CPSC 532D: Assignment 4 – due Saturday, 10 Dec 2022, **11:59pm**

As before: use LaTeX, either with the template I give or your own document if you prefer.

You can do this with a partner if you'd like (there's a "find a group" post on Piazza), but please **make sure you understand everything you're submitting** – don't just split an assignment in half. If you do parts of the assignment with a partner and parts separately, submit separate solutions, and say in each part you did together who you did it with.

If you look stuff up anywhere other than in **SSBD, MRT, Telgarsky, or Wainwright**, **cite your sources**: just say in the answer to that question where you looked. If you ask anyone else for help, **cite that too**. Please do not look at solution manuals / search for people proving the things we're trying to prove / etc. If you accidentally come across a solution while looking for something related, still write the argument up in your own words, link to wherever you found it, and be clear about what happened.

1 Proving kerneldom [25 + 5 challenge + 2 bonus points]

Prove that the following functions are kernels, i.e. that they are positive definite functions.

Hint: Recall that we proved you can do so by directly proving all kernel matrices are psd, by writing an explicit feature mapping $k(x, x') = \langle \phi(x), \phi(x') \rangle$ where ϕ maps into any Hilbert space (including $\mathbb{R}^d$), or by using steps known to produce new kernels out of old ones.

You could also use Bochner's theorem, which we did not cover in class, if you're a Fourier buff: a kernel $k(x, y) = \psi(x - y)$ with $\psi(0) = 1$ is psd iff it is the Fourier transform of a probability measure.

Hint: Here are two Hilbert spaces that might be useful to you. First, the space ℓ^2 of square-summable sequences $(a_k)_{k=1}^\infty$ with inner product $\langle (a_k), (b_k) \rangle_{\ell^2} = \sum_k a_k b_k$. Second, the space $L^2(\mathcal{X})$ of square-integrable functions[1] on $\mathcal{X}$, with inner product $\langle f, g \rangle_{L^2} = \int_{\mathcal{X}} f(x) g(x) \mathrm{d}x$.

(a) [5 points] $k(x, y) = \cos(x - y)$ on $\mathbb{R}$.

Hint: The list of trigonometric identities makes for good bedtime reading.

Answer: TODO

(b) [5 points] $k_n(x, y) = \frac{1}{2\pi} \left[1 + 2 \sum_{k=1}^n \cos(k(x - y))\right] = \frac{\sin\left(\left(n + \frac{1}{2}\right)(x - y)\right)}{2\pi \sin\left(\frac{1}{2}(x - y)\right)}$ on $\mathbb{R}$ for any $n \geq 0$.

(This is called the Dirichlet kernel*; it's a continuous kernel which converges to the Dirac delta function $\delta(x - y)$ as $n \to \infty$.)*

Answer: TODO

(c) [5 points] $k(x, y) = \min(x, y)$ on $[0, 1]$.

Hint: You could consider the integral $\int_{\mathbb{R}} \mathbb{1}(t \in [0, x]) \mathbb{1}(t \in [0, y]) \mathrm{d}t$.

Answer: TODO

(d) [5 points] $k(X, Y) = \sum_{x \in X} \sum_{y \in Y} k_0(x, y)$ on finite sets with elements in $\mathcal{X}$, where k_0 is a kernel on $\mathcal{X}$.

Answer: TODO

(e) [5 points] $k(x, y) = 1/\sqrt{1 - xy}$ on $(-1, 1)$.

For [2 bonus points], you can instead show $1/\sqrt{1 - x^\mathsf{T} y}$ on $\{x \in \mathbb{R}^d : \|x\| < 1\}$.

Hint: It might help to use the following expansion (see e.g. here), which converges for $|z| < 1$:

$$\frac{1}{\sqrt{1 - z}} = \sum_{k=0}^\infty c_k z^k \quad \textit{for} \quad c_k := \frac{1}{2^{2k}} \binom{2k}{k}.$$

Answer: TODO

(f) [5 challenge points] The distance kernel $k(x, y) = \|x\| + \|y\| - \|x - y\|$, where $\|\cdot\|$ is the norm of any Hilbert space.

