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

   Answer:

   (a) $\varphi(x) = x$.

   (b) $\varphi(x) = \sqrt{c}\varphi_1(x)$, where φ_1 is the feature of k_1.

   (c) $\varphi(x) = f(x)\varphi_1(x)$

   (d) Positive definite as

   $$\sum_{i,j} \alpha_i\alpha_j(k_1(x_i, x_j) + k_2(x_i, x_j)) = \sum_{i,j} \alpha_i\alpha_j k_1(x_i, x_j) + \sum_{i,j} \alpha_i\alpha_j k_2(x_i, x_j) \geq 0.$$

   The feature is obtained by "stacking" vectors φ_1 and φ_2 together.

   (e) By writing φ_1, φ_2 for the features of k_1 and k_2, we have

   $$\begin{aligned} k_1(x, x')k_2(x, x') = \varphi_1(x)^\top\varphi_1(x')\varphi_2(x')^\top\varphi_2(x) &= \operatorname{Tr}\left(\varphi_1(x')\varphi_2(x')^\top\varphi_2(x)\varphi_1(x)^\top\right) \\ &= \operatorname{Tr}\left(\Phi(x')\Phi(x)^\top\right) \\ &= \langle\Phi(x'), \Phi(x)\rangle, \end{aligned}$$

   where the feature is the outer product matrix $\Phi(x) = \varphi_1(x)\varphi_2(x)^\top$.

   (f) From (b), (d) and (e), since addition and multiplication preserves positive definiteness and since all the coefficients in the Taylor series expansion of the exponential function are non-negative, $\kappa_m(x, x') = \sum_{r=1}^{m} \frac{k_1^r(x,x')}{r!}$ is a valid kernel $\forall m \in \mathbb{N}$. Fix α and $\{x_i\}$. Then $a_m = \sum_{i,j} \alpha_i\alpha_j\kappa_m(x_i, x_j) \geq 0\ \forall m$. But $a_m \to \sum_{i,j} \alpha_i\alpha_j \exp(k_1(x_i, x_j))$ as $m \to \infty$, so $\sum_{i,j} \alpha_i\alpha_j \exp(k_1(x_i, x_j)) \geq 0$ as well.

   (g) By (a), (b), (f), $\exp\left(\frac{1}{\gamma^2}x^\top x'\right)$ is a valid kernel, but then by (c) so is $\exp\left(-\frac{1}{2\gamma^2}\|x - x'\|_2^2\right) = \exp\left(-\frac{1}{2\gamma^2}\|x\|_2^2\right)\exp\left(\frac{1}{\gamma^2}x^\top x'\right)\exp\left(-\frac{1}{2\gamma^2}\|x'\|_2^2\right)$

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* $\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} a_i a_j k(x_i, x_j)$]

Answer: Assume $f(x_{n+1}) \neq 0$. Then $\forall a_{n+1}$

$$\begin{aligned} 0 \leq \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \\ &\quad + 2a_{n+1} \sum_{i=1}^{n} a_i k(x_i, x_{n+1}) + a_{n+1}^2 k(x_{n+1}, x_{n+1}) \\ &= 2a_{n+1} f(x_{n+1}) + a_{n+1}^2 k(x_{n+1}, x_{n+1}). \end{aligned}$$

To minimise the expression in the last line, take $a_{n+1} = -f(x_{n+1})/k(x_{n+1}, x_{n+1})$. But this gives

$$0 \leq -f^2(x_{n+1})/k(x_{n+1}, x_{n+1}).$$

Since $k(x_{n+1}, x_{n+1}) > 0$, it must be that $f(x_{n+1}) = 0$. The conclusion about functions of the form $f(x) = \sum_{i=1}^n a_i k(x_i, x)$ is immediate since $\|f\|_{\mathcal{H}_k}^2 = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j)$.

Another way to show this is by simply applying the Cauchy-Schwarz inequality in $\mathcal{H}_k$:

$$|f(x)| = |\langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}| \leq \|f\|_{\mathcal{H}_k} \sqrt{k(x, x)}.$$

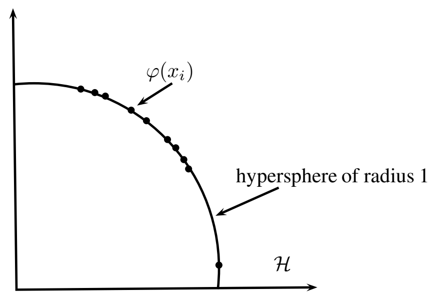
Thus $\|f\|_{\mathcal{H}_k} = 0$ implies $f(x) = 0, \ \forall x$.

