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	X							

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: IN2378 / Endterm **Date:** Thursday 18th August, 2022
Examiner: Prof. Debarghya Ghoshdastidar **Time:** 13:45 – 15:15

	P 1	P 2	P 3	P 4	P 5
I					

Working instructions

- This exam consists of **10 pages** with a total of **5 problems**.
  Please make sure now that you received a complete copy of the exam, and all pages are correctly printed.
- You need to answer **only 4 out of 5 problems**.
  If you attempt all questions, then the 4 problems with most points will be considered.
- The total amount of achievable credits in this exam is **40 points**.
- 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.
- Allowed resources: Printed lecture notes or on an electronic device.
- Do not write with red or green colors nor use pencils.

Left room from __________ to __________ / Early submission at __________

Problem 1 **Risk and Bayes Risk (10 credits)**

Consider a binary classification problem in 2 dimension, where the joint distribution of the features $x = (x_1, x_2)$ and label y is such that

$$x = (x_1, x_2) \sim \mathit{Uniform}[0,1] \times \mathit{Uniform}[0,1] \qquad \eta(x) = \mathbb{P}(y = 1|x) = \begin{cases} 0.1 & \text{if } x_1 < 0.5 \& x_2 < 0.5 \\ 0.9 & \text{if } x_1 \geq 0.5 \& x_2 \geq 0.5 \\ 0.6 & \text{otherwise.} \end{cases}$$

0 1 2

a) State the Bayes classifier and compute the Bayes risk for the above problem.

0 1 2 3

b) What is the optimal axis-aligned linear classifier for the above problem? Also state minimal risk achieved by axis-aligned linear classifiers.

Note: There must be an argument about why the presented axis-aligned classifier is optimal.

0 1 2 3 4 5

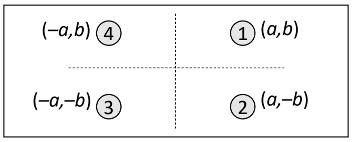
c) Can we achieve a risk lower than axis-aligned linear classifiers if we use axis aligned rectangles, that is,

$$h(x) = \begin{cases} 1 & \text{for } a \leq x_1 \leq b, c \leq x_2 \leq d \\ 0 & \text{otherwise.} \end{cases}$$

where $a, b, c, d \in [0,1]$ are parameters of the model?

Problem 2 **Clustering and Hierarchical Clustering (10 credits)**

Consider the following configuration of four points in $\mathbb{R}^2$. The coordinates of the points are noted—coordinates of point 1 is (a, b) where $a > b > 0$.



a) Assume $a > b > 0$. Compute the 2-means cost for all possible 2-way clustering. Based on your computation, state the optimal 2-means clustering.

b) Assume $a > b$ and consider k-means++ with $k = 2$.

- What is the expected cost of k-means++?
- What is the probability that k-means++ returns the optimal 2-means clustering?

Note: Probability and expectation are with respect to the randomisation in k-means++.

c)* Assume $a > b$ and consider the distance $d(x, x') = \|x - x'\|^2$.

- Draw the hierarchy (tree) returned by average linkage clustering.
- Compute the value function of the tree for $d(x, x') = \|x - x'\|^2$.

Problem 3 **Algorithmic Stability (10 credits)**

Given a loss function ℓ, we say that a learner $\mathcal{A}$ is asymptotically on-average-replace-one stable with respect loss function ℓ and distribution $\mathcal{D}$ if

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m,(x',y')\sim\mathcal{D},i\sim\text{Uniform}(m)} \left[\ell\big(\mathcal{A}_{S^i}(x_i), y_i\big) - \ell\big(\mathcal{A}_S(x_i), y_i\big)\right] = 0.$$

Notations are same as the ones used in lecture slides for on-average-replace-one stability: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ is the training sample and S^i is obtained by replacing (x_i, y_i) in S with (x', y'). $i \sim \text{Uniform}(m)$ denotes that i is chosen uniformly at random from $\{1, 2, \dots, m\}$.

We call a learner $\mathcal{A}$ universally asymptotically on-average-replace-one stable with respect loss function ℓ if it is asymptotically on-average-replace-one stable for every distribution $\mathcal{D}$.

0 1 2 3 4 5 6

a) Let $\mathcal{X} = \{x \ : \ \|x\|_2 \le B\}$. Consider linear ridge regression learner $\mathcal{A}$ that outputs a linear function $\mathcal{A}_S(x) = \widehat{w}^\top x$, where

$$\widehat{w} = \underset{w\,:\,\|w\|_2\le B}{\arg\min}\ \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\|w\|_2^2.$$

Prove or disprove: The linear ridge regressor $\mathcal{A}$ is universally asymptotically on-average-replace-one stable with respect to squared loss $(y - w^\top x)^2$.

