MATH 547 Homework 2

Qilin Ye

June 11, 2022

Problem 1

Let M be a $k \times k$ real symmetric matrix. Show that M is PSD if and only if there exist a real $k \times k$ matrix R such that $M = RR^T$. In either case, if r^i denotes the i^{th} row of R, we have

$$m_{i,j} = \langle r^i, r^j \rangle \qquad \text{for } 1 \leqslant i, j \leqslant k.$$

Proof. If $M = RR^T$ for some R, the for all $v \in \mathbb{R}^k$, $x^T M x = x^T RR^T x = \|R^T x\|^2 \geqslant 0$. Conversely, since M is symmetric, spectral theorem gives a decomposition $A = Q^T D Q^T$. Define $D^{1/2}$ to be the diagonal matrix with $D^{1/2}_{i,i} = \sqrt{D_{i,i}}$. Define $R := D^{1/2}Q$ and we have $M = (D^{1/2}Q)^T(D^{1/2}Q) = RR^T$.
Note that the claim can be generalized: M is PSD of ran $n \leqslant k$ if and only if there exists a $k \times k$ matrix R with rank n such that $M = RR^T$. □

Problem 2

Let μ be a Borel measure on $\mathbb{R}^n$ that assigners positive measure to open sets in $\mathbb{R}^n$. Let $m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be continuous with $\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} |m(x,y)|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y) < \infty$. Show that TFAE:

(1) For all $p \geqslant 1$, for all $z^{(1)}, \dots, z^{(p)} \in \mathbb{R}^n$, and for all $\beta_1, \dots, \beta_p \in \mathbb{R}$, we have

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

(2) For all $f \in L^2(\mu)$, we have

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

Proof. We first assume (1). Let $\epsilon > 0$, $f \in L^2(\mu)$ be given. Since simple functions are dense in L^2, there exists a (compactly supported, L^2-integrable) $g = \sum_{i=1}^{n} c_i 1_{E_i}$ with $\|f - g\|_2 < \epsilon$. Since the support of g is compact, m is uniformly continuous on $\operatorname{supp}(g)$ [support of g], so there exists $\delta > 0$ such that

$$\|(x_1, y_1) - (x_2, y_2)\| < \delta \implies |m(x_1, y_1) - m(x_2, y_2)| < \frac{\epsilon}{\mu(\operatorname{supp}(g))} \qquad \text{for } (x_i, y_i) \in \operatorname{supp}(g).$$

By compactness $\operatorname{supp}(g)$ can be covered by finitely many balls $\tilde{B}_1, \dots, \tilde{B}_k$ with radii $\delta/2$. By recursively deleting the overlapping parts (i.e., define $B_1 := \tilde{B}_1$ and $B_i := \tilde{B}_i \setminus \bigcup_{j=1}^{i-1} B_i$), we may assume that $\operatorname{supp}(g)$ is a union of finitely many disjoint sets, each with "radius" $\leqslant \delta/2$.

Let $b_1, ..., b_k$ be any points in $B_1, ..., B_k$, respectively, and define

$$\ell(x,y) := \begin{cases} m(b_i, b_j) & \text{if } x = b_i, y = b_j \text{ for some } i, j \\ 0 & \text{otherwise.} \end{cases}$$

Having crafted all recipes, we now bound the integral in (2). We first approximate f by g:

