Chair of Theoretical Computer Science
School of Computation, Information, and Technology
Technical University of Munich

1								
2								
3								
4								
5								
6								
7								
8								
9								
0	☒							

Registration number

Signature

Note:
- Cross your Registration number(with leading zero). It will be evaluated automatically.
- Sign in the corresponding signature field.

Statistical Foundations of Learning

Exam: CIT4230004 / Endterm **Date:** Friday 11th August, 2023
Examiner: Prof. Debarghya Ghoshdastidar **Time:** 11:00 – 13:00

	P 1	P 2	P 3	P 4	P 5	P 6
I						

Working instructions

- This exam consists of **12 pages** with a total of **6 problems**.
  Please make sure now that you received a complete copy of the exam, and all pages are correctly printed.
- You need to answer all problems.
- The total amount of achievable credits in this exam is 42 credits.
- Sub-problems. marked * can be solved without solving the previous parts
- **Answers are only accepted if the solution approach is documented.**
  - Give a reason for each answer in the solution box of the respective subproblem.
  - If you use additional space for answer (given at end of paper), mention this in the solution box.
- You are allowed to use the lecture slides, assignments solutions or reference texts (either in printed form or on an electronic device).
- For iPads/laptops, you are only allowed to browse using a mouse/trackpad/touchscreen, but should not use the internet, any mode of typing (physical or virtual keyboard), or any means of communication.
- Do not write with red or green colours, nor use pencils.

Left room from ________ to ________	/	Early submission at ________

Problem 1 VC Dimension (6 credits)

Let $v_1, \dots, v_n \in \mathbb{R}^d$ for some $n < d$. Define the hypothesis class

$$\mathcal{H} = \left\{ x \mapsto sign\left(\sum_{i=1}^{n} \alpha_i \langle v_i, x \rangle + b \right) \mid \alpha_1, \dots, \alpha_n, b \in \mathbb{R} \right\}$$

0
1
2
3

a) Show that $\text{VCdim}(\mathcal{H}) \leq n + 1$

0
1
2
3

b) State a necessary and sufficient condition on $v_1, \dots, v_n$ such that $\text{VCdim}(\mathcal{H}) = n + 1$. You need to state a proof, but can use any result from lecture (clearly state the results you use).

Problem 2 **Explainable Clustering (9 credits)**

Consider a dataset $\mathcal{X} \subset \mathbb{R}^d$. Given two centers $a, b \in \mathbb{R}^d$, define the subsets

$$A = \{x \in \mathcal{X} : \|x - a\|_2 \leq \|x - b\|_2\} \quad \text{and} \quad B = \{x \in \mathcal{X} : \|x - b\|_2 \leq \|x - a\|_2\}$$

and define the 2-centers cost of clustering into A, B accordingly as

$$cost(A, B) = \max_{x \in \mathcal{X}} \min \{\|x - a\|_2, \|x - b\|_2\}.$$

We now construct a decision tree to approximate the clustering into A, B by partitioning $\mathcal{X}$ into two leaves

$$C_1 = \{x : x_i > \theta\} \quad \text{and} \quad C_2 = \{x : x_i \leq \theta\}$$

where we threshold at $i = \text{argmax}_{j \in [d]} |a_j - b_j|$ and $\theta = \frac{a_i + b_i}{2}$.

a) Give an example in 2-dimension ($d = 2$), where the above algorithm does not return the decision tree with minimum cost

0 1 2

b)* Prove that if a point $x \in A$ is split from a in the decision tree, then $\|x - b\|_2 \leq (1 + 2\sqrt{d}) \cdot \|x - a\|_2$

0 1 2 3 4

c) Using the argument in part-(b), show that $cost(C_1, C_2) \leq (1 + 2\sqrt{d}) \cdot cost(A, B)$

0 1 2 3

Problem 3 **Stability of bagged Tikhonov learners (5 credits)**

Given a training sample S, consider the Tikhonov regularised loss minimisation

$$\widehat{w} = \underset{w \in \mathcal{H}}{\arg\min}\ L_S(w) + \lambda \|w\|^2$$

Let us call $\widehat{w}$ a Tikhonov learner. In this problem, we study the on-average-replace-one stability of an ensemble of Tikhonov learners.
Suppose that the training sample S, of size m, is equally split into k sub-samples $C_1, \dots, C_k$ (assume m is a multiple of k). Using each sub-sample C_j (of size $\frac{m}{k}$), we obtain a Tikhonov learner $\widehat{w}_{j,S} = \underset{w \in \mathcal{H}}{\arg\min}\ L_{C_j}(w) + \lambda \|w\|^2$.

Define $\widehat{w}_{bag,S} = \frac{1}{k} \sum_{j=1}^{k} \widehat{w}_{j,S}$, and assume that the loss is convex and ρ-Lipschitz.

0 1 2 3

a) Prove that the on-average-replace-one stability of the bagged learner $\widehat{w}_{bag,S}$ is smaller than

$$\frac{\rho}{k} \cdot \mathbb{E}_{S \sim \mathcal{D}^m, (x', y') \sim \mathcal{D}, i \sim \text{Uniform}\{1,\dots,m\}} \left[\| w_{j,S^i} - w_{j,S} \| \right] \qquad \text{for a particular } j \in \{1, \dots, k\}.$$

Note that S^i above refers to the standard notation of one-replaced training sample used in lecture.

