Sample midterm questions

May 6, 2022

Question : PAC learning. Consider the concept class C formed by threshold functions on the real line, $C = \{[c, \infty) : \forall c \in \mathbb{R}\} \cup \{(-\infty, c] : \forall c \in \mathbb{R}\}$. Give a PAC-learning algorithm for C. The analysis is similar to that of the axis-aligned rectangles given in class, but you should carefully present and justify your proof.

Solution. Let $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ denote the labeled sample of size m. Without loss of generality, assume that the true concept is $[c, \infty)$ for some unknown $c \in \mathbb{R}$. Define

$$\begin{aligned}\widehat{l} &= \max\{x_i : (x_i, y_i) \in S, y_i = -1\}\\ \widehat{r} &= \min\{x_i : (x_i, y_i) \in S, y_i = 1\}\end{aligned}$$

By definition, $\hat{l} \leq c \leq \widehat{r}$. The algorithm returns the concept $R_S = [\widehat{c}, \infty)$ with $\widehat{c} = (\widehat{l} + \widehat{r})/2$. The error region of R_S is the interval $[\widehat{c}, c)$ when $\widehat{c} < c$, and $[c, \widehat{c})$ otherwise. In both cases, the error region is a subset of $(\widehat{l}, \widehat{r})$. Therefore,

$$\begin{aligned}\Pr[R(R_S) > \epsilon] &\leq \Pr[R((\widehat{l}, \widehat{r})) > \epsilon]\\ &\leq (1-\epsilon)^m \leq e^{-m\epsilon}\end{aligned}$$

Setting δ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$. □

Question : Growth function. A linearly separable labeling of a set X of vectors in $\mathbb{R}^d$ is a classification of X into two sets $X^+$and $X^-$with $X^+ = \{\mathbf{x} \in X : \mathbf{w} \cdot \mathbf{x} > 0\}$ and $X^- = \{\mathbf{x} \in X : \mathbf{w} \cdot \mathbf{x} < 0\}$ for some $\mathbf{w} \in \mathbb{R}^d$. Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$.

(a) Let $\{X^+, X^-\}$be a dichotomy of X and let $\mathbf{x}_{m+1} \in \mathbb{R}^d$. Show that $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if $\{X^+, X^-\}$is linearly separable by a hyperplane going through the origin and $\mathbf{x}_{m+1}$.

Solution. $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if there exists $\mathbf{w}_1 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in X^+, \mathbf{w}_1 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in X^-, \mathbf{w}_1 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_1 \cdot \mathbf{x}_{m+1} > 0 \tag{1}$$

and there exists $\mathbf{w}_2 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in X^+, \mathbf{w}_2 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in X^-, \mathbf{w}_2 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_2 \cdot \mathbf{x}_{m+1} < 0. \tag{2}$$

For any $\mathbf{w}_1, \mathbf{w}_2$, the function $f : (t \mapsto t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \cdot \mathbf{x}_{m+1}$ is continuous over $[0, 1]$. (1) and (2) hold iff $f(0) < 0$ and $f(1) > 0$, that is iff there exists $\mathbf{w} = t_0\mathbf{w}_1 + (1-t_0)\mathbf{w}_2$ linearly separating $\{X^+, X^-\}$and such at $\mathbf{w} \cdot \mathbf{x}_{m+1} = 0$. □

(b) Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$ such that any k-element subset of X with $k \leq d$ is linearly independent. Then, the number of linearly separable labelings of X is $C(m, d) = 2\sum_{k=0}^{d-1}\binom{m-1}{k}$.

Solution. Repeating the formula, we obtain $C(m, d) = \sum_{k=0}^{m-1}\binom{m-1}{k}C(1, d-k)$. Since, $C(1, n) = 2$ if $n \geq 1$ and $C(1, n) = 0$ otherwise, the result follows. □

Question : Growth function. Consider the family H of threshold functions over $\mathbb{R}^N$ defined by $\{\mathbf{x} = (x_1, \ldots, x_N) \mapsto \operatorname{sgn}(x_i - \theta) : i \in [1, N], \theta \in \mathbb{R}\}$, where $\operatorname{sgn}(z) = +1$ if $z \geq 0, \operatorname{sgn}(z) = -1$ otherwise. Give an explicit upper bound on the growth function $\Pi_H(m)$ of H that is in $O(mN)$.

