SC4/SM8 Advanced Topics in Statistical Machine Learning
Problem Sheet 2

1. Let k_1 and k_2 be positive definite kernels on $\mathbb{R}^p$. Verify that the following are also valid kernels.

   [*Hint: it suffices to identify the corresponding feature.*]

   (a) $x^\top x'$,

   (b) $ck_1(x, x')$, for $c \geq 0$,

   (c) $f(x)k_1(x, x')f(x')$ for any function $f : \mathbb{R}^p \to \mathbb{R}$,

   (d) $k_1(x, x') + k_2(x, x')$,

   (e) $k_1(x, x')k_2(x, x')$,

   (f) $\exp(k_1(x, x'))$,

   (g) $\exp\left(-\frac{1}{2\gamma^2}\|x - x'\|_2^2\right)$.

2. Assume that kernel k is not strictly positive definite, but that there exist $\{a_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$, such that
$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) = 0.$$
   Show that then
$$f(x) = \sum_{i=1}^n a_i k(x_i, x) = 0 \quad \forall x \in \mathcal{X}.$$
   Hence conclude that the RKHS functions of the form $f(x) = \sum_{i=1}^n a_i k(x_i, x)$ have zero norm if and only if they are identically equal to zero. [*Hint: assume contrary for some $x = x_{n+1}$ and consider $\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} a_i a_j k(x_i, x_j)$*]

3. (**One-Class SVM**) A Gaussian RBF kernel on $\mathcal{X} = \mathbb{R}^p$ is given by
$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right). \tag{1}$$

   (i) What is $k(x, x)$ for this kernel? What can you conclude about the norm of the features $\varphi(x)$ of x? What values can the angles between $\varphi(x)$ and $\varphi(x')$ take? Sketch the set $\{\varphi(x) : x \in \mathcal{X}\}$ as if the features lived in a 2D space.

   (ii) Let $\{x_i\}_{i=1}^n$ be a set of points in $\mathcal{X} = \mathbb{R}^p$ (no labels are given). The one-class Support Vector Machine (SVM) is a method for outlier detection which in its primal form is defined as
$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho, \quad \text{subject to } \langle w, \varphi(x_i)\rangle \geq \rho - \xi_i,\ \xi_i \geq 0,$$
   where ν is a given SVM parameter, features $\varphi(x)$ correspond to the RBF kernel in (1), and ξ_i's are the non-negative slack variables. The fitted hyperplane $\langle w, \varphi(x)\rangle - \rho$ in the feature space separates the majority of points from the origin (while pushing away from the origin as much as possible) and is used to determine "atypical" x-instances.

   Using the 2D intuition from (i), sketch the corresponding hyperplane in the feature space and annotate with ρ, w and a non-zero slack ξ_j for an "outlier" x_j. Would it make sense to use the one-class SVM with a linear kernel?

(iii) Write the dual form of the one-class SVM, using Lagrangian duality.
[*Hint: setting to zero the derivative of the Lagrangian with respect to* w *should give* $w = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$, *where* $\alpha_i \geq 0$ *are the Lagrange multipliers of the constraints* $\langle w, \varphi(x_i) \rangle \geq \rho - \xi_i$]

4. Derive the Gram matrix $\tilde{\mathbf{K}}$ of centred features $\tilde{\varphi}(x_i) = \varphi(x_i) - \frac{1}{n}\sum_{r=1}^{n} \varphi(x_r)$ as a function of kernel values $\mathbf{K}_{i,j} = k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j)$. Show that it takes the form $\mathbf{HKH}$, where $\mathbf{H}$ is a matrix you should specify. Verify that $\mathbf{H}$ is symmetric and idempotent, i.e., $\mathbf{H}^2 = \mathbf{H}$.

5. Show that

$$\mathrm{MMD}_k\,(P, Q) = \sup_{f \in \mathcal{H}_k :\, \|f\|_{\mathcal{H}_k} \leq 1} \left| \mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y) \right|.$$

6. Consider a multilayer perceptron with 1 hidden layer consisting of N hidden units. The MLP is given by the function $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \sum_{i=1}^{N} w_j h(a_j^\top x + b_j)$$

with nonlinearity h and parameters initialised iid as:

$$\begin{aligned} w_j &\sim \mathcal{N}(0, \sigma_w^2/N) \\ a_j &\sim \mathcal{N}(0, \sigma_a^2 I_d) \\ b_j &\sim \mathcal{N}(0, \sigma_b^2) \end{aligned} \tag{2}$$

Assume that the nonlinearity has bounded second moment, $\mathbb{E}[h(a^\top x + b)^2] \leq V < \infty$ for all $x \in \mathbb{R}^d$. We will consider the behaviour of f at initialisation, in case of a very wide MLP, i.e. $N \to \infty$.

(a) Show that for each $x \in \mathbb{R}^d$, $f(x)$ is normally distributed as $N \to \infty$, with zero mean and variance $\sigma_w^2 \mathbb{E}[h(a^\top x + b)^2]$ where $\mathbb{E}$ is expectation with respect to random parameters a and b given by the initialisation. Why is the division by N important in (2)?

(b) Show that for $x, x' \in \mathbb{R}^d$, the pair $f(x), f(x')$ is also jointly normally distributed as $N \to \infty$, with zero mean and variance $\sigma_w^2 \mathbb{E}[h(a^\top x + b) h(a^\top x' + b)]$.

(c) For input-output pair (x, y) and square loss, derive the gradients with respect to w_j, a_j and b_j.

(d) What do you notice about the typical scales of these gradients at the first step of SGD? For a wide MLP with very large N, how would SGD behave at the first iteration? Specifically, would the first layer parameters (a_j, b_j) change much relative to the second layer parameters (w_j)? How about for subsequent iterations?

(e) Would ADAM behave differently?

(f) Suppose we parameterise our MLP slightly differently:

$$f'(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_j' h((a_j')^\top x + b_j')$$

with parameters initialised iid as:

$$
\begin{aligned}
w'_j &\sim \mathcal{N}(0, \sigma_w^2) \\
a'_j &\sim \mathcal{N}(0, \sigma_a^2 I_d) \\
b'_j &\sim \mathcal{N}(0, \sigma_b^2)
\end{aligned}
\tag{3}
$$

Explain why this does not change the MLP model. How does this change the behaviour of SGD in the first and subsequent iterations?