STATISTICAL MACHINE LEARNING (WS2019)
EXAM T2 (90 MIN / 25P)

Assignment 1. **(4p)** Consider a linear classifier $h\colon \mathcal{X} \to \mathcal{Y}$ assigning inputs $x \in \mathcal{X}$ into two classes $\mathcal{Y} = \{-1, +1\}$ based on the rule

$$h(x; w, b) = \begin{cases} +1 & \text{if } \langle \phi(x), w \rangle + b \geq 0\,, \\ -1 & \text{if } \langle \phi(x), w \rangle + b < 0\,, \end{cases} \tag{1}$$

where $\phi\colon \mathcal{X} \to \mathbb{R}^n$ is a feature map and $(w, b) \in \mathbb{R}^{n+1}$ are parameters. Assume that you cannot evaluate the feature map $\phi(x)$ explicitly, however, you can evaluate a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$, $\forall x, x' \in \mathcal{X}$. Let $T^m = \{(x^j, y^j) \in (\mathcal{X} \times \mathcal{Y}) \mid j = 1, \ldots, m\}$ be a set of training examples. Describe a variant of the Perceptron algorithm which finds the parameters $(w, b) \in \mathbb{R}^{n+1}$ so that the classifier (1) predicts all examples from T^m correctly provided such parameters exist. The algorithm has to access the inputs $x \in \mathcal{X}$ only via the kernel function.

Assignment 2. **(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 for each $x \in \mathcal{X}$ returns a sequence $h(x) = (h_1(x), \ldots, h_n(x))$. Assume that we want to measure the prediction accuracy of $h(x)$ by the expected Hamming distance $R(h) = \mathrm{E}_{(x, y_1, \ldots, y_n) \sim p}\left(\sum_{i=1}^{n} [\![h_i(x) \neq y_i]\!]\right)$ where $p(x, y_1, \ldots, y_n)$ is a p.d.f. defined over $\mathcal{X} \times \mathcal{Y}$. As the distribution $p(x, y_1, \ldots, y_n)$ is unknown we estimate $R(h)$ by the test error

$$R_S(h) = \frac{1}{l} \sum_{j=1}^{l} \sum_{i=1}^{n} [\![y_i^j \neq h_i(x^j)]\!]$$

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

a) Assume that the sequence length is $n = 10$ and that we compute the test error from $l = 1000$ examples. Use the Hoeffding inequality to compute the minimal probability that $R(h)$ will be in the interval $(R_S(h) - 1, R_S(h) + 1)$.

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

Assignment 3. **(6p)** A real valued random variable has a Laplace distribution with location parameter μ if its probability density function is

$$p_\mu(x) = \frac{1}{2} \exp\Big(-|x - \mu|\Big).$$

a) Sketch the graph of this pdf. Show that given i.i.d. samples $x_1, x_2, \ldots, x_m$, the maximum likelihood estimate of μ is the sample median.

b) The probability density of the real valued random variable Y is a mixture of n Laplace distributions

$$p(y) = \sum_{k=1}^{n} \pi_k \, p_{\mu_k}(y).$$

Explain how to estimate their parameters $\mu_k \in \mathbb{R}$ and mixture weights π_k given an i.i.d. sample $y_1, y_2, \ldots, y_m$ by using the EM algorithm. Give a formula for the auxiliary variables $\alpha_i(k)$ in the E-step. Show that the optimisation task in the M-step decomposes into independent tasks for each μ_k and for the π-s. Give a formula for the optimal re-estimated π-s.

Assignment 4. (3p) Let us consider a standard Markov chain model for sequences $s = (s_1, \ldots, s_n)$ of length n with states $s_i \in K$ given by:

$$p(s) = p(s_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1}).$$

The conditional probabilities $p(s_i \mid s_{i-1})$ and the marginal probability $p(s_1)$ for the first element are known. We want to find the most probable sequence s without consecutive repetitions, i.e. $s_i \neq s_{i-1}$ for all $i = 2, \ldots, n$. Describe an efficient algorithm for solving this task. What complexity has it?

Assignment 5. (4p) Define a neural module (layer) joining a linear layer and a Softplus layer which is a smooth approximation of the ReLU layer. Give the forward, backward and parameter messages. Consider n inputs, K units of the linear layer and K units of the Softplus layer each processing the output of the corresponding unit of the preceding linear layer. Each Softplus unit applies the non-linearity

$$f(x) = \ln(1 + e^x).$$

- 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 6. (3p) A convolutional layer transforms an input volume $W_{\text{in}} \times H_{\text{in}} \times C$ into an output volume $W_{\text{out}} \times H_{\text{out}} \times D$, where W_{in} and H_{in} define spatial dimensions of the input and C is the number of input channels. Similarly W_{out} and H_{out} denote spatial dimensions of the output and D the number of filters. Consider stride S, zero padding P and filters having receptive field of $F \times F$ units.

(1) Give types and total number of parameters of the layer.
(2) Consider padding P preserving the size of the output in the W dimension, i.e., $W_{\text{in}} = W_{\text{out}}$. Give P as a function of F, S and W_{in}.