Solution. For each feature, x_j, there at most $m+1$ ways of selecting the threshold (between any two feature values or beyond or below all values). Thus, the total number of thresholds functions for a sample of size m is at most $(m+1)N$. Thus, the growth function is upper bounded by $(m+1)N$. $\square$

Question : VC dimension. Let C_1 and C_2 be two concept classes. Show that for any concept class $C = \{c_1 \cap c_2 : c_1 \in C_1, c_2 \in C_2\}$,

$$\Pi_C(m) \leq \Pi_{C_1}(m)\Pi_{C_2}(m). \tag{1}$$

Solution. Fix a set X of m points. Let $Y_1, \ldots, Y_k$ be the set of intersections of the concepts of C_1 with X. By definition of $\Pi_{C_1}(X), k \leq \Pi_{C_1}(X) \leq \Pi_{C_1}(m)$. By definition of $\Pi_{C_2}(Y_i)$, the intersection of the concepts of C_2 with Y_i are at most $\Pi_{C_2}(Y_i) \leq \Pi_{C_2}(m)$. Thus, the number of sets intersections of concepts of C with X is at most

$$k\Pi_{C_2}(Y_i) \leq \Pi_{C_1}(m)\Pi_{C_2}(m).$$

$\square$

Question : VC dimension. Let C be a concept class with VC dimension d and let C_s be the concept class formed by all intersections of s concepts from C, $s \geq 1$. Show that the VC dimension of C_s is bounded by $2ds\log_2(3s)$

Solution. In view of the result proved in the previous question, $\Pi_{C_s}(m) \leq (\Pi_{C_1}(m))^s$. By Sauer's lemma, this implies

$$\Pi_{C_s}(m) \leq \left(\frac{em}{d}\right)^{sd}.$$

If $\left(\frac{em}{d}\right)^{sd} < 2^m$, then the VC dimension of C_s is less than m. Thus, it suffices to show this inequality holds with $m = 2ds\log_2(3s)$. Plugging in that value for m and taking the $\log_2$ yield:

$$\begin{aligned}
&ds\log_2(2es\log_2(3s)) < 2ds\log_2(3s) \\
&\Leftrightarrow \log_2(2es\log_2(3s)) < 2\log_2(3s) = \log_2\left(9s^2\right) \\
&\Leftrightarrow 2es\log_2(3s) < 9s^2 \\
&\Leftrightarrow \log_2(3s) < \frac{9s}{2e}
\end{aligned}$$

This last inequality holds for $s = 2 : \log_2(6) \approx 2.6 < 9/(2e) \approx 3.3$. Since the functions corresponding to the left-hand-side grows more slowly than the one corresponding to the right-hand-side (compare derivatives for example), this implies that the inequality holds for all $s \geq 2$ $\square$

Question : VC dimension. Let H and H' be two families of functions mapping from X to $\{0, 1\}$ with finite VC dimensions. Show that

$$\operatorname{VCdim}(H \cup H') \leq \operatorname{VCdim}(H) + \operatorname{VCdim}(H') + 1$$

Use that to determine the VC dimension of the hypothesis set formed by the union of axis-aligned rectangles and triangles in dimension 2 .

Solution. The number of ways m particular points can be classified using $H \cup H'$ is at most the number of classifications using H plus the number of classifications using H'. This gives immediately the following inequality for growth functions for any $m \geq 0$:

$$\Pi_{H' \cup H}(m) \leq \Pi_H(m) + \Pi_{H'}(m).$$

Let VCdim$(H) = d$ and VCdim $(H') = d'$. Then, by Sauer's lemma,

$$\Pi_{H' \cup H}(m) \leq \sum_{i=0}^{d} \binom{m}{i} + \sum_{i=0}^{d'} \binom{m}{i}$$

Using the identity $\binom{m}{i} = \binom{m}{m-i}$ and a change of variable, this can be rewritten as

$$\Pi_{H' \cup H}(m) \leq \sum_{i=0}^{d} \binom{m}{i} + \sum_{i=0}^{d'} \binom{m}{m-i} \leq \sum_{i=0}^{d} \binom{m}{i} + \sum_{i=m-d'}^{m} \binom{m}{i}$$

Now, if $m - d' > d + 1$, that is $m \geq d + d' + 2$,

$$\Pi_{H' \cup H}(m) \leq \sum_{i=0}^{m} \binom{m}{i} - \binom{m}{d+1} = 2^m - \binom{m}{d+1} < 2^m$$

Thus, the VC dimension of $H \cup H'$ cannot be greater than or equal to $d + d' + 2$, which implies VCdim $(H \cup H') \leq d + d' + 1$.

