Assignment 1

Due: 13.05.2025, 23:59

Points: 17

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 1.1: Bayes Risk

Assume $\alpha \leq \frac{1}{3}$. Let $\mathcal{X} = \{1, 2, \ldots, 30\}$, $\mathcal{Y} = \{\pm 1\}$ and class probability be

$$\eta(x) = \mathbb{P}(y = +1|x) = \begin{cases} 1-\alpha & \text{if } x \in \{11, 12, \ldots, 20\} \\ \alpha & \text{otherwise.} \end{cases}$$

You may assume that x is sampled from a probability mass function $\mathbf{p} = (p_1, p_2, \ldots, p_{30})$.

a) Compute the Bayes risk for the problem, and show that the Bayes classifier is in the class of signed intervals

$$\mathcal{H}_{int} = \left\{ h_{s,t,b}(x) = b \text{ for } x \in (s,t), \text{ and } -b \text{ otherwise } : \ b \in \{\pm 1\}, s, t \in \mathbb{R}, s < t \right\}$$

b) Define $q_1 = (p_1 + \ldots + p_{10})$, $q_2 = (p_{11} + \ldots + p_{20})$, $q_3 = (p_{21} + \ldots + p_{30})$. Find the minimal risk achieved by the class of decision stumps

$$\mathcal{H}_{ds} = \left\{ h_{t,b}(x) = b \text{ for } x < t, \text{ and } -b \text{ otherwise } : \ b \in \{\pm 1\}, t \in \mathbb{R} \right\}.$$

The solution will depend on q_1, q_2, q_3. Give possible optimal decision stumps.

(2 + 3 = 5 points)

Exercise 1.2: OLS is not universally consistent.

In the lecture, we mentioned that OLS is not universally consistent. Consider the following non-linear model, where $x \sim \mathcal{N}(0, I)$, $y = (x^\top w^*)^2 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\|w^*\| > 0$.

a) Consider the square loss, find the Bayes predictor for this model and compute the Bayes risk.

b) Prove that OLS is not consistent w.r.t. this distribution, i.e., show that the expected risk of the OLS predictor does **not** converge to the Bayes risk as the number of training samples $m \to \infty$.

Hints:

- For any v and $z \sim \mathcal{N}(0, \Sigma)$, the fourth moment satisfies $\mathbb{E}[(z^\top v)^4] = 3(v^\top \Sigma v)^2$.
- All odd-order moments of zero-mean Gaussians vanish, i.e., $\mathbb{E}[z^{\otimes k}] = 0$ for odd k.

(2+3=5 points)

Exercise 1.3: Universal consistency of ϵ-neighbourhood classifiers

Consider a domain $\mathcal{X} \subseteq \mathbb{R}$. Given training samples $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{X} \times \{\pm 1\}$ and some $\epsilon > 0$, we define the ϵ-neighbourhood classifier $h_{S,\epsilon} : \mathcal{X} \to \{\pm 1\}$ as

$$h_{S,\epsilon}(x) = \operatorname{sign}\left(\sum_{i\,:\,|x_i - x| \leq \epsilon} y_i \right)$$

a) Fix $\epsilon > 0$ and consider an arbitrary training sample S. Assume that for any $x \in \mathcal{X}$, there is at least one sample in an ϵ-neighborhood of it. Express $h_{S,\epsilon}$ as a plug-in classifier with a weighted average estimator $\widehat{\eta}$.

b) In the next two subproblems, we prove universal consistency in a specific setting: Let $\mathcal{X} = \{0, 1\}$ and $\epsilon < 1$, and assume that $P(x) > 0$ for both $x \in \mathcal{X}$. In this setting, simplify the weighted average estimator $\widehat{\eta}$ from part a) and show that $\forall x \in \{0, 1\}$, the estimator $\widehat{\eta}(x)$ converges to $\eta(x)$ in probability as $m \to \infty$.

Hint: If $Z \sim \text{Binomial}(n, p)$ then $Z/n \to p$ in probability as $n \to \infty$.

c) Use the previous subproblem to show that the ϵ-neighbourhood classifier is universally consistent on $\mathcal{X} = \{0, 1\}$ for any $\epsilon < 1$ without using Stone's theorem.

