STATISTICAL MACHINE LEARNING (WS2021)
EXAM T2 (90MIN / 22P)

Assignment 1 (5p). Let $\mathcal{X}$ be a set of input observations and $\mathcal{Y} = \mathcal{A}^n$ a set of sequences of length n defined over a finite alphabet $\mathcal{A}$. Let $h\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule that returns a sequence $h(x) = (h_1(x), \dots, h_n(x))$ for each $x \in \mathcal{X}$. Assume that we want to measure the prediction accuracy by a loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ defined as

$$\ell\big((y_1, \dots, y_n), (y'_1, \dots, y'_n)\big) = \min\left\{5, \sum_{i=1}^{n} [\![y_i \neq y'_i]\!]\right\},$$

that is, we penalize the prediction by the Hamming distance but we pay penalty at most 5. The performance of the prediction rule is measured by the expected risk $R(h) = \mathbb{E}_{(x,y_1,\dots,y_n)\sim p}\ell((y_1, \dots, y_n), h(x))$ where $p(x, y_1, \dots, y_n)$ is a p.d.f. defined over $\mathcal{X} \times \mathcal{Y}$. As the distribution $p(x, y_1, \dots, y_n)$ is unknown, we estimate $R(h)$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{j=1}^{l} \ell\big((y_1^j, \dots, y_n^j), h(x^j)\big) ,$$

where $\mathcal{S}^l = \{(x^i, y_1^i, \dots, y_n^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l\}$ is a set of examples drawn from i.i.d. random variables with distribution $p(x, y_1, \dots, y_n)$.

a) Assume that the sequence length is $n = 10$ and that we compute the test error from $l = 50$ examples. What is the minimal probability that $R(h)$ will be in the interval $(R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1)$?

b) What is the minimal number l of test examples which we need to collect in order to guarantee that $R(h)$ is in the interval $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability γ at least? Write l as a function of ε, n and γ.

Assignment 2 (3p). Consider a random variable $x \in \mathbb{R}$ that follows a distribution $p(x)$ with expectation $\mathbb{E}_{x\sim p(x)}[x] = \mu_0$ and variance $\mathbb{V}_{x\sim p(x)}[x] = \sigma_0^2$. We want to approximate $p(x)$ by a Gaussian distribution

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

Find the parameters μ, σ of the Gaussian distribution that provides the best approximation of $p(x)$ w.r.t. the Kullback-Leibler divergence

$$D_{KL}(p(x) \parallel q(x)) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Assignment 3 (5p). Let $s = (s_1, \ldots, s_n)$ denote sequences of length n over the finite alphabet $\mathcal{A} = \{a, b, c, \ldots, z\}$. Let $p(s)$ be a Markov chain model on them with probability given by

$$p(s) = p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-s}).$$

We want to find the most probable sequence s among all sequences which have the letter "q" in position k. Assume that $1 < k < n$. Give an algorithm for solving this task. What is its run-time complexity?

Assignment 4 (5p). Define a neural module (layer) joining a linear layer and an ELU (Exponential Linear Unit) layer. Give the forward, backward and parameter messages. Consider n inputs, K units of the linear layer and K units of the ELU layer each processing the output of the corresponding unit of the preceding linear layer. Each ELU unit applies the non-linearity:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ e^x - 1, & \text{if } x \le 0. \end{cases}$$

- The forward message is defined as a function of layer outputs w.r.t. to its inputs.
- The backward message is defined as the set of derivatives of all layer outputs w.r.t. to all layer inputs.
- Finally, the parameter message is defined as the set of derivatives of all layer outputs w.r.t. to all layer parameters.

Assignment 5 (4p). Consider a regression problem with multiple training datasets $\mathcal{T}^m = \{(x_i, y_i) \mid i = 1, \ldots, m\}$ of size m generated by using

$$y = f(x) + \epsilon, \tag{1}$$

where ϵ is noise with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Derive the bias-variance decomposition for the 1-nearest-neighbour regression. The response of the 1-NN regressor is defined as:

$$h_m(x) = y_{n(x)} = f(x_{n(x)}) + \epsilon,$$

where $n(x)$ gives the index of the nearest neighbour of x in $\mathcal{T}^m$. For simplicity assume that all x_i are the same for all training datasets $\mathcal{T}^m$ in consideration, hence, the randomness arises from the noise ϵ, only.

Give the squared bias:

$$\mathbb{E}_x\left[\big(g_m(x) - f(x)\big)^2\right] = \mathbb{E}_x\left[\Big(\mathbb{E}_{\mathcal{T}^m}\big(h_m(x)\big) - f(x)\Big)^2\right]$$

and variance:

$$\mathbb{V}_{x,\mathcal{T}^m}\big(h_m(x)\big).$$