Hint: For all $n \geq 1$, for all $x_1, \ldots, x_n$ and $c_1, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$, it holds that

$$\sum_{i=1}^n \sum_{j=1}^n c_i \|x_i - x_j\| c_j \leq 0.$$

Answer: TODO

[1] Really, this should be a space of equivalence classes of functions, since a function that's zero only almost everywhere will have norm zero. That won't matter for this question.

2 Maximizing differences [25 points]

Let's consider learning a kernel classifier with the somewhat unusual *linear loss*, $\ell(h, (x, y)) = -yh(x)$, where $y \in \{-1, 1\}$. Take the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated RKHS $\mathcal{F}$.

(a) [10 points] Find the regularized loss minimizer

$$\hat{h}_\lambda = \operatorname*{arg\,min}_{h \in \mathcal{F}} L_S(h) + \tfrac{1}{2}\lambda \|h\|_{\mathcal{F}}^2, \tag{RLM}$$

for a training sample $S = ((x_1, y_1), \ldots, (x_n, y_n))$.

Answer: TODO

(b) [5 points] Show that $L_S(\hat{h}_\lambda) = -\frac{1}{\lambda} \left\| \frac{1}{n} \sum_{i:y_i=1} k(x_i, \cdot) - \frac{1}{n} \sum_{i:y_i=-1} k(x_i, \cdot) \right\|_{\mathcal{F}}^2$.

Answer: TODO

(c) [5 points] Find a (data-dependent) value of λ, call it $\hat{\lambda}$, such that $\|\hat{h}_{\hat{\lambda}}\|_{\mathcal{F}} = 1$, and simplify the expression for $L_S(\hat{h}_{\hat{\lambda}})$.

Answer: TODO

(d) [5 points] Argue that $\hat{h}_{\hat{\lambda}}$ is a solution to

$$\min_{h \in \mathcal{F} : \|h\| \le 1} L_S(h). \tag{ERM}$$

Further argue that solving (ERM) is equivalent to solving

$$\max_{h \in \mathcal{F} : \|h\| \le 1} \sum_{i:y_i=1} h(x_i) - \sum_{i:y_i=-1} h(x_i), \tag{MAX}$$

i.e. finding a function high on the positively-labeled points and low on the negatively-labeled ones.

Answer: TODO

3 Kernel metrics on distributions [15 + 5 challenge points]

Here's a slight tweak to the object we developed in Question 2, called the *maximum mean discrepancy*:

$$\operatorname{MMD}(\mathcal{P}, \mathcal{Q}) = \sup_{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \le 1} \mathop{\mathbb{E}}_{X \sim \mathcal{P}} f(X) - \mathop{\mathbb{E}}_{Y \sim \mathcal{Q}} f(Y).$$

Assume for this question that the kernel k of $\mathcal{F}$ is continuous and satisfies $\sup_{x \in \mathcal{X}} k(x, x) \le \kappa^2$ for some $\kappa < \infty$. Define the *kernel mean embedding* of a (Borel) distribution $\mathcal{P}$ as $\mu_{\mathcal{P}} = \mathbb{E}_{X \sim \mathcal{P}} k(X, \cdot)$; for bounded kernels, this is guaranteed to exist, and you can move the expectation inside or outside of inner products: for any $f \in \mathcal{F}$,

$$\langle \mu_{\mathcal{P}}, f \rangle_{\mathcal{F}} = \langle \mathop{\mathbb{E}}_{X \sim \mathcal{P}} k(X, \cdot), f \rangle_{\mathcal{F}} = \mathop{\mathbb{E}}_{X \sim \mathcal{P}} \langle k(X, \cdot), f \rangle_{\mathcal{F}} = \mathop{\mathbb{E}}_{X \sim \mathcal{P}} f(X).$$

(a) [10 points] Prove that

$$\operatorname{MMD}(\mathcal{P}, \mathcal{Q}) = \|\mu_{\mathcal{P}} - \mu_{\mathcal{Q}}\|_{\mathcal{F}}$$

and

$$\operatorname{MMD}^2(\mathcal{P}, \mathcal{Q}) = \mathop{\mathbb{E}}_{\substack{X, X' \sim \mathcal{P} \\ Y, Y' \sim \mathcal{Q}}} \Big[k(X, X') - 2k(X, Y) + k(Y, Y') \Big].$$

