MATH 547, STATISTICAL LEARNING THEORY SELECTED HOMEWORK SOLUTIONS

STEVEN HEILMAN

Contents

1. Homework 1

Exercise 1.8. Let $a>0$. Let $X^{(1)},\ldots,X^{(k)}\in\mathbb{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(a,0,\ldots,0)$ and identity covariance matrix). Let $X^{(k+1)},X^{(k+2)},\ldots,X^{(2k)}\in\mathbb{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(-a,0,\ldots,0)$, where $a>0$ is known. As in our analysis of the perceptron algorithm, define

$$\mathcal{B}:=\max_{i=1,\ldots,2k}\left\|X^{(i)}\right\|$$

$$\Theta:=\min\Big\{\,\|w\|:\forall\,1\leq i\leq 2k\;\; y_i\langle w,X^{(i)}\rangle\geq 1\Big\}.$$

(If the minimum w does not exist, instead define $\Theta:=\infty$.)

Define $y_1=\cdots=y_k:=1$, and $y_{k+1}=\cdots=y_{2k}:=-1$.

Give some reasonable estimates for $\mathbf{E}\mathcal{B}$ and $\mathbf{E}(1/\Theta)$ as a function of a.

Solution. Let $t>a$. Then from the union bound

$$\mathbf{P}(\mathcal{B}>t)\leq\sum_{i=1}^{2k}\mathbf{P}(\left|X^{(i)}\right|>t)=2k\mathbf{P}(\left|X^{(1)}\right|>t)\leq 2ke^{-(t-a)^2/2}.$$

Therefore, $\mathbf{E}\mathcal{B}\leq 100ka$.

Using independence,

$$\mathbf{P}(\forall\,1\leq i\leq 2k\;\; y_i\langle w,X^{(i)}\rangle\geq 1)=[\mathbf{P}(y_1\langle w,X^{(1)}\rangle\geq 1)]^{2k}=[\mathbf{P}(\langle w,X^{(1)}\rangle\geq 1)]^{2k}$$
$$=[\mathbf{P}(\langle(1,0,\ldots,0),X^{(1)}\rangle\geq 1/\|w\|)]^{2k}=[\int_{1/\|w\|}^{\infty}e^{-(t-a)^2/2}dt/\sqrt{2\pi}]^{2k}.$$

In particular, if $1/\|w\|\leq a$ (i.e. $\|w\|\geq 1/a$), then

$$\mathbf{P}(\forall\,1\leq i\leq 2k\;\; y_i\langle w,X^{(i)}\rangle\geq 1)\geq 2^{-2k}.$$

So,

$$\mathbf{P}\Big(\min\Big\{\,\|w\|:\forall\,1\leq i\leq 2k\;\; y_i\langle w,X^{(i)}\rangle\geq 1\Big\}\leq 1/a\Big)\geq 2^{-2k}.$$

Date: October 30, 2021

And

$$\mathbf{E}(1/\Theta) \geq a2^{-2k}.$$

$\square$

2. Homework 2

Exercise 2.2. Let μ be a Borel measure on $\mathbb{R}^n$ such that the measure of any open set in $\mathbb{R}^n$ is positive. Let $m\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be continuous with $\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |m(x,y)|^2\, d\mu(x)d\mu(y) < \infty$. Show that the following two positive semidefinite conditions on m are equivalent:

- $\forall\ p \geq 1$, for all $z^{(1)}, \ldots, z^{(p)} \in \mathbb{R}^n$, for all $\beta_1, \ldots, \beta_p \in \mathbb{R}$ we have

$$\sum_{i,j=1}^{p} \beta_i \beta_j m(z^{(i)}, z^{(j)}) \geq 0.$$

- $\forall\ f \in L_2(\mu)$, we have

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)d\mu(x)d\mu(y) \geq 0.$$

From either condition, we should see that the converse of Mercer's Theorem holds. We should also be able to deduce various properties of positive semidefinite (PSD) kernels. For example, a nonnegative linear combination of PSD kernels is PSD.