Now, the VC dimension of axis-aligned rectangles in dimension 2 is 4 and the VC dimension of triangles (3-gones) is 7 . Thus, the VC dimension of the union of these sets is bounded by $4 + 7 + 1 = 12$. □

Question VC dimension. Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i$ VCdim $(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$.

(a) Prove that

$$\text{VCdim}\left(\cup_{i=1}^{r} \mathcal{H}_i\right) \leq 4d \log(2d) + 2 \log(r)$$

Solution. We may assume w.l.o.g. that for each $i \in [r]$, VCdim $(\mathcal{H}_i) = d \geq 3$. Let $\mathcal{H} = \bigcup_{i=1}^{r} \mathcal{H}_i$. Let $k \in [d]$, such that $\Pi_{\mathcal{H}}(k) = 2^k$. We will show that $k \leq 4d \log(2d) + 2 \log r$. By definition of the growth function, we have

$$\Pi_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \Pi_{\mathcal{H}_i}(k)$$

Since $d \geq 3$, by applying Sauer's lemma on each of the terms $\Pi_{\mathcal{H}_i}$, we obtain

$$\Pi_{\mathcal{H}}(k) < r m^d$$

It follows that $k < d \log m + \log r$. Lemma A.2 implies that $k < 4d \log(2d) + 2 \log r$. □

(b) Prove that for $r = 2$ it holds that

$$\text{VCdim}\left(\mathcal{H}_1 \cup \mathcal{H}_2\right) \leq 2d + 1$$

Solution. A direct application of the result above yields a weaker bound. We need to employ a more careful analysis. As before, we may assume w.l.o.g. that VCdim $(\mathcal{H}_1) =$ VCdim $(\mathcal{H}_2) =$ d. Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. Let k be a positive integer such that $k \geq 2d + 2$. We show that $\Pi_{\mathcal{H}}(k) < 2^k$. By Sauer's lemma,

$$\begin{aligned}
\Pi_{\mathcal{H}}(k) &\leq \Pi_{\mathcal{H}_1}(k)+\Pi_{\mathcal{H}_2}(k) \\
&\leq \sum_{i=0}^{d}\binom{k}{i}+\sum_{i=0}^{d}\binom{k}{i} \\
&= \sum_{i=0}^{d}\binom{k}{i}+\sum_{i=0}^{d}\binom{k}{k-i} \\
&= \sum_{i=0}^{d}\binom{k}{i}+\sum_{i=k-d}^{k}\binom{k}{i} \\
&\leq \sum_{i=0}^{d}\binom{k}{i}+\sum_{i=d+2}^{k}\binom{k}{i} \\
&< \sum_{i=0}^{d}\binom{k}{i}+\sum_{i=d+1}^{k}\binom{k}{i} \\
&= \sum_{i=0}^{k}\binom{k}{i} \\
&= 2^k
\end{aligned}$$

□

Question : Rademacher complexity. Non-negativity of empirical Rademacher complexity: Show that for any hypothesis set $\mathcal{H}$ and sample S, we have $\widehat{\mathcal{R}}_S(\mathcal{H}) \geq 0$.

Solution. By the sub-additivity of supremum, we can write:

$$\begin{aligned}
\widehat{\mathcal{R}}_S(\mathcal{H}) &= \frac{1}{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h\left(x_i\right)\right] \\
&\geq \frac{1}{m}\sup_{h\in\mathcal{H}}\sigma\left[\sum_{i=1}^{m}\sigma_i h\left(x_i\right)\right] \\
&= \frac{1}{m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\underset{\boldsymbol{\sigma}}{\mathbb{E}}\left[\sigma_i\right]h\left(x_i\right)=0.
\end{aligned}$$

