STATISTICAL MACHINE LEARNING (WS2016)
TEST (120 MIN / 26P)

Assignment 1. (3p) We are given a prediction strategy $h\colon \mathcal{X} \to \{0, 1, \ldots, 100\}$ estimating human age $y \in \{0, 1, \ldots, 100\}$ from an image $x \in \mathcal{X}$. The task is to estimate the expected absolute deviation between the predicted age and the true age

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

We are going to estimate $R^{\mathrm{MAE}}(h)$ by computing the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{i=1}^{l} |y^i - h(x^i)|$$

where $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y)$. What is the minimal number of test examples l which guarantees that $R^{\mathrm{MAE}}(h)$ is in the interval $[R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1]$ with probability at least 95%?

Hint: Use the Hoeffding inequality which states that for all $\varepsilon > 0$ it holds that

$$\mathbb{P}\left(\left| \frac{1}{l} \sum_{i=1}^{l} z^i - \mu \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(b-a)^2}}$$

where $\{z^1, \ldots, z^l\} \in [a, b]^l$ are realizations of independent random variables with the same expected value μ.

Assignment 2. (4p) Let us consider a non-homogeneous polynomial kernel of degree two, that is, a function $k\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined such that

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^2 \,.$$

(a) Define a feature map $\boldsymbol{\phi}\colon \mathbb{R}^n \to \mathbb{R}^d$ such that

$$\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle = k(\boldsymbol{x}, \boldsymbol{x}') \,.$$

(b) For $\boldsymbol{\phi}\colon \mathbb{R}^n \to \mathbb{R}^d$ defined in (a) write the feature space dimension d as a function of the input data dimension n.

Assignment 3. (5p) Consider a linear classifier $h\colon \mathcal{X} \times \mathbb{R}^n \times \mathbb{R} \to \{+1, -1\}$ defined as

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

where $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are parameters of the classifier and $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ is a feature map. Let $\mathcal{T}^m = \{(x^i, y^i) \in \mathcal{X} \times \{+1, -1\} \mid i = 1, 2, \ldots, m\}$ be a training set containing m examples of inputs and corresponding hidden states.

(a) Show that the function $F\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ defined as

$$F(\boldsymbol{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y^i \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i) \rangle\}$$

is an upper bound of the training error

$$R_{\mathcal{T}^m}(\boldsymbol{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i, \boldsymbol{w}, b))$$

of the classifier $h(x, \boldsymbol{w}, b)$ for every setting of parameters $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$.

(b) Show that learning $\boldsymbol{w}$ and b from $\mathcal{T}^m$ based on minimisation of the upper bound $F(\boldsymbol{w}, b)$ can be written as a linear program.

Assignment 4. (4p) The probability density function of an exponential distribution is $p_\lambda(x) = \lambda e^{-\lambda x}$ for $x \in \mathbb{R}_+$. It is parametrised by a single parameter $\lambda > 0$. You are given an i.i.d. sample $\mathcal{T}^m = \{x_i \in \mathbb{R}_+ \mid i = 1, \ldots, m\}$ generated from such a distribution with unknown λ. The task is to estