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

Assignment 1. (3p) 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

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

where $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ is a feature map and $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ are parameters.

Let $\mathcal{T}^m = \{(x^j, y^j) \in (\mathcal{X} \times \mathcal{Y}) \mid j = 1, \ldots, m\}$ be a set of training examples. Describe a variant of the Perceptron algorithm which finds the parameters $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ so that the classifier (1) predicts all examples from $\mathcal{T}^m$ correctly provided such parameters exists.

Assignment 2. (5p) We are given a set $\mathcal{H} = \{h_i\colon \mathcal{X} \to \{0, \ldots, 100\} \mid i = 1, \ldots, 100\}$ containing 100 Convolution Neural Networks, each being trained to predict a biological age $y \in \mathcal{Y} = \{0, \ldots, 100\}$ from a facial image $x \in \mathcal{X}$. The goal is to select a CNN with the minimal expected absolute deviation between the predicted and the true age

$$R(h) = \mathbb{E}_{(x,y)\sim p}(|y - h(x)|)\,,$$

where the expectation is w.r.t. an unknown distribution $p(x, y)$ generating the images. Because $p(x, y)$ is unknown, we approximate $R(h)$ by the empirical risk

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^{m} |y^j - h(x^j)|\,,$$

computed from a set $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ containing m examples i.i.d. drawn from $p(x, y)$.

a) Define a method based on the Empirical Risk Minimization which uses $\mathcal{T}^m$ to select the best CNN out of the given options $\mathcal{H}$.

b) What is the minimal number of the training examples m we need to collect in order to have a guarantee that $R(h)$ is in the interval $(R_{\mathcal{T}^m}(h) - 1, R_{\mathcal{T}^m}(h) + 1)$ for every $h \in \mathcal{H}$ with probability at least 95%?

Assignment 3. (5p) Consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}$ of fixed length n. Each sequence is a combination of a leading part $i \leqslant k$ and a trailing part $i > k$. The boundary $k = 0, \ldots, n$ is random with some categorical distribution $\boldsymbol{\pi} \in \mathbb{R}_+^{n+1}$, $\sum_k \pi_k = 1$. The values x_i, in the leading and trailing part are statistically independent and distributed with some probability density function $p_1(x)$ and $p_2(x)$ respectively. Altogether the distribution for pairs $(\boldsymbol{x}, k)$ reads

$$p(\boldsymbol{x}, k) = \pi_k \prod_{i=1}^{k} p_1(x_i) \prod_{j=k+1}^{n} p_2(x_j)\,. \tag{2}$$

The densities p_1 and p_2 are known. Given an i.i.d. sample of sequences $\mathcal{T}^m = \{\boldsymbol{x}^\ell \in \mathbb{R}^n \mid \ell = 1, \ldots, m\}$, the task is to estimate the unknown boundary distribution $\boldsymbol{\pi}$ by the EM-algorithm.

a) The E-step of the algorithm requires to compute the values of auxiliary variables $\alpha_\ell^{(t)}(k) = p(k \mid \boldsymbol{x}^\ell)$ for each example $\boldsymbol{x}^\ell$ given the current estimate $\boldsymbol{\pi}^{(t)}$ of the boundary distribution. Give a formula for computing these values from model (2).

b) The M-step requires to solve the optimisation problem

$$\frac{1}{m} \sum_{\ell=1}^{m} \sum_{k=0}^{n} \alpha_\ell^{(t)}(k) \log p(\boldsymbol{x}^\ell, k) \to \max_{\boldsymbol{\pi}} .$$

Substitute the model (2) and solve the optimisation task.

Assignment 4. (3p) Let us consider the following 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 assumed to be known.

Describe an efficient algorithm for computing the marginal probabilities $p(s_i)$ for all $i = 2, \ldots, n$. What complexity has it?

Assignment 5. (4p) Consider the squared logarithmic loss:

$$\ell\big(y, h(x)\big) = \Big(\log\big(1 + y\big) - \log\big(1 + h(x)\big)\Big)^2,$$

where y is the target and $h(x)$ the output of the regression for input x. Give the pseudo code for the corresponding Gradient Boosting Machine using this loss and discuss differences to the squared loss GBM.

Assignment 6. (5p) Define a neural module (layer) joining a linear layer and a PReLU (Parametric Rectified Linear Unit) layer. Give the forward, backward and parameter messages. Consider n inputs, K units of the linear layer and K units of the PReLU layer each processing the output of the corresponding unit of the preceding linear layer. Each PReLU unit applies the non-linearity $f(s) = \max(s, as)$, where $a \in \mathbb{R}$ is a trainable parameter.

- 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.