□

Question : Rademacher complexity. What is the Rademacher complexity of a hypothesis set reduced to a single hypothesis? An alternative definition of the Rademacher is based on absolute values: $\mathcal{R}'(H) = \frac{1}{m}\mathrm{E}_{\boldsymbol{\sigma},S}\left[\sup_{h\in H}\left|\sum_{i=1}^{m}\sigma_i h\left(x_i\right)\right|\right]$. Show the following upper bound for a hypothesis set reduced to a single hypothesis h :

$$\mathcal{R}'(\{h\}) \leq \sqrt{\frac{\mathrm{E}_{x\sim D}\left[h^2(x)\right]}{m}}.$$

Solution. Let h be that single hypothesis. By definition,

$$\mathcal{R}(\{h\}) = \frac{1}{m}\underset{\boldsymbol{\sigma},S}{\mathrm{E}}\left[\sum_{i=1}^{m}\sigma_i h\left(x_i\right)\right] = \frac{1}{m}\underset{S}{\mathrm{E}}\left[\sum_{i=1}^{m}\underset{\boldsymbol{\sigma}}{\mathrm{E}}\left[\sigma_i\right]h\left(x_i\right)\right]=0,$$

since $\mathrm{E}_\sigma\left[\sigma_i\right]=0$ for all $i\in[1,m]$. Using Jensen's inequality, with the alternative definition the Rademacher complexity can be bounded as follows:

$$
\begin{aligned}
\mathcal{R}'(\{h\}) &= \frac{1}{m} \underset{\boldsymbol{\sigma},S}{\mathrm{E}}\left[\left|\sum_{i=1}^{m} \sigma_i h\left(x_i\right)\right|\right] \\
&= \frac{1}{m} \underset{\boldsymbol{\sigma},S}{\mathrm{E}}\left[\sqrt{\left|\sum_{i=1}^{m} \sigma_i h\left(x_i\right)\right|^2}\right] \\
&\leq \frac{1}{m} \sqrt{\underset{\boldsymbol{\sigma},S}{\mathrm{E}}\left[\left|\sum_{i=1}^{m} \sigma_i h\left(x_i\right)\right|^2\right]} && \text{(by Jensen's inequality)} \\
&= \frac{1}{m} \sqrt{\underset{\boldsymbol{\sigma},S}{\mathrm{E}}\left[\sum_{i,j=1}^{m} \sigma_i \sigma_j h\left(x_i\right) h\left(x_j\right)\right]} \\
&= \frac{1}{m} \sqrt{\underset{S}{\mathrm{E}}\left[\sum_{i=1}^{m} h\left(x_i\right)^2\right]} && (\mathrm{E}\left[\sigma_i \sigma_j = 0\right] \text{ for } i \neq j\ ) \\
&= \frac{1}{m} \sqrt{m \underset{S}{\mathrm{E}}\left[h\left(x_1\right)^2\right]} = \sqrt{\frac{\mathrm{E}_x\left[h^2(x)\right]}{m}}. && \text{(i.i.d. sample)}
\end{aligned}
$$

□

Question : Rademacher complexity. Consider the trivial hypothesis set $\mathcal{H} = \{h_0\}$.

(a) Show that $\mathcal{R}_m(\mathcal{H}) = 0$ for any $m > 0$.

Solution. By definition of the Rademacher complexity:

$$
\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{H}) &= \underset{\sigma}{\mathbb{E}}\left[\max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h\left(x_i\right)\right] \\
&= \underset{\sigma}{\mathbb{E}}\left[\frac{1}{m} \sum_{i=1}^{m} \sigma_i h_0\left(x_i\right)\right] && (\mathcal{H} = \{h_0\}) \\
&= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\sigma_i\right] h_0\left(x_i\right) && (\sigma \text{ is independent of } h_0) \\
&= 0. && (\mathbb{E}\left[\sigma_i\right] = 0)
\end{aligned}
$$

□

(b) Use a similar construction to show that Massart's lemma is tight.

Solution. By Massart's lemma, for each finite hypothesis set $\mathcal{H}$, we have:

$$
\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{H}) &\leqslant \sqrt{\frac{2 \ln |\mathcal{H}|}{m}} \\
&= \sqrt{\frac{2 \ln |1|}{m}} && (|\mathcal{H}| = |\{h_0\}| = 1) \\
&= 0.
\end{aligned}
$$

□

Question : Hard versus soft SVM. Prove or refute the following claim: There exists $\lambda > 0$ such that for every sample S of $m > 1$ examples, which is separable by the class of homogenous halfspaces, the hard-SVM and the soft-SVM (with parameter λ) learning rules return exactly the same weight vector.