$$\begin{aligned}
\left| \iint f(x)f(y)m(x,y) - \iint g(x)g(y)m(x,y) \right| &= \left| \int f(x)f(y)m(x,y) - \int f(x)g(y)m(x,y) \right. \\
&\quad \left. + \iint f(x)g(y)m(x,y) - \iint g(x)g(y)m(x,y) \right| \\
&\leqslant \iint |f(x)||f(y) - g(y)||m(x,y)| + \iint |f(x) - g(x)||g(y)||m(x,y)| \\
&\leqslant \int |f(y) - g(y)| \left[\int |f(x)||m(x,y)| \, \mathrm{d}\mu(x) \right] \mathrm{d}\mu(y) + ... \\
&\leqslant \int |f(y) - g(y)| \cdot \|f\|_2 \cdot \left(\int |m(x,y)|^2 \, \mathrm{d}\mu(x) \right)^{1/2} \mathrm{d}\mu(y) + ... \\
&\leqslant \|f\|_2 \|f - g\|_2 \left(\iint |m(x,y)|^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \\
&\quad + \|g\|_2 \|f - g\|_2 \left(\iint |m(x,y)|^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \\
&< 3\epsilon \|f\|_2 \left(\iint |m(x,y)|^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \qquad (2.1)
\end{aligned}$$

if we WLOG assume $\epsilon < \|f\|_2$ so that $\|f\|_2 + \|g\|_2 < 3\|f\|_2$.

On the other hand, approximating m by ℓ gives

$$\begin{aligned}
\left| \iint g(x)g(y)m(x,y) - \iint g(x)g(y)\ell(x,y) \right| &\leqslant \iint |g(x)||g(y)||m(x,y) - \ell(x,y)| \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\
&\leqslant \|g\|_2^2 \left(\iint |m(x,y) - \ell(x,y)|^2 \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \\
&< \|g\|_2^2 \left(\iint_{\mathrm{supp}(g)^2} |m(x,y) - \ell(x,y)| \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \right)^{1/2} \\
&< \|g\|_2^2 \left(\sum_{i,j=1}^{k} \frac{\epsilon^2}{\mu(\mathrm{supp}(g))^2} \mu(B_i)\mu(B_j) \right)^{1/2} \\
&\leqslant \|g\|_2^2 \frac{\epsilon}{\mu(\mathrm{supp}(g))} \left(\sum_{i,j=1}^{k} \mu(B_i)\mu(B_j) \right)^{1/2} \\
&\leqslant \|g\|_2^2 \frac{\epsilon}{\mu(\mathrm{supp}(g))} \mu\Big(\bigcup_{i=1}^{k} B_i\Big)^{1/2} \mu\Big(\bigcup_{j=1}^{k} B_j\Big)^{1/2} \\
&= \epsilon \|g\|_2^2 < 4\epsilon \|f\|_2^2 \qquad (2.2)
\end{aligned}$$

where we again assume $\epsilon < \|f\|_2$ so $\|g\|_2^2 < 4\|f\|_2^2$. Combining the two inequalities above,

$$\left| \iint f(x)f(y)m(x,y) - \iint g(x)g(y)\ell(x,y) \right| < 4\epsilon \|f\|_2^2 + 3\epsilon \|f\|_2 \left(\iint |m(x,y)^2| \right)^{1/2}.$$

By assumption,

$$\begin{aligned}
\iint g(x)g(y)\ell(x,y) \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) &= \iint_{\mathrm{supp}(g)^2} g(x)g(y)\ell(x,y) \, \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\
&= \sum_{i,j=1}^{k} \mu(B_i)\mu(B_j)m(b_i, b_j) \geqslant 0.
\end{aligned}$$

Therefore $\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} f(x)f(y)m(x,y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y) \geqslant 0$, as claimed.

Conversely, let $z^{(1)}, \dots, z^p \in \mathbb{R}^n$ and $\beta_1, \dots, \beta_p \in \mathbb{R}$ be given. Define

$$f_n(x) := \sum_{i,j=1}^{p} \frac{\beta_i}{\mu(B(z^{(i)}, 1/n)} 1_{B(z^{(i)},1/n)}(x).$$

By definition,

$$0 \leqslant \iint f_n(x)f_n(y)m(x,y) = \sum_{i,j=1}^{p} \frac{\beta_i\beta_j}{\mu(B(z^{(i)}, 1/n)\mu(B(z^{(j)}, 1/n))} \iint_{B(z^{(i)},1/n)\times B(z^{(j)},1/n)} m(x,y).$$