3. (**One-Class SVM**) A Gaussian RBF kernel on $\mathcal{X} = \mathbb{R}^p$ is given by

$$k(x, x') = \exp\left(-\frac{1}{2\sigma^2} \left\|x - x'\right\|^2\right). \tag{1}$$

   (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.

   Answer: $k(x, x) = \|\varphi(x)\|_2^2 = 1$, and $k(x, x') = \langle \varphi(x), \varphi(x')\rangle > 0$, so the angle between any two feature vectors is not larger than $\pi/2$.



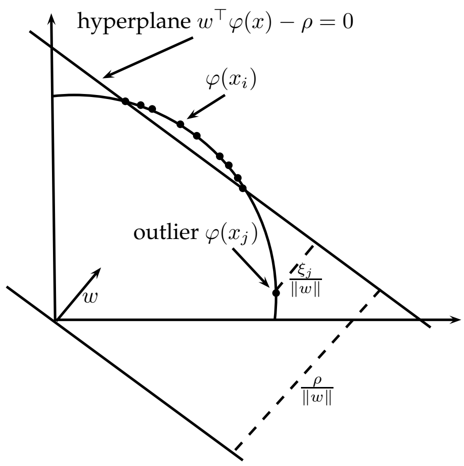
(ii) Let $\{x_i\}_{i=1}^n$ be a set of points in $\mathcal{X} = \mathbb{R}^p$ (no labels are given). The one-class Support Vector Machine (SVM) is a method for outlier detection which in its primal form is defined as

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho, \quad \text{subject to } \langle w, \varphi(x_i)\rangle \geq \rho - \xi_i,\ \xi_i \geq 0,$$

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?

Answer: The hyperplane that separates majority of points from the origin is useful for outlier detection precisely because all feature vectors lie on the unit hypersphere. One-class SVM therefore relies on the properties of the RBF kernel and would not make sense with a linear kernel. With linear kernel, gross outliers in the same half-space as majority of data would still be allowed.



(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 $\langle w, \varphi(x_i)\rangle \geq \rho - \xi_i$*]

Answer: Lagrangian is given by

$$\begin{aligned} L(w,\xi,\rho,\alpha,\beta) &= \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho \\ &\qquad - \sum_{i=1}^n \alpha_i\left(\langle w, \varphi(x_i)\rangle - \rho + \xi_i\right) - \sum_{i=1}^n \beta_i\xi_i, \end{aligned}$$

for Lagrange multipliers $\alpha_i \geq 0$, $\beta_i \geq 0$. Differentiating w.r.t. w, ξ, ρ and setting to zero gives

$$w = \sum_{i=1}^n \alpha_i\varphi(x_i), \qquad \alpha_i + \beta_i = \frac{1}{\nu n}, \qquad \sum_{i=1}^n \alpha_i = 1.$$

Substituting back into Lagrangian gives the dual:

$$\max_{\alpha} -\frac{1}{2}\alpha^\top K\alpha, \quad \textit{subject to} \quad \sum_{i=1}^{n} \alpha_i = 1, \; \alpha_i \le \frac{1}{\nu n}.$$

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., $\mathbf{H}^2 = \mathbf{H}$.

   Answer: To get the centred features we need

$$\begin{aligned}
\tilde{\mathbf{K}}_{i,j} &= \left\langle \varphi(x_i) - \frac{1}{n}\sum_{r=1}^{n}\varphi(x_r),\, \varphi(x_j) - \frac{1}{n}\sum_{r=1}^{n}\varphi(x_r) \right\rangle \\
&= \langle\varphi(x_i), \varphi(x_j)\rangle + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\langle\varphi(x_r), \varphi(x_s)\rangle \\
&\quad -\frac{1}{n}\sum_{r=1}^{n}\langle\varphi(x_i), \varphi(x_r)\rangle - \frac{1}{n}\sum_{r=1}^{n}\langle\varphi(x_r), \varphi(x_j)\rangle \\
&= \mathbf{K}_{i,j} + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\mathbf{K}_{r,s} - \frac{1}{n}\sum_{r=1}^{n}\mathbf{K}_{i,r} - \frac{1}{n}\sum_{r=1}^{n}\mathbf{K}_{r,j},
\end{aligned}$$

   which depends only on $\mathbf{K}$. In matrix form, $\tilde{\mathbf{K}} = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{K}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$, where the centering matrix $\mathbf{H} = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is clearly symmetric. To check idempotence:

$$\begin{aligned}
(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top) &= (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top - \cancel{\frac{1}{n}\mathbf{1}\mathbf{1}^\top} + \cancel{\frac{1}{n^2}\mathbf{1}\underbrace{\mathbf{1}^\top\mathbf{1}}_{n}\mathbf{1}^\top}) \\
&= I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top.
\end{aligned}$$