(2+2+3 = 7 points)

Assignment 2

Due: 27.05.2025, 23:59

Points: 15

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 1.1: Bayes Risk

Let $\mathcal{X} = \mathcal{Y} = \{1, 2, 3\}$. Assume the labels Y follow the distribution

$$P(Y = j) = \begin{cases} 1/4 & \text{if } j = 1, 2 \\ 1/2 & \text{if } j = 3 \end{cases}$$

Conditioned on the labels, the features X are distributed as

$$P(X = i | Y = 1) = \begin{cases} 1/3 & \text{if } i = 2 \\ 2/3 & \text{if } i = 3 \end{cases}$$
$$P(X = i | Y = 2) = \begin{cases} 1/2 & \text{if } i = 1 \\ 1/2 & \text{if } i = 3 \end{cases}$$
$$P(X = i | Y = 3) = \begin{cases} 2/3 & \text{if } i = 1 \\ 1/3 & \text{if } i = 2 \end{cases}$$

(a) Compute the Bayes classifier (i.e. the classifier h^* that maximizes the probability that $h^*(x) = y$ for any given x).

(b) Compute the Bayes risk.

Hint: You don't have to compute the marginals of X.

(3+2 = 5 points)

Exercise 1.2: Rademacher Complexity of Sets

The empirical Rademacher complexity of a set $X \subset \mathbb{R}^m$ is defined as

$$\mathcal{R}_m(X) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{x \in X}\langle \sigma, x\rangle\right]$$

where the expectation is with respect to m independent Rademacher random variables $\sigma = (\sigma_1, \ldots, \sigma_m) \in \{\pm 1\}^m$. The convex hull of a set X is defined as

$$conv(X) = \left\{ \sum_{i=1}^{N} \lambda_i x_i \middle| x_i \in X, \lambda_i \geq 0, \sum_{i=1}^{N} \lambda_i = 1, N > 0 \right\}$$

Show that $\mathcal{R}_m(X) = \mathcal{R}_m(conv(X))$.

(4 points)

Exercise 1.3: Uniform Convergence in Transfer Learning

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

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

You may 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).

Let $\hat{h}$ minimise $L_{S,\alpha}(h)$. This exercise derives a bound on $L_{\mathcal{D}_1}(\hat{h})$, i.e. generalisation bounds for $\hat{h}$, in three steps.

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

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

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

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

3. Use the bounds from previous parts, and optimality of $\hat{h}$ to conclude that, with probability $1-\delta$,

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

(1 + 3 + 2 = 6 points)

Assignment 3

Due: 10.06.2025, 23:59

Points: 15

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 2.1: VC Dimension I

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

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

1. Show that $\text{VCdim}(\mathcal{H}) \leq n + 1$
2. Prove a necessary and sufficient condition on $v_1, \ldots, v_n$ such that $\text{VCdim}(\mathcal{H}) = n+1$.

Hint: You can answer this question based on results from the lecture, and some linear algebra.

(3 + 2 = 5 points)

Exercise 2.2: VC Dimension II

Consider the set $\mathcal{X}_n = \{1, 2, 3, \ldots, n\}$. For any $k \in \mathcal{X}_n$, define the binary classifier

$$h_k : \mathcal{X}_n \to \{0, 1\},\ h_k(x) = \begin{cases} 1 & \text{if } x \text{ is a multiple of } k \\ 0 & \text{otherwise} \end{cases}$$

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

1. For $n = 7$, compute $\text{VCdim}(H_7)$. **Hint:** There's a tight upper bound based on $|\mathcal{H}_7|$.
2. What is the maximum value of n such that $\text{VCdim}(\mathcal{H}_n) = 2$? Justify your answer.

(2 + 2 = 4 points)

Exercise 2.3: VC dimensions

In this exercise, we will see that the number of parameters of a hypothesis class need not be equal to the VC dimension.

1. For any $k \geq 1$, derive the VC dimension of

$$\mathcal{H} = \left\{ \sum_{i=0}^{k} \mathbf{1}\left\{ t_{2i} \leq x < t_{2i+1} \right\}, \quad 0 \leq t_0 < \cdots < t_{2k+1} \leq 1 \right\}$$

2. $\mathcal{X} = \mathbb{R}$. Consider the hypothesis class $\mathcal{H} = \{\text{sign}(\sin(ax)), \quad a \in \mathbb{R}\}$. Derive $\text{VCdim}(\mathcal{H})$.