0 1 2

b) Uaing part-(a), show that the bagged learner $\widehat{w}_{bag,S}$ is on-average-replace-one stable with rate $\frac{2\rho^2}{\lambda m}$.

Problem 4 **Bayes Risk (8 credits)**

Suppose that the feature space $\mathcal{X}$ can be written as $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, that is each feature vector $x \in \mathcal{X}$ can be written as $x = (u, v)$, where $u \in \mathcal{U}$ and $v \in \mathcal{V}$ are smaller feature vectors. In this problem, we will study the risk of binary classifiers that can only see part of the data, that is, h is a function of only u instead of $x = (u, v)$.

a) Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{0, 1\}$ that is characterised as $\mathcal{D}_\mathcal{X} \times \eta$, where $\eta(x) = \mathbb{P}(y = 1|x)$. Define $\mathcal{H} \subset \{0, 1\}^\mathcal{X}$ as the set of all binary predictors that only consider information in u, that is, if $x = (u, v)$ and $x' = (u, v')$, then any $h \in \mathcal{H}$ satisfies $h(x) = h(x')$.
Let $L_\mathcal{D}(h)$ denote the risk with respect to the 0-1 loss.

0 1 2 3 4 5

- Compute the minimum risk $\min_{h \in \mathcal{H}} L_\mathcal{D}(h)$ achieved by any classifier in $\mathcal{H}$.
- What is the optimal classifier that achieves the above risk?

Hint: It may help to write $\eta(x)$ more explicitly as $\eta(u, v)$ for any $x = (u, v)$, and write expectations over $x = (u, v) \sim \mathcal{D}_\mathcal{X}$ in terms of $u \sim \mathcal{D}_\mathcal{U}$, $v \sim \mathcal{D}_{\mathcal{V}|u}$ (first u is sampled, and then v sampled given u).

b) Consider the problem where $\mathcal{X} = \{0, 1\}^3$, $\mathcal{D}_\mathcal{X}$ is uniform over $\mathcal{X}$ and, for every $x = (a, b, c)$, $\eta(x) = \dfrac{a + 2b + 3c}{6}$.
Use part (a) to derive the optimal axis-aligned classifier for this problem.
Note: There must be an argument about why the presented axis-aligned classifier is optimal.

0 1 2 3

Problem 5 **Robust risk of 1-nearest neighbour classifier (9 credits)**

Assume $\mathcal{X} = \mathbb{R}^p$. For any binary classifier, $h : \mathcal{X} \to \{0, 1\}$, we define δ-robust risk in the following way:

- For any test pair $(x, y) \in \mathcal{X} \times \{0, 1\}$, let $\widetilde{x}$ be sampled from a normal distribution centred at x and covariance $\delta^2 I$, that is, $\widetilde{x} \sim \mathcal{N}(x, \delta^2 I)$.
- The δ-robust 0-1 loss is computed at (x, y) as $\mathbb{E}_{\widetilde{x} \sim \mathcal{N}(x, \delta^2 I)} \left[\mathbf{1} \{ h(\widetilde{x}) \neq y \} \right]$.
- The δ-robust risk is defined as $L_{\mathcal{D}}^{rob}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{E}_{\widetilde{x} \sim \mathcal{N}(x, \delta^2 I)} \left[\mathbf{1} \{ h(\widetilde{x}) \neq y \} \right]$

0 1 2 3 4 5

a) For the 1-nearest neighbour classifier $h = h^{NN}$, prove that the asymptotic δ-robust risk $\lim_{m \to \infty} \mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}^{rob}(h)]$ can be computed as

$$\lim_{m \to \infty} \mathbb{E}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}^{rob}(h^{NN}) \right] = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \mathbb{E}_{\widetilde{x} \sim \mathcal{N}(x, \delta^2 I)} \left[\eta(x) + \eta(\widetilde{x}) - 2\eta(x)\eta(\widetilde{x}) \right] .$$

Note: You may assume η is uniformly continuous.

0 1 2 3 4

b)* Assume $\mathcal{X} = \mathbb{R}^p$ and η is ρ-Lipschitz, that is, $|\eta(x') - \eta(x)| \leq \rho \|x' - x\|$ for all x, x'. Show that the asymptotic δ-robust risk is smaller than $L_{\mathcal{D}}^{NN} + \rho\delta\sqrt{p}$, where $L_{\mathcal{D}}^{NN}$ is the asymptotic risk of the 1-nearest neighbour classifier.
Hint: If z is a p-dimensional standard normal vector, then $\mathbb{E}[\|z\|^2] = p$.

Problem 6 Universal Kernels (5 credits)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a postive semidefinite kernel, and let $\phi : [0, 1] \to \mathcal{H}$ be the feature map into its RKHS $\mathcal{H}$. In this problem, we show that if the kernel k is universal, then ϕ is injective (that is, for every $x \neq x'$, ϕ satisfies $\phi_x \neq \phi_{x'}$).

a) Assume kernel k is universal and consider any compact set $C \subset \mathcal{X}$. Show that, for every continuous function $f : \mathcal{X} \to \mathbb{R}$ and $\epsilon > 0$, there exists a function $h \in \text{span}\{\phi_x : x \in \mathcal{X}\}$ such that

0
1
2

$$|f(x) - f(x')| \leq |h(x) - h(x')| + \epsilon \qquad \text{for every } x, x' \in C.$$

b)* Using the statement of part-(a), prove by contradiction that ϕ must be injective.

0
1
2
3

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.