

**STATISTICAL MACHINE LEARNING (WS2025)**  
**EXAM T3 (90MIN / 26P)**

**Assignment 1 (6p).** Let  $\mathcal{X}$  denote a set of images of handwritten digits and let  $\mathcal{Y} = \{0, 1, \dots, 9\}$  be the corresponding set of digit labels. Consider a prediction rule  $h: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Y} \times \mathcal{Y}$  which, for each pair of images  $(x_A, x_B) \in \mathcal{X} \times \mathcal{X}$ , outputs a pair of predicted digit labels  $h(x_A, x_B) = (h_A(x_A), h_B(x_B)) \in \mathcal{Y} \times \mathcal{Y}$ . We measure the prediction accuracy of  $h$  using the absolute deviation between the sum of the true digit labels and the sum of the predicted digit labels. To this end, we define the loss function  $\ell(y_A, y_B, \hat{y}_A, \hat{y}_B) = |y_A + y_B - \hat{y}_A - \hat{y}_B|$ . The expected risk of the predictor  $h$  with respect to a joint probability density function  $p(x_A, x_B, y_A, y_B)$  defined on  $\mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$  is given by  $R(p, h) = \mathbb{E}_{(x_A, x_B, y_A, y_B) \sim p} [\ell(y_A, y_B, h_A(x_A), h_B(x_B))]$ . Since the distribution  $p(x_A, x_B, y_A, y_B)$  is unknown, we estimate  $R(p, h)$  using the test error

$$\widehat{R}(S_l, h) = \frac{1}{l} \sum_{j=1}^l \ell(y_{A,j}, y_{B,j}, h_A(x_{A,j}), h_B(x_{B,j})),$$

where  $S_l = ((x_{A,j}, x_{B,j}, y_{A,j}, y_{B,j}) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \mid j = 1, \dots, l)$  is a test set consisting of  $l$  i.i.d. samples drawn from the distribution  $p(x_A, x_B, y_A, y_B)$ .

a) What is the minimal number of test examples  $l$  that must be collected in order to guarantee that

$$R(p, h) \in (\widehat{R}(S_l, h) - \varepsilon, \widehat{R}(S_l, h) + \varepsilon)$$

with probability at least  $1 - \delta$ ? Express  $l$  as a function of  $\varepsilon$  and  $\delta$ .

b) Provided the number of test examples is  $l = 10,000$ , what is the probability that  $R(p, h)$  is in the interval  $(\widehat{R}(S_l, h) - 1, \widehat{R}(S_l, h) + 1)$ ?

**Assignment 2 (6p).** Consider the vector of binary observations  $\mathbf{x} = (x_1, \dots, x_n) \in \{0, 1\}^n$  and the hidden state  $y \in \mathcal{Y} = \{+1, -1\}$  generated from the joint probability

$$p(\mathbf{x}, y) = p(y) \prod_{i=1}^n p(x_i \mid y), \quad (1)$$

where  $p(x_i \mid y)$ ,  $i = 1, \dots, n$ ,  $y \in \mathcal{Y}$ , are conditional probabilities of the binary observations given  $y$ , and  $p(y)$  is a prior probability. Assume that the conditional and the prior probabilities are unknown, and the objective is to learn a strategy  $h \in \mathcal{X} \rightarrow \mathcal{Y}$  which minimizes the probability of misclassification. A learning algorithm  $A: \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$  is employed. It returns a strategy  $h$  from the hypothesis space  $\mathcal{H} = \{h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b) \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$  based on samples generated from (1).

a) Find the VC-dimension of  $\mathcal{H}$ .

b) If the algorithm discovers a classifier with minimal classification error on training examples, analyse whether the algorithm is a successful PAC learner.

c) Determine the approximation error of the learning algorithm  $A$ .

Provide explanations for your answers. Yes/No answers are not sufficient.

**Assignment 3** (4p). Consider a dataset  $T_m = (x_i \in (0, \infty) \mid i = 1, \dots, m)$  consisting of  $m$  statistically independent observations drawn from a log-normal distribution with parameters  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$ . The probability density function of the distribution is given by

$$p(x; \mu, \sigma^2) = \frac{1}{x \sigma \sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right). \quad (2)$$

**Task:** Using the dataset  $T_m$ , derive the maximum likelihood estimates (MLEs) of the parameters  $\mu$  and  $\sigma^2$ .

**Assignment 4** (6p). Consider a regression problem with training datasets

$$\mathcal{T}^m = \{(x_i, y_i) \mid i = 1, \dots, m\}$$

generated by

$$y = f(x) + \epsilon,$$

where the inputs  $x_i$  are drawn i.i.d. from a distribution  $p(x)$ , the noise satisfies  $\mathbb{E}(\epsilon) = 0$  and  $\text{Var}(\epsilon) = \sigma^2$ , and  $\epsilon$  is independent of  $x$ .

Consider the regression model which ignores the input  $x$  and predicts the output of a single uniformly chosen training example. In other words, we uniformly sample the index of a training sample and use the observed label  $y_J$  as the prediction:

$$h(x) = y_J, \quad J \sim \text{Unif}\{1, \dots, m\},$$

where  $J$  is independent of the dataset.

(1) Compute the expected predictor

$$g(x) = \mathbb{E}_{\mathcal{T}^m \sim p(\mathcal{T}^m), J} [h(x)].$$

(2) Give the squared bias

$$\mathbb{E}_x \left[ (g(x) - f(x))^2 \right].$$

(3) Compute the variance

$$\text{Var}_{x, \mathcal{T}^m \sim p(\mathcal{T}^m), J} (h(x)).$$

Simplify all expressions as much as possible.

**Assignment 5** (4p). Consider a Hidden Markov Model (HMM) with two hidden states  $s_t \in \{A, B\}$  and observations  $x_t \in \{0, 1\}$ . The initial distribution is uniform:  $p(s_1 = A) = p(s_1 = B) = \frac{1}{2}$ . Transitions are

$$p(s_t \mid s_{t-1}) = \begin{bmatrix} \beta & 1 - \beta \\ 1 - \beta & \beta \end{bmatrix},$$

and emissions are

$$p(x_t = 1 \mid A) = \gamma, \quad p(x_t = 0 \mid A) = 1 - \gamma, \quad p(x_t = 1 \mid B) = \delta, \quad p(x_t = 0 \mid B) = 1 - \delta.$$

**a)** Write down the general factorized expression for the joint probability  $p(x, s)$  for  $x = (x_1, \dots, x_n)$  and  $s = (s_1, \dots, s_n)$ .

**b)** Let  $n = 2$  and  $x = (1, 0)$ . Compute  $p(x = (1, 0), s = (A, A))$  and  $p(x = (1, 0), s = (A, B))$ . Simplify your answers.