Solution. The claim is wrong. Fix some integer $m > 1$ and $\lambda > 0$. Let $\mathbf{x}_0 = (0, \alpha) \in \mathbb{R}^2$, where $\alpha \in (0,1)$ will be tuned later. For $k = 1, \ldots, m-1$, let $\mathbf{x}_k = (0, k)$. Let $y_0 = \ldots = y_{m-1} = 1$. Let $S = \{(\mathbf{x}_i, y_i) : i \in \{0, 1, \ldots, m-1\}\}$. The solution of hard-SVM is $\mathbf{w} = (0, 1/\alpha)$ (with value $1/\alpha^2$). However, if

$$\lambda \cdot 1 + \frac{1}{m}(1 - \alpha) \leq \frac{1}{\alpha^2},$$

the solution of soft-SVM is $\mathbf{w} = (0, 1)$. Since $\alpha \in (0, 1)$, it suffices to require that $\frac{1}{\alpha^2} > \lambda + 1/m$. Clearly, there exists $\alpha_0 > 0$ s.t. for every $\alpha < \alpha_0$, the desired inequality holds. Informally, if α is small enough, then soft-SVM prefers to "neglect" $\mathbf{x}_0$. $\square$

Question : Kernel. Let N be any positive integer. For every $x, x' \in \{1, \ldots, N\}$ define

$$K(x, x') = \min\{x, x'\}.$$

Prove that K is a valid kernel; namely, find a mapping $\psi : \{1, \ldots, N\} \to H$ where H is some Hilbert space, such that

$$\forall x, x' \in \{1, \ldots, N\}, K(x, x') = \langle \psi(x), \psi(x') \rangle$$

Solution. Define $\psi : \{1, \ldots, N\} \to \mathbb{R}^N$ by

$$\psi(j) = \left(\mathbf{1^j}; \mathbf{0^{N-j}}\right),$$

where $\mathbf{1^j}$ is the vector in $\mathbb{R}^j$ with all elements equal to 1 , and $\mathbf{0^{N-j}}$ is the zero vector in $\mathbb{R}^{N-j}$. Then, assuming the standard inner product, we obtain that $\forall (i,j) \in [N]^2$,

$$\langle \psi(i), \psi(j) \rangle = \left\langle \left(\mathbf{1^i}; \mathbf{0^{N-i}}\right), \left(\mathbf{1^j}; \mathbf{0^{N-j}}\right) \right\rangle = \min\{i, j\} = K(i, j).$$

$\square$

Question : Kernel. Let $\mathcal{X}$ be an instance set and let ψ be a feature mapping of $\mathcal{X}$ into some Hilbert feature space V. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function that implements inner products in the feature space V.

Consider the binary classification algorithm that predicts the label of an unseen instance according to the class with the closest average. Formally, given a training sequence $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, for every $y \in \{\pm 1\}$ we define

$$c_y = \frac{1}{m_y} \sum_{i : y_i = y} \psi(\mathbf{x}_i).$$

where $m_y = |\{i : y_i = y\}|$. We assume that $m_+$and $m_-$are nonzero. Then, the algorithm outputs the following decision rule:

$$h(\mathbf{x}) = \begin{cases} 1 & \|\psi(\mathbf{x}) - c_+\| \leq \|\psi(\mathbf{x}) - c_-\| \\ 0 & \text{otherwise.} \end{cases}$$

1. Let $\mathbf{w} = c_+ - c_-$and let $b = \frac{1}{2}\left(\|c_-\|^2 - \|c_+\|^2\right)$. Show that

$$h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \psi(\mathbf{x}) \rangle + b).$$

2. Show how to express $h(\mathbf{x})$ on the basis of the kernel function, and without accessing individual entries of $\psi(\mathbf{x})$ or $\mathbf{w}$.

Solution. We will work with the label set $\{\pm 1\}$.