Let $n \to 0$. Since m is continuous,

$$\lim_{n\to\infty} \frac{1}{\mu(B(z^{(i)}, 1/n)\mu(B(z^{(j)}, 1/n))} \iint_{B(z^{(i)},1/n)\times B(z^{(j)},1/n)} m(x,y) = m(z^{(i)}, z^{(j)}).$$

Therefore $\sum_{i,j=1}^{p} \beta_i\beta_j m(z^{(i)}, z^{(j)}) \geqslant 0$, completing the proof. □

Problem 3

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

$$m(x,y) = \langle \varphi(x), \varphi(y) \rangle_C \qquad \text{for all } x, y \in \mathbb{R}^n.$$

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

(1) $m(x,y) := 1 + \langle x, y \rangle$ for $x, y \in \mathbb{R}^n$,

(2) $m(x,y); = (1 + \langle x, y \rangle)^d$ for $x, y \in \mathbb{R}^n$ where d is a fixed integer, and

(3) $m(x,y) := \exp(-\|x-y\|^2)$.

Solution. (1) Define $\varphi : \mathbb{R}^n \to \mathbb{R}^{n+1}$ by $(x_1, \dots, x_n) \mapsto (x_1, \dots, x_n, 1)$. Then

$$\langle \varphi(x), \varphi(y) \rangle = \langle (x_1, \dots, x_n, 1), (y_1, \dots, y_n, 1) \rangle = \langle x, y \rangle + 1.$$

(2) Define $\varphi : \mathbb{R}^n \to \mathbb{R}^{(n+1)^d}$ by $(x_1, \dots, x_n) \mapsto (x_1, \dots, x_n, 1)^{\otimes d}$, where $\otimes$ takes vectors of lengths m and k and outputs a vector of length mk:

$$(v_1, \dots, v_m) \otimes (w_1, \dots, w_k) := (v_1w_1, \dots, v_1w_k, v_2w_1, \dots, v_mw_1, \dots, v_mw_k).$$

Then $\left\langle u^{\otimes d}, v^{\otimes d} \right\rangle = \langle u, v \rangle^d$, so $\langle \varphi(x), \varphi(y) \rangle = \langle (x_1, \dots, x_n, 1), (y_1, \dots, y_n, 1) \rangle^d = (\langle x, y \rangle + 1)^d$.

(3) We note that $\exp(-\|x-y\|^2) = \exp(-\|x\|^2 - \|y\|^2 + 2\langle x, y \rangle) = \exp(-\|x\|^2)\exp(-\|y\|^2) \sum_{d=0}^{\infty} \frac{(2\langle x, y \rangle)^d}{d!}$. To this end, using (2), we define $\varphi : \mathbb{R}^n \to \bigotimes_{d=0}^{\infty} \mathbb{R}^{n^d}$ by

$$\varphi(x) := \exp(-\|x\|^2)\left(1, \sqrt{2^1/1!} \cdot x, \sqrt{2^2/2!} \cdot x^{\otimes 2}, \sqrt{2^3/3!} \cdot x^{\otimes 3}, \dots\right)$$

with inner product

$$\left\langle \{u_i\}_{i=1}^{\infty}, \{v_j\}_{j=1}^{\infty} \right\rangle := \sum_{d=0}^{\infty} \langle u_d, v_d \rangle$$

where $\langle u_d, v_d\rangle$ uses the standard inner product in $\mathbb{R}^{n^d}$. Then from (2)

$$\begin{aligned}\langle \varphi(x), \varphi(y)\rangle &= \exp(-\|x\|^2 - \|y\|^2)\sum_{d=0}^{\infty}\frac{2^d}{d!}\left\langle x^{\otimes d}, y^{\otimes d}\right\rangle \\ &= \exp(-\|x\|^2 - \|y\|^2)\sum_{d=0}^{\infty}\frac{2^d\langle x, y\rangle^d}{d!} = \\ &= \exp(-\|x\|^2 - \|y\|^2)\exp(2\langle x, y\rangle) = \exp(-\|x-y\|^2).\end{aligned}$$

Problem 4

Show that the set of conjunctions is contained in the set of linear threshold functions.