Answer: TODO

From the $\|\mu_{\mathcal{P}} - \mu_{\mathcal{Q}}\|_{\mathcal{F}}$ form, we can see that MMD satisfies all the conditions of a metric on probability distributions except that we might have $\operatorname{MMD}(\mathcal{P}, \mathcal{Q}) = 0$ for $\mathcal{P} \ne \mathcal{Q}$. *The* energy distance *is equivalent to the MMD with the distance kernel from Question 1 part (f).*

(b) [5 challenge points] Let $\mathcal{X}$ be a compact metric space, and $C(\mathcal{X})$ denote the set of continuous bounded functions on $\mathcal{X}$. A *universal* kernel k has an RKHS $\mathcal{F}$ such that for any $g \in C(\mathcal{X})$, for all $\varepsilon > 0$ there exists an $f \in \mathcal{F}$ with $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)| \le \varepsilon$. Prove that if k is universal, then $\operatorname{MMD}(\mathcal{P}, \mathcal{Q}) = 0$ implies $\mathcal{P} = \mathcal{Q}$.

Hint: The following result will probably be helpful:

Lemma 3.1. *Two Borel probability measures $\mathcal{P}$ and $\mathcal{Q}$ on a metric space $\mathcal{X}$ are equal if and only if for all $f \in C(\mathcal{X})$, $\mathbb{E}_{X \sim \mathcal{P}} f(X) = \mathbb{E}_{Y \sim \mathcal{Q}} f(Y)$.*

Answer: TODO

Let U and V be (potentially dependent) random variables with values in $\mathcal{U}$ and $\mathcal{V}$ respectively. Let k_U be a kernel over $\mathcal{U}$, k_V a kernel over $\mathcal{V}$, and $k_{UV}((u, v), (u', v')) = k_U(u, u') k_V(v, v')$ a kernel over $\mathcal{U} \times \mathcal{V}$.

Define $C_{UV} = \mathbb{E}[k_U(U, \cdot) \otimes k_V(V, \cdot)] - \mathbb{E}[k_U(U, \cdot)] \otimes \mathbb{E}[k_V(V, \cdot)]$; recalling that the outer product of $a \in \mathcal{A}$ and $b \in \mathcal{B}$ is a linear operator from $\mathcal{B}$ to $\mathcal{A}$ defined by $[a \otimes b]x = a \langle b, x \rangle_{\mathcal{B}}$ for all $x \in \mathcal{B}$.

(c) [5 points] Show that $\langle f, C_{UV} g \rangle_{\mathcal{F}_U} = \operatorname{Cov}(f(U), g(V))$ for all $f \in \mathcal{F}_U$, $g \in \mathcal{F}_V$.

Answer: TODO

Using a similar argument to part (b), one can show that if k_{UV} is universal (or various somewhat weaker conditions), then U and V are independent if and only if $C_{UV} = 0$. This can be checked using something called the Hilbert-Schmidt independence criterion *(HSIC), which estimates the squared Hilbert-Schmidt norm of C_{UV}. With the distance kernel of Question 1 part (f), this becomes proportional to the* distance covariance.

4 One way to do semi-supervised learning [25 points]

Semi-supervised learning is when you're given not only a training set of (x, y) pairs, but also a set of unlabeled x samples from the marginal distribution of x. (For instance, maybe you have a really big dataset scraped from the web, and have only paid for human annotation of a small, random selection from it.) Even though there aren't any labels, this can be useful for determining the optimal decision function under some reasonable assumptions: for instance, if you have a clear cluster structure, it's perhaps more likely that the labeling function is constant on that cluster.

One way to try to implement this is to penalize the *gradient norm* of the decision function, evaluated at the data points – the decision function should be smooth where there's data. (You might be familiar with this type of gradient penalty from GANs.) It turns out that the special structure of RKHSes will allow for this. Specifically, let's let $\mathcal{F}$ be an RKHS for some continuously twice-differentiable kernel k on $\mathbb{R}$, i.e. $f \in \mathcal{F}$ maps $\mathbb{R}$ to $\mathbb{R}$. (Everything here will work for $\mathbb{R}^d$, the notation just gets a little messier.)