1) Observe that

$$\begin{aligned}
h(\mathbf{x}) &= \operatorname{sign}\left(\|\psi(\mathbf{x})-c_-\|^2 - \|\psi(\mathbf{x})-c_+\|^2\right) \\
&= \operatorname{sign}\left(2\langle\psi(\mathbf{x}),c_+\rangle - 2\langle\psi(\mathbf{x}),c_-\rangle + \|c_-\|^2 - \|c_+\|^2\right) \\
&= \operatorname{sign}(2(\langle\psi(\mathbf{x}),\mathbf{w}\rangle + b)) \\
&= \operatorname{sign}(\langle\psi(\mathbf{x}),\mathbf{w}\rangle + b)
\end{aligned}$$

2) Simply note that

$$\begin{aligned}
\langle\psi(\mathbf{x}),\mathbf{w}\rangle &= \langle\psi(\mathbf{x}),c_+ - c_-\rangle \\
&= \frac{1}{m_+}\sum_{i:y_i=1}\langle\psi(\mathbf{x}),\psi(\mathbf{x}_i)\rangle + \frac{1}{m_-}\sum_{i:y_i=-1}\langle\psi(\mathbf{x}),\psi(\mathbf{x}_i)\rangle \\
&= \frac{1}{m_+}\sum_{i:y_i=1}K(\mathbf{x},\mathbf{x}_i) + \frac{1}{m_-}\sum_{i:y_i=-1}K(\mathbf{x},\mathbf{x}_i)
\end{aligned}$$

□

Question : Kernel. Let σ be a positive real number. K is defined by $K(x,y) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}$ for all $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$ (Hint: you could show that K is the normalized kernel of a kernel K' and show that K' is PDS using the following equality: $\|\mathbf{x}-\mathbf{y}\| = \frac{1}{2\Gamma\left(\frac{1}{2}\right)}\int_0^{+\infty}\frac{1-e^{-t\|\mathbf{x}-\mathbf{y}\|^2}}{t^{\frac{3}{2}}}dt$ valid for all $\mathbf{x},\mathbf{y}$).

Solution. It suffices to show that K is the normalized kernel associated to the kernel K' defined by

$$\forall(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N, K'(\mathbf{x},\mathbf{y}) = e^{\phi(\mathbf{x},\mathbf{y})}$$

where $\phi(\mathbf{x},\mathbf{y}) = \frac{1}{\sigma}[\|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|]$, and to show that K' is PDS. For the first part, observe that

$$\frac{K'(\mathbf{x},\mathbf{y})}{\sqrt{K'(\mathbf{x},\mathbf{x})K'(\mathbf{y},\mathbf{y})}} = e^{\phi(\mathbf{x},\mathbf{y})-\frac{1}{2}\phi(\mathbf{x},\mathbf{x})-\frac{1}{2}\phi(\mathbf{y},\mathbf{y})} = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}.$$

To show that K' is PDS, it suffices to show that ϕ is PDS, since composition with a power series with non-negative coefficients (here exp) preserve the PDS property. Now, for any $c_1,\ldots,c_n \in \mathbb{R}$, let $c_0 = -\sum_{i=1}^n c_i$, then, we can write

$$\begin{aligned}
\sum_{i,j=1}^n c_ic_j\phi(\mathbf{x}_i,\mathbf{x}_j) &= \frac{1}{\sigma}\sum_{i,j=1}^n c_ic_j\left[\|\mathbf{x}_i\| + \|\mathbf{x}_j\| - \|\mathbf{x}_i-\mathbf{x}_j\|\right] \\
&= \frac{1}{\sigma}\left[-\sum_{i=1}^n c_0c_i\|\mathbf{x}_i\| + -\sum_{i=1}^n c_0c_j\|\mathbf{x}_j\| - \sum_{i,j=1}^n c_ic_j\|\mathbf{x}_i-\mathbf{x}_j\|\right] \\
&= -\frac{1}{\sigma}\sum_{i,j=0}^n c_ic_j\|\mathbf{x}_i-\mathbf{x}_j\|
\end{aligned}$$

with $\mathrm{x}_0 = 0$. Now, for any $z \in \mathbb{R}$, the following equality holds:

$$z^{\frac{1}{2}} = \frac{1}{2\Gamma\left(\frac{1}{2}\right)}\int_0^{+\infty}\frac{1-e^{-tz}}{t^{\frac{3}{2}}}dt$$