Proof. Let let $I, J \subset \{1, \dots, n\}$ and let $f : \{0,1\}^n \to \{0,1\}$ be the corresponding boolean conjunction:

$$f(x_1, \dots, x_n) := \prod_{i\in I} x_i \prod_{j\in J}(1 - x_j).$$

We define $w \in \mathbb{R}^n$ by $w_k = 1$ if $k \in I$, $w_k = -1$ if $k \in J$, and $w_k = 0$ otherwise, and we define t to be the cardinality of I. Then

$$\langle x, w\rangle = \sum_{i\in I} x_i - \sum_{j\in J} x_j \leqslant |I|$$

with equality attained if and only if $\sum_{i\in I} x_i = |I|$ and $\sum_{j\in J} x_j = 0$, that is, $x_i = 1$ for all $i \in I$ and $x_j = 0$ for all $j \in J$. This is equivalent to saying $f(x_1, \dots, x_n) = 1$. □

Problem 5

Let X be a real-valued random variable and let $X_1, X_2, \dots$ be independent copies of X. Let $a < b$ and suppose $\mathbb{P}(a \leqslant X \leqslant b) > 3/4$. For $n \in \mathbb{N}$, let Y_n be a median of $X_1, \dots, X_n$. Show that

$$\mathbb{P}(a \leqslant Y_n \leqslant b) \geqslant 1 - \sum_{j=\lfloor n/2\rfloor}^{n}\binom{n}{j}\alpha^j$$

where $\alpha := \mathbb{P}(X \notin [a, b])$.

Show additionally that

$$\mathbb{P}(a \leqslant Y_n \leqslant b) \geqslant 1 - (1 + o(1))\frac{\sqrt{\pi}}{\sqrt{2n}}\frac{2^n\alpha^{\lfloor n/2\rfloor}}{1-\alpha} \geqslant 1 - (4\alpha)^{\lfloor n/2\rfloor}\cdot O(1).$$

Proof. If $Y_n \notin [a, b]$, then either $Y_n < a$ or $Y_n > b$. That is, either at least $\lfloor n/2\rfloor$ samples are $< a$ or at least $\lfloor n/2\rfloor$ are $> b$. Such events are a subset of "at least $\lfloor n/2\rfloor$ samples are $< a$ or $> b$" which translates to the RHS. □

Problem 6

Explain why taking the expected value of the inequality for the average number of mis-classifications of Adaboost does not guarantee PAC learning.

Solution. The probability distribution $\mathbb{P}$ is not given, so we do not know the exact distribution of X, and so we cannot compute $\mathbb{P}(g(X) \neq f(X))$ as required in PAC's definition. Instead we can only analyze empirical risk as stated in the notes.

Problem 7

Show that the Sauer-Shelah lemma is sharp for all n, d. That is, find $\mathcal{F}$ with $d := \text{VCdim}(\mathcal{F})$ such that

$$|\mathcal{F}| = \sum_{i=0}^{d} \binom{n}{i}.$$

Proof. Let n, d be given. Following the hint, we define $\mathcal{F}$ as the collection of functions $f : \{0,1\}^n \to \{0,1\}$ of the following form:

$$f(x) = \prod_{i \in I} x_i \prod_{j \notin I} (1 - x_j)$$

where $I \subset \{1, \ldots, n\}$ with $|I| \leqslant d$. It is clear that different I's give rise to different functions, so $|\mathcal{F}| = \sum_{i=0}^{d} \binom{n}{i}$. It remains to verify that $\text{VCdim}(\mathcal{F}) = d$. Clearly any set of size d is shattered by $\mathcal{F}$. If $B \subset \{0,1\}^n$ is one such set and $g : B \to \{0,1\}$, then the function

$$f(x) = f(x_1, \ldots, x_n) := \begin{cases} g(x) & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$$

is a valid Boolean function mapping at most d inputs to 1 and is therefore in $\mathcal{F}$. Conversely, if $B' \subset \{0,1\}^n$ has $> d$ elements, the constant function 1 on B' cannot be extended to any function in $\mathcal{F}$. Therefore $\text{VCdim}(\mathcal{F}) = d$, and we are done. □