5. Show that

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

   Answer:

$$\begin{aligned}
\sup_{f\in\mathcal{H}_k:\, \|f\|_{\mathcal{H}_k}\le 1} \left|\mathbb{E}_{X\sim P}f(X) - \mathbb{E}_{Y\sim Q}f(Y)\right| &= \sup_{f\in\mathcal{H}_k:\, \|f\|_{\mathcal{H}_k}\le 1} \left|\langle f, \mu_k(P) - \mu_k(Q)\rangle_{\mathcal{H}_k}\right| \\
&\le 1 \cdot \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}.
\end{aligned}$$

   by Cauchy-Schwarz. Moreover, the equality holds if f is colinear with $\mu_k(P) - \mu_k(Q)$, i.e. the supremum is attained at

$$f = \frac{\mu_k(P) - \mu_k(Q)}{\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}}$$

   which is called the witness function.

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)?

Answer: Each of the N terms in $f(x)$ are iid, with bounded variance, so we can apply the Central Limit Theorem. It is easy to check that the mean and variance are as given.

Division by N is important as otherwise the scale of $f(x)$ will grow as $N^{1/2}$ which can get too large if N is large.

This initialisation is known as Xavier initialisation.

(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 $\sigma_w^2 \mathbb{E}[h(a^\top x + b)h(a^\top x' + b)]$.

Answer: We can express the vector $[f(x), f(x')]^\top$ as a sum of N iid terms as well, then apply CLT. We need to check that $[h(a^\top x + b), h(a^\top x' + b)]^\top$ has bounded second moment:

$$\mathbb{E}[\|[h(a^\top x + b), h(a^\top x' + b)]^\top\|^2] = \mathbb{E}[|h(a^\top x + b)|^2 + |h(a^\top x' + b)|^2] \leq 2V$$

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

Answer: The gradients are:

$$\begin{aligned} \frac{\partial L}{\partial f(x)} &= (f(x) - y) \\ \frac{\partial L}{\partial w_j} &= (f(x) - y) h(a_j^\top x + b_j) \\ \frac{\partial L}{\partial h(a_j^\top x + b_j)} &= (f(x) - y) w_j \\ \frac{\partial L}{\partial a_j} &= (f(x) - y) w_j h'(a_j^\top x + b_j) x \\ \frac{\partial L}{\partial b_j} &= (f(x) - y) w_j h'(a_j^\top x + b_j) \end{aligned}$$

where $h'(z)$ is the derivative of h at z.

(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?

Answer: At initialisation the typical scale of w_j is $N^{-1/2}$. So the gradients for a_j, b_j are $N^{1/2}$ times smaller than for w_j. At the first iteration, a_j, b_j will change much less than for w_j.

In subsequent iterations, some w_j's might become much larger than $N^{-1/2}$, and the resulting gradients for a_j, b_j will be larger as well. This can lead to "sparsity" where some parameters are much larger than others.

(e) Would ADAM behave differently?

Answer: ADAM rescales the gradient of each parameter by $1/\sqrt{v}$ where v is the average of squared gradients over past iterations. This ensures that the $N^{-1/2}$ factor is removed, so updates to a_j, b_j will not be much smaller than those for w_j.

This is an example where the choice optimisation algorithm can significantly impact the learning.

(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?

Answer: This does not change the MLP model since the division by $\sqrt{N}$ just appears elsewhere.

The gradients are changed as follows:

$$\begin{aligned} \frac{\partial L}{\partial f'(x)} &= (f'(x) - y) \\ \frac{\partial L}{\partial w'_j} &= N^{-1/2}(f'(x) - y)h((a'_j)^\top x + b'_j) \\ \frac{\partial L}{\partial h((a'_j)^\top x + b'_j)} &= N^{-1/2}(f'(x) - y)w'_j \\ \frac{\partial L}{\partial a'_j} &= N^{-1/2}(f'(x) - y)w'_j h'((a'_j)^\top x + b'_j)x \\ \frac{\partial L}{\partial b'_j} &= N^{-1/2}(f'(x) - y)w'_j h'((a'_j)^\top x + b'_j) \end{aligned}$$

Now all gradients are very small and scale similarly as $N^{-1/2}$.

It is possible to make the learning rate larger to offset the small gradients so that updates are substantial. However for large N the number of parameters is very large and this might lead to a very large change to $f'(x)$ itself and causing unstable learning. It turns out that in practice even if all parameters are updated very little, the resulting total update to $f'(x)$ can be significant.