Solution. We denote $\|f\|_2 := (\int_{\mathbb{R}^n} |f(x)|^2\, d\mu(x))^{1/2}$. Let $f, g \in L_2(\mu)$. From the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)d\mu(x)d\mu(y) - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)m(x,y)d\mu(x)d\mu(y)\right| \\
&\qquad = \Big|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)d\mu(x)d\mu(y) - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)g(y)m(x,y)d\mu(x)d\mu(y) \\
&\qquad + \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)g(y)m(x,y)d\mu(x)d\mu(y) - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)m(x,y)d\mu(x)d\mu(y)\Big| \\
&\leq \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(x)|\,|f(y)-g(y)|\,|m(x,y)|\, d\mu(x)d\mu(y) \\
&\qquad\qquad + \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(x)-g(x)|\,|g(y)|\,|m(x,y)|\, d\mu(x)d\mu(y) \\
&\leq \int_{\mathbb{R}^n} (|f(x)|^2\, |m(x,y)|^2\, d\mu(x))^{1/2}\, |f(y)-g(y)|\, d\mu(y) \\
&\qquad\qquad + \int_{\mathbb{R}^n} (|f(y)|^2\, |m(x,y)|^2\, d\mu(y))^{1/2}\, |f(x)-g(x)|\, d\mu(x) \\
&\leq 2\,\|f\|_2\,\|f-g\|_2\,(\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |m(x,y)|^2\, d\mu(x)d\mu(y))^{1/2}. \qquad (*)
\end{aligned}$$

Similarly, from the Cauchy-Schwarz inequality, if $\overline{m}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)m(x,y)d\mu(x)d\mu(y) - \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)\overline{m}(x,y)d\mu(x)d\mu(y)\right| \\
&\qquad = \left|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)[m(x,y)-\overline{m}(x,y)]d\mu(x)d\mu(y)\right| \\
&\qquad \leq \|g\|_2^2 \,(\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |m(x,y)-\overline{m}(x,y)|^2 \, d\mu(x)d\mu(y))^{1/2}. \qquad (**)
\end{aligned}$$

Assume the first condition holds. Let $f \in L_2(\mu)$. Let $\varepsilon > 0$. Let g be a simple function of the form $g = \sum_{i=1}^k \alpha_i 1_{A_i}$ such that $\|f-g\|_2 < \varepsilon$, where $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ and $A_1, \ldots, A_k \subseteq \mathbb{R}^n$ are disjoint (measurable) sets with compact closure.

Now, our aim is to show that second property holds for g. Since the support $C := \cup_{i=1}^k A_i$ of g has compact closure, and since m is continuous, m is uniformly continuous on $C \times C$. So, for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \in \mathbb{R}^n \times \mathbb{R}^n$, if $\left\|(x^{(1)}, y^{(1)}) - (x^{(2)}, y^{(2)})\right\| < \delta$, then $\left\|m(x^{(1)}, y^{(1)}) - m(x^{(2)}, y^{(2)})\right\| < \varepsilon$.

For any subset $A \subseteq \mathbb{R}^n$, define $\mathrm{diam}(A) := \sup_{x,y\in A}\|x-y\|$. Since $A_1, \ldots, A_n$ have compact closure, we can rewrite g in the form

$$g = \sum_{i=1}^{\ell} \gamma_i 1_{B_i},$$

where $\gamma_1, \ldots, \gamma_\ell \in \mathbb{R}$ and $\mathrm{diam}(B_i) < \delta/2$. For every $1 \leq i, j \leq \ell$, let $(x^{(i)}, y^{(j)})$ be any point in $B_i \times B_j$. By choice of ε, δ, we have

$$\left|m(x,y) - m(x^{(i)}, y^{(j)})\right| < \varepsilon, \qquad \forall\,(x,y) \in B_i \times B_j.$$

Define $\overline{m}(x,y) := \sum_{i,j=1}^{\ell} m(x^{(i)}, y^{(j)}) 1_{B_i}(x) 1_{B_j}(y)$. Then

$$\begin{aligned}
\int_C\int_C |m(x,y)-\overline{m}(x,y)|^2 \, d\mu(x)d\mu(y) &= \sum_{i,j=1}^{\ell} \int_{B_i}\int_{B_j} |m(x,y)-\overline{m}(x,y)|^2 \, d\mu(x)d\mu(y) \\
&\leq \varepsilon^2 \sum_{i,j=1}^{\ell} \mu(B_i)\mu(B_j).
\end{aligned}$$

The combination of $(*)$ and $(**)$ implies (i.e. first choosing g so that $\|f-g\|_2 < \varepsilon$, and then choosing $B_1, \ldots, B_\ell$ such that $\int_C\int_C |m(x,y)-\overline{m}(x,y)|^2 \, d\mu(x)d\mu(y) < \varepsilon$) that

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)d\mu(x)d\mu(y) > 0 \qquad \text{if } \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)\overline{m}(x,y)d\mu(x)d\mu(y) > 0.$$

By assumption,

$$\begin{aligned}
\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} g(x)g(y)\overline{m}(x,y)d\mu(x)d\mu(y) &= \sum_{i,j=1}^{\ell} \gamma_i\gamma_j \int_{B_i}\int_{B_j} m(x^{(i)}, y^{(j)})d\mu(x)d\mu(y) \\
&= \sum_{i,j=1}^{\ell} [\gamma_i\mu(B_i)][\gamma_j\mu(B_j)]m(x^{(i)}, y^{(j)}) \geq 0.
\end{aligned}$$

It follows that $\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)d\mu(x)d\mu(y) \geq 0$ as well.

The converse follows by reversing the above reasoning. Suppose the second condition holds. Let $\varepsilon > 0$. Consider the function

$$f_\varepsilon := \sum_{i=1}^{p} \frac{\beta_i}{\mu(B(z^{(i)}, \varepsilon))} 1_{B(z^{(i)},\varepsilon)}(x).$$

(By assumption a division by zero does not occur.) Then

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f_\varepsilon(x) f_\varepsilon(y) m(x,y) d\mu(x) d\mu(y) = \sum_{i,j=1}^{p} \beta_i\beta_j \frac{\int_{B(z^{(i)},\varepsilon)}\int_{B(z^{(j)},\varepsilon)} m(x,y) d\mu(x) d\mu(y)}{\mu(B(z^{(i)},\varepsilon))\mu(B(z^{(j)},\varepsilon))}.$$

Since m is continuous,

$$\lim_{\varepsilon\to 0^+} \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f_\varepsilon(x) f_\varepsilon(y) \overline{m}(x,y) d\mu(x) d\mu(y) = \sum_{i,j=1}^{p} \beta_i\beta_j m(z^{(i)}, z^{(j)}).$$

Since the second condition holds, we conclude that $\sum_{i,j=1}^{p} \beta_i\beta_j m(z^{(i)}, z^{(j)}) \geq 0$. □

Exercise 2.3. For each kernel function $m\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ below, find an inner product space C and a map $\phi\colon \mathbb{R}^n \to C$ such that

$$m(x,y) = \langle \phi(x), \phi(y)\rangle_C, \qquad \forall\, x, y \in \mathbb{R}^n.$$

Conclude that each such m is a positive semidefinite function, in the sense stated in Mercer's Theorem.

- $m(x,y) := 1 + \langle x, y\rangle\ \forall\ x, y \in \mathbb{R}^n$.
- $m(x,y) := (1 + \langle x, y\rangle)^d\ \forall\ x, y \in \mathbb{R}^n$, where d is a fixed positive integer.
- $m(x,y) := \exp(-\|x-y\|^2)$.

Hint: it might be helpful to consider d-fold iterated tensor products of the form $x^{\otimes d} = x \otimes x \otimes \cdots \otimes x$, along with their corresponding inner products.

Solution. In the first case, we use $\phi(x) := (x, 1)$, $\phi\colon \mathbb{R}^n \to \mathbb{R}^{n+1}$, where $C := \mathbb{R}^{n+1}$ has the standard inner product. Then

$$\langle \phi(x), \phi(y)\rangle_C = \langle (x,1), (y,1)\rangle = \langle x, y\rangle + 1.$$

In the second case, we use $\phi(x) := (x,1)^{\otimes d}$, $\phi\colon \mathbb{R}^n \to \mathbb{R}^{d(n+1)}$, where $C := \mathbb{R}^{d(n+1)}$ has the standard inner product (so that $\langle x^{\otimes d}, y^{\otimes d}\rangle = \langle x, y\rangle^d$.) Then

$$\langle \phi(x), \phi(y)\rangle_C = \langle (x,1)^{\otimes d}, (y,1)^{\otimes d}\rangle_C = \langle (x,1), (y,1)\rangle^d = (\langle x, y\rangle + 1)^d.$$

In the final case, we let $C := \bigoplus_{d=0}^{\infty} \mathbb{R}^{dn}$, where for any $a = (a^{(0)}, a^{(1)}, \ldots)$, $b = (b^{(0)}, b^{(1)}, \ldots) \in C$, we define

$$\langle a, b\rangle_C := \sum_{d=0}^{\infty} \langle a^{(i)}, b^{(i)}\rangle_{\mathbb{R}^{dn}}.$$

Then, for any $x \in \mathbb{R}^n$, define $\phi\colon \mathbb{R}^n \to C$ by

$$\phi(x) := e^{-\|x\|^2}\Big(1, \frac{2^{1/2}}{\sqrt{1!}}x, \frac{2^{2/2}}{\sqrt{2!}}x^{\otimes 2}, \frac{2^{3/2}}{\sqrt{3!}}x^{\otimes 3}, \ldots\Big).$$

That is, the d^{th} coordinated of ϕ satisfies

$$\phi(x)_d = e^{-\|x\|^2}\frac{2^{d/2}}{\sqrt{d!}}x^{\otimes d}.$$

Then

$$\begin{aligned}\langle\phi(x),\phi(y)\rangle_C &= e^{-\|x\|^2-\|y\|^2}\sum_{d=0}^{\infty}\frac{2^d}{d!}\langle x^{\otimes d},y^{\otimes d}\rangle = e^{-\|x\|^2-\|y\|^2}\sum_{d=0}^{\infty}\frac{(2\langle x,y\rangle)^d}{d!}\\ &= e^{-\|x\|^2-\|y\|^2}e^{2\langle x,y\rangle} = e^{-\|x-y\|^2}\end{aligned}$$

□

Exercise 2.9. For any $f\in\mathcal{F}$, show that

$$\mathrm{VCdim}(\mathcal{F}) = \mathrm{VCdim}(D(f)).$$

(Recall: $\mathcal{F}$ is a subset of $\{0,1\}$-valued functions on a set A. Let $f,g\in\mathcal{F}$. Since $f = 1_{\{f=1\}}$, we can identify f with the set where it is 1 and extend set operations to functions in $\mathcal{F}$. For example, $f\Delta g := 1_{\{f=1\}\Delta\{g=1\}}$, where Δ denotes symmetric difference. And we define

$$D(f) := \{f\Delta g\colon g\in\mathcal{F}\}.)$$

Solution. Let $B\subseteq A$ be a set shattered by $\mathcal{F}$. Then, for any function $h\colon B\to\{0,1\}$, there exists $g\in\mathcal{F}$ such that $g|_B = h$. In particular, for any $q\in\mathcal{F}$, any function of the form $(f\Delta q)|_B$ has some $p\in\mathcal{F}$ such that $p|_B = (f\Delta q)|_B$. That is, if B is shattered by $D(f)$, then B is shattered by $\mathcal{F}$. It follows that

$$\mathrm{VCdim}(\mathcal{F}) \geq \mathrm{VCdim}(D(f)).$$

We now prove the other inequality. If B is shattered by $D(f)$, then for any $h\colon B\to\{0,1\}$, there exists $g\in\mathcal{F}$ such that $(f\Delta g)|_B = h|_B$. In particular, for any $q\in\mathcal{F}$, any function of the form $q|_B$ has some $p\in\mathcal{F}$ such that $(f\Delta p)|_B = q|_B$. That is, if B is shattered by $\mathcal{F}$, then B is shattered by $D(f)$. It follows that

$$\mathrm{VCdim}(\mathcal{F}) \leq \mathrm{VCdim}(D(f)).$$

□

USC Department of Mathematics, Los Angeles, CA
E-mail address: stevenmheilman@gmail.com