Thus,

$$\begin{aligned}
-\frac{1}{\sigma}\sum_{i,j=0}^n c_ic_j\|\mathbf{x}_i-\mathbf{x}_j\| &= \frac{1}{2\Gamma\left(\frac{1}{2}\right)}\int_0^{+\infty} -\frac{1}{\sigma}\sum_{i,j=0}^n c_ic_j\frac{1-e^{-t\|\mathbf{x}_i-\mathbf{x}_j\|^2}}{t^{\frac{3}{2}}}dt \\
&= \frac{1}{2\Gamma\left(\frac{1}{2}\right)}\int_0^{+\infty}\frac{1}{\sigma}\frac{\sum_{i,j=0}^n c_ic_je^{-t\|\mathbf{x}_i-\mathbf{x}_j\|^2}}{t^{\frac{3}{2}}}dt.
\end{aligned}$$

Since a Gaussian kernel is PDS, the inequality $\sum_{i,j=0}^{n} c_i c_j e^{-t\|\mathrm{x}_i-\mathrm{x}_j\|^2} \geq 0$ holds and the right-hand side is non-negative. Thus, the inequality $-\frac{1}{\sigma}\sum_{i,j=0}^{n} c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\| \geq 0$ holds, which shows that ϕ is PDS. $\square$

Question : Kernel. Show that K defined by $K(x, x') = \frac{1}{\sqrt{1-(\mathbf{x}\cdot\mathbf{x}')}}$ for all $\mathbf{x}, \mathbf{x}' \in X = \left\{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_2 < 1\right\}$ is a PDS kernel. Bonus point: show that the dimension of the feature space associated to K is infinite (hint: one method to show that consists of finding an explicit expression of a feature mapping Φ).

Solution. $f : x \mapsto \frac{1}{\sqrt{1-x}}$ admits the Taylor series expansion

$$f(x) = \sum_{n=0}^{\infty} \binom{1/2}{n} (-1)^n x^n$$

for $|x| < 1$, where $\binom{1/2}{n} = \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\cdots\left(\frac{1}{2}-n+1\right)}{n!}$. Observe that $\binom{1/2}{n}(-1)^n > 0$ for all $n \geq 0$, thus, the coefficients in the power series expansion are all positive. Since the radius of the convergence of the series is one and that by the Cauchy-Schwarz inequality $|\mathbf{x}' \cdot \mathbf{x}| \leq \|\mathbf{x}'\| \|\mathbf{x}\| < 1$ for $\mathbf{x}, \mathbf{x}' \in X$, by the closure property theorem, $(\mathbf{x}, \mathbf{x}') \mapsto f(\mathbf{x}' \cdot \mathbf{x})$ is a PDS kernel. Now, let $a_n = \binom{1/2}{n}(-1)^n$. Then, for $\mathbf{x}, \mathbf{x}' \in X$,

$$\begin{aligned}
f(\mathbf{x}' \cdot \mathbf{x}) &= \sum_{n=0}^{\infty} a_n \left(\sum_{i=1}^{N} x_i x_i'\right)^n \\
&= \sum_{n=0}^{\infty} a_n \sum_{s_1+\cdots+s_N=n} \binom{n}{s_1, \ldots, s_N} \left(x_{i_1} x_{i_1}'\right)^{s_1} \cdots \left(x_{i_N} x_{i_N}'\right)^{s_N} \\
&= \sum_{s_1,\ldots,s_N \geq 0} a_{s_1+\cdots+s_N} \binom{s_1+\cdots+s_N}{s_1, \ldots, s_N} \left(x_{i_1} x_{i_1}'\right)^{s_1} \cdots \left(x_{i_N} x_{i_N}'\right)^{s_N}
\end{aligned}$$

where the sums can be permuted since the series is absolutely summable. Thus, we can write $f(\mathbf{x}' \cdot \mathbf{x}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$ with

$$\Phi(\mathbf{x}) = \left(\sqrt{a_{s_1+\cdots+s_N} \binom{s_1+\cdots+s_N}{s_1, \ldots, s_N}} x_{i_1}^{s_1} \cdots x_{i_N}^{s_N}\right)_{s_1,\ldots,s_N \geq 0}$$

Φ is a mapping to an infinite-dimensional space. $\square$