Our goal will be to minimize the following regularized loss over all of $\mathcal{F}$, where both ν and λ are positive scalars:

$$J(h) = \frac{1}{n}\sum_{i=1}^{n}\big(h(x_i) - y_i\big)^2 + \frac{\nu}{m}\sum_{i=1}^{m}\big(h'(x_i)\big)^2 + \lambda\|h\|_{\mathcal{F}}^2. \tag{J}$$

Here we assume that we have our usual sample set $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$, but we *also* have an unlabeled sequence $\big(x_{n+1}, \ldots, x_m\big)$, so that we have m total samples for x (of which the first n are labeled).

Kernels are two-argument functions, so differentiation notation can be slightly awkward. For brevity, we will use ∂_1 to refer to differentiating with respect to the first argument and ∂_2 the second, so that $\partial_1 k(x, y)$ means $\frac{\partial}{\partial x}k(x, y)$, $\partial_2^2 k(x, y)$ means $\frac{\partial^2}{\partial y^2}k(x, y)$, and $\partial_1\partial_2 k(x, x)$ means $\frac{\partial^2}{\partial z_1 \partial z_2}k(z_1, z_2)|_{\substack{z_1=x\\ z_2=x}}$ – note that differentiation happens "before" passing the arguments in (this is *not* $\frac{\partial^2}{\partial x^2}k(x, x)$).

The following result will be useful for us:

Lemma 4.1 (Special case of Steinwart/Christmann Lemma 4.34)**.** *Let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a kernel such that both $\partial_1\partial_2 k(x, y)$ and $\partial_1^2\partial_2^2 k(x, y)$ exist and are continuous. Then, for all $x \in \mathbb{R}$, $\partial_1 k(x, \cdot)$ and $\partial_1^2 k(x, \cdot)$ are functions in $\mathcal{F}$ such that for all $f \in \mathcal{F}$ we have*

$$\langle \partial_1 k(x, \cdot), f\rangle_{\mathcal{F}} = f'(x) \quad \text{and} \quad \langle \partial_1^2 k(x, \cdot), f\rangle_{\mathcal{F}} = f''(x).$$

For example, this also means that

$$\langle \partial_1 k(x, \cdot), \partial_1 k(x', \cdot)\rangle_{\mathcal{F}} = \partial_1\partial_2 k(x, x') \quad \text{and} \quad \langle \partial_1^2 k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{F}} = \partial_1^2 k(x, x').$$

(a) [13 points] Show a representer theorem for $\arg\min_{h\in\mathcal{H}} J(h)$, i.e. that you can write the optimal h as a linear combination of some set of vectors in $\mathcal{H}$.

Hint: The representer theorem we showed in class won't *directly apply, because J depends on the derivatives of h. You'll need to make an analogous argument, taking advantage of the lemma above.*

Answer: TODO

Define the following matrices:

$$K = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$G = \begin{bmatrix} \partial_1 k(x_1, x_1) & \dots & \partial_1 k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \partial_1 k(x_m, x_1) & \dots & \partial_1 k(x_m, x_n) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$H = \begin{bmatrix} \partial_1 \partial_2 k(x_1, x_1) & \dots & \partial_1 \partial_2 k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ \partial_1 \partial_2 k(x_m, x_1) & \dots & \partial_1 \partial_2 k(x_m, x_m) \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

(b) [12 points] Write an explicit form for $J(h)$ in terms of usual matrix and vector operations on the K, G, and H matrices and the vector $y \in \mathbb{R}^n$ of labels, as well as the parameters of your linear combination (and λ, ν, n, and m).

Hint: It will probably help to start by writing out $h(x_i)$, $h'(x_i)$, and $\|h\|^2_{\mathcal{F}}$, then plugging those together. It might be helpful in intermediate steps to use the standard basis vectors e_i, which have a one in the ith entry and zero in all others. Be careful about shapes matching.

Answer: TODO

If you did it right, the final form for J should be a quadratic form of your coefficients in terms of the matrices K, G, and H. Thus, setting the gradient to zero will give an analytical solution written as the solution to a certain linear system. (No need for you to write out that solution, since it's a little messy.)