a Kernel given by $K(u, v) = \exp(-||u - v||^2)$ is a valid kernel. [Hint: Use the last two parts' results.]

**Solution:**



**Solution:**

$$\begin{aligned} k(x,z) &= \exp\left(\frac{-||x-z||^2}{\sigma^2}\right) = \exp\left(\frac{-||x||^2 - ||z||^2 + 2x^T z}{\sigma^2}\right) \\ &= \exp\left(\frac{-||x||^2}{\sigma^2}\right)\exp\left(\frac{-||z||^2}{\sigma^2}\right)\exp\left(\frac{2x^T z}{\sigma^2}\right) = \\ &= (g(x)g(z))\exp\left(k_1(x,z)\right) \end{aligned}$$

Clearly, $g(x)g(z)$ is a kernel according to (b), and $\exp\left(k_1(x,z)\right)$ is a kernel according to (d). All these imply that $k(x,z) = \exp\left(\frac{-||x-z||^2}{\sigma^2}\right)$ is a valid kernel.