If you disprove, then for which distributions $\mathcal{D}$ is the learner asymptotically on-average-replace-one stable?

0 1 2 3 4

b)* Prove or disprove: 1-nearest neighbour classifier is universally asymptotically on-average-replace-one stable with respect to 0-1 loss.

If you disprove, then for which distributions $\mathcal{D}$ is the learner asymptotically on-average-replace-one stable?

Problem 4 VC Dimension (10 credits)

Consider the set $\mathcal{X}_n = \{1, 2, 3, \dots, n\}$. For any $k \in \mathcal{X}_n$, define the binary classifier $h_k : \mathcal{X}_n \to \{0, 1\}$ as

$h_k(x) = 1$ if x is a multiple of k, and 0 otherwise.

Let $\mathcal{H}_n = \{h_k \ : \ k \in \mathcal{X}_n\}$ be the hypothesis class of all binary classifiers of above form.

a) For $n = 7$, compute VCdim(H_7).
Hint: You can get a tight upper bound based on $|\mathcal{H}_7|$.

0
1
2
3
4

b) What is the maximum value of n such that VCdim($\mathcal{H}_n$) = VCdim($\mathcal{H}_7$).
Hint: Take d = VCdim($\mathcal{H}_7$) + 1. For a set $\{x_1, \dots, x_d\}$ to be shattered by $\mathcal{H}_n$, identify conditions that $x_1, \dots, x_d$ should satisfy. From this, you can get the smallest possible value $\max\{x_1, \dots, x_d\}$ should have.

0
1
2
3
4
5
6

Problem 5 Uniform Convergence (10 credits)

In transfer learning, the goal is to minimise the risk with respect to a target distribution $\mathcal{D}_1$, that is, $\min_{h\in\mathcal{H}} L_{\mathcal{D}_1}(h)$. However, we have access to few training samples from $\mathcal{D}_1$ and many training samples from a source distribution $\mathcal{D}_2$. Formally let $\beta \in (0,1)$ and assume that the training set S, of size m, is split into βm samples from $\mathcal{D}_1$ and rest from $\mathcal{D}_2$, that is, $S = S_1 \cup S_2$, where $S_1 \sim \mathcal{D}_1^{\beta m}$, $S_2 \sim \mathcal{D}_2^{(1-\beta)m}$.

We aim to minimise a weighted empirical risk. For $\alpha \in (0,1)$, define the weighted empirical risk of classifier h as

$$L_{S,\alpha}(h) \;=\; \alpha L_{S_1}(h) + (1-\alpha) L_{S_2}(h) \;=\; \frac{\alpha}{\beta m}\sum_{(x,y)\in S_1} \mathbf{1}\{h(x) \neq y\} + \frac{1-\alpha}{(1-\beta)m}\sum_{(x,y)\in S_2} \mathbf{1}\{h(x) \neq y\}$$

Let $\widehat{h}$ minimise $L_{S,\alpha}(h)$. The following sub-problem derive a bound on $L_{\mathcal{D}_1}(\widehat{h})$, generalisation bounds for $\widehat{h}$.

Note: To solve the sub-problems, assume the following:

- $\mathcal{H}$ has a finite number of hypotheses;
- There is a target predictor $h^* \in \mathcal{H}$ such that $L_{D_1}(h^*) = 0$ (equivalently, $\mathcal{D}_1$ is realisable).

0 1 2

a) Define a $\mathcal{H}$-distance between two distributions $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_{\mathcal{D}'}(h)|$. Show that for any h,

$$L_{D_1}(h) \;\leq\; \mathbb{E}_S[L_{S,\alpha}(h)] + (1-\alpha) d_{\mathcal{H}}(\mathcal{D}_1, \mathcal{D}_2).$$

0 1 2 3 4 5

b)* Use Hoeffding's inequality and union bound to show that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}} |L_{S,\alpha}(h) - \mathbb{E}_S[L_{S,\alpha}(h)]| \leq \sqrt{\frac{1}{2m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}\right)\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

c) Use the bounds from previous parts, and optimality of $\widehat{h}$ to conclude that, with probability $1-\delta$,

0
1
2
3

$$L_{\mathcal{D}_1}(\widehat{h}) \;\leq\; (1-\alpha)\big(L_{\mathcal{D}_2}(h^*) + d_{\mathcal{H}}(\mathcal{D}_1,\mathcal{D}_2)\big) + \sqrt{\frac{2}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}\right)\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

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