   Hint: For your proof it is helpful to consider the set of points $x_i = 10^{-i}$. Then, for any labels $y_1, \ldots, y_n$ choose

$$a = \pi \left(1 + \sum_{i=1}^{n} \frac{(1 - y_i)10^i}{2} \right)$$

(2+4 = 6 points)

Assignment 4

Due: 25.06.2025, 23:59

Points: 16

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 4.1: Some applications of Paley-Zygmund

Recall the Paley-Zygmund inequality that we proved on Sheet 1: Let Z be a non-negative random variable with finite variance. Then, for any scalar $\theta \in [0,1]$, it holds that

$$\mathbb{P}\big(Z > \theta\mathbb{E}[Z]\big) \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$$

(a) Prove the following: If Z is a non-negative random variable with finite variance, and $\mathbb{E}[Z]^2 > c\mathbb{E}[Z^2]$ for some constant $0 < c \leq 1$, then $\mathbb{E}[\sqrt{Z}]$ can be sandwiched as follows

$$\sqrt{\mathbb{E}\left[Z\right]} \geq \mathbb{E}\left[\sqrt{Z}\right] \geq c'\sqrt{\mathbb{E}[Z]}$$

where $c' \in (0,1]$ is a constant.

Hint: It may be convenient to use Paley-Zygmund inequality and Markov inequality.

(b) Apply the above inequality to derive a lower bound on $\mathbb{E}[Y]$, where Y is defined as follows:

(a) $Y = \sqrt{S}$, where $S \sim \text{Binomial}(n,p)$ with $np \geq 1$

(b) $Y = \left|\sum_{i=1}^{n} \sigma_i\right|$, where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher variables

(3+3 = 6 points)

Exercise 4.2: Lower bound on Rademacher complexity

Let $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$ be a hypothesis class with $\text{VCdim}(\mathcal{H}) = d < \infty$. Let $\mathcal{D}_{\mathcal{X}}$ denote the uniform distribution on some set of d points in $\mathcal{X}$ shattered by $\mathcal{H}$. Prove that there exists $c > 0$ such that the expected Rademacher complexity satisfies

$$R_{\mathcal{D}_{\mathcal{X}},m}(\mathcal{H}) \geq c\sqrt{\frac{d}{m}} \text{ for all } m \geq 1$$

(5 points)

Exercise 4.3: Lower bound on uniform convergence

Suppose ℓ is a bounded loss function, that is, there is $c > 0$ such that $0 \leq \ell(y,y') \leq c$ for all $y, y' \in \mathcal{Y}$.

For a hypothesis class $\mathcal{H} \in \mathcal{Y}^{\mathcal{X}}$ and distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, let $R_{\mathcal{D},m}(\ell \circ \mathcal{H})$ denote the (expected) Rademacher complexity of $\mathcal{H}$ with respect to loss ℓ.

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sup_{h\in\mathcal{H}}\big|L_{\mathcal{D}}(h) - L_S(h)\big|\right] \geq \frac{1}{2}R_{\mathcal{D},m}(\ell \circ \mathcal{H}) - \frac{c}{2\sqrt{m}}$$

(5 points)

Assignment 5

Due: 09.07.2025, 23:59

Points: 14

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 5.1: Rademacher Complexity of Neural Networks

Let $\mathcal{X} \subset \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be a set of size n. Consider the following neural network $h : \mathcal{X} \to \mathbb{R}$ with one hidden layer and ReLU activation, given by

$$h(x) = v^T \phi(Wx)$$

where $\phi(z) = \max(z, 0)$ and $W \in \mathbb{R}^{m \times d}$ has rows $w_1, \ldots, w_m$. Hence, the hidden layer has m neurons. Give a bound on the empirical Rademacher complexity of this class of neural networks (denoted $\mathcal{H}$). Assuming that $\|w_j\|_2 \leq C_1$ for all $j \in [m]$ and $\|v\|_2 \leq C_2$, show that

$$\mathcal{R}_{\mathcal{X}}(\mathcal{H}) \leq \frac{2C_1 C_2 \sqrt{m}}{\sqrt{n}}$$

Hints: You will need to show that

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right] \leq 2\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(x_i) \right]$$

Also, remember the contraction principle for Lipschitz functions.