Problem 8

Show that both notions of ϵ-net agree up to changing the constant ϵ in the following sense. Let A be a metric space, $\mathbb{P}$ a probability law on A, Ω the collection of balls with arbitrary center and radius. Assume there exist $a, b, c_1, c_2 > 0$ such that $c_1 r^a \leqslant \mathbb{P}(B(x,r)) \leqslant c_2 r^b$ for all $x \in A, r > 0$. Then S is a measure theoretic ϵ-net for Ω if and only if there is an ϵ'-net for Ω with respect to the metric d on A.

Proof. We first show (ϵ-net $\Rightarrow$ measure theoretic ϵ-net). Let $\{x_i\}_{i \in I}$ be an ϵ-net and let $B(x,r) \in \Omega$. By assumption

$$c_1 \epsilon^a \leqslant \mathbb{P}(B(x,\epsilon)) \leqslant c_2 \epsilon^b, \tag{*}$$

so if $\mathbb{P}(B) > c_2 \epsilon^b$, it is guaranteed that the radius of $B > \epsilon$. Then by definition of ϵ-net there exists $x_i \in \{x_i\}$ whose distance to the center of the ball is $< \epsilon$.

Conversely, if $\{x_i\}$ is a measure theoretic "$c_1\epsilon^a$-net", by assumption, if $\mathbb{P}(B) > c_1 \epsilon^a$ there exists $x_i \in B$. By (*) this means that for every ball with radius ϵ, there exists some x_i within distance ϵ from the center of the ball. Since the centers are arbitrary we see $\bigcup_{i \in I} B(x_i, \epsilon)$ covers A, completing our proof. □

Problem 9

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

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

Proof. For convenience we denote VCdim($\mathcal{F}$) by d. Let $B \subset A$ be a set with cardinality b shattered by $\mathcal{F}$, and let $\varphi : B \to \{0,1\}$ be any Boolean function. We define a function $g_0 : B \to \{0,1\}$ by

$$g_0(b) := \begin{cases} f(b) & \text{if } \varphi(b) = 0 \\ 1 - f(b) & \text{if } \varphi(b) = 1. \end{cases}$$

By doing so, $(f \Delta g_0)(b) = 1$ if and only if $\varphi(b) = 1$. Since $\mathcal{F}$ shatters B, we can extend g_0 to a Boolean function $g \in \mathcal{F}$ on A. Thus $\varphi \equiv (f\Delta g)\big|_B$, B is shattered by $D(f)$, and $\text{VCdim}(D(f)) \geqslant d = \text{VCdim}(\mathcal{F})$.
Conversely, suppose $C \subset A$ is shattered by $D(f)$. Let $\psi : C \to \{0,1\}$ be any Boolean function. By assumption there exists $g \in \mathcal{F}$ with $(f\Delta g)\big|_C \equiv \psi$. We claim that this implies $g\big|_C = f\big|_C \Delta \psi$. This can be easily verified via brute force:

f	ψ	g	$f\Delta\psi$
0	0	0	0
0	1	1	1
1	0	1	1
1	1	0	0

Of course, $g\big|_C$ is a Boolean function on C, so by assumption there exists $\psi' \in \mathcal{F}$ such that $g\big|_C = (f\Delta\psi')\big|_C$. From this we see $\psi'\big|_C = \psi$, so indeed ψ admits an extension with domain on all of A. Therefore whatever is shattered by $D(f)$ is also shattered by $\mathcal{F}$, i.e., $\text{VCdim}(D(f)) \leqslant d = \text{VCdim}(\mathcal{F})$. □