(5 points)

Exercise 5.2: Stability and Generalization

We call a learner $\mathcal{A}$ *"strong replace-one stable"* with rate β_m if for all $i \in [m]$

$$\mathbb{E}_{S \sim \mathcal{D}^m, (x',y') \sim \mathcal{D}} \left[|\ell(\mathcal{A}_{S^i}(x_i), y_i) - \ell(\mathcal{A}_S(x_i), y_i)| \right] \leq \beta_m$$

where $S^i = S \cup \{(x', y')\} \setminus \{(x_i, y_i)\}$, as defined in lecture. Assume $\exists c > 0$ such that

$$\forall\, y, y' : \quad 0 \leq \ell(y, y') \leq c$$

Show that for a *"strong replace-one stable"* learner $\mathcal{A}$,

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[\left(L_{\mathcal{D}}(\mathcal{A}_S) - L_S(\mathcal{A}_S) \right)^2 \right] \leq \frac{c^2}{m} + 6c\beta_m$$

(5 points)

Exercise 5.3: Hard SVM vs. Soft SVM

Prove or disprove the following statement: There exists a universal $\lambda > 0$ such that for every set of separable training data the solution of Soft SVM with parameter λ is identical to the solution of Hard SVM.

(4 points)

Assignment 6

Due: 23.07.2025, 23:59

Points: 15

The solutions have to be handed in via Moodle. We do not accept late submissions.

We would recommend using LaTeX for writing your submission but also accept handwritten solutions, but please note that if we can not read or understand it, we cannot grade it.

To get full points, always provide the steps in your derivation/proofs and make clear when/how you use known results, for example, from the lecture (e.g. already proven concentration inequalities).

Exercise 6.1: Universality of the Gaussian kernel

Let $\mathcal{X} \subset \{x \in \mathbb{R}^p\}$ be a bounded set. In the lecture, we saw that the exponential kernel $k(x, y) = \exp(\langle x, y\rangle)$ is universal on $\mathcal{X}$. Use this to conclude that the Gaussian kernel $k(x, y) = \exp(-\frac{1}{2}\|x - y\|^2)$ is also universal on $\mathcal{X}$.

(5 points)

Exercise 6.2: Feature maps of universal kernels are injective

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a universal kernel. Denote by $\mathcal{H}$ its RKHS, and by $\phi : \mathbb{R}^d \to \mathcal{H}$ its feature map. Recall the reproducing property: For every $f \in \mathcal{H}$ and any $x \in \mathbb{R}^d$, we have $f(x) = \langle \phi(x), f\rangle$. Prove that ϕ is injective.

(5 points)

Exercise 6.3: Kernel Mean Classifier

Consider data $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, where $y_i = -1$ for $i \leq m$ and $y_i = 1$ for all $i \geq m+1$. Define the group means

$$\alpha = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \beta = \frac{1}{n}\sum_{i=m+1}^{n+m} x_i.$$

Consider the mean classifier

$$f(x) = \begin{cases} -1, & \text{if } \|x - \alpha\| \leq \|x - \beta\| \\ +1, & \text{else} \end{cases} \tag{1}$$

1. Show that this classifier produces **linear** decision boundaries, i.e. f is of the form $f(x) = sgn(\langle x, v\rangle + b)$, for some $v \in \mathbb{R}^d$ and some $b \in \mathbb{R}$.
2. In order to allow for nonlinear decision boundaries, we kernelize f. To this end, let k be a positive semi-definite kernel on $\mathbb{R}^d$, and denote ϕ for its feature map. Define

$$\alpha = \frac{1}{m}\sum_{i=1}^{m} \phi(x_i), \quad \beta = \frac{1}{n}\sum_{i=m+1}^{n+m} \phi(x_i).$$

   Kernelize the classifier f from (1) by replacing the Euclidean distances by distances in the feature space. Show that the output of this kernelized classifier can be evaluated on a new point x by evaluating the kernel $k(x, x_i)$, **without explicitly computing the feature map** $\phi(x)$.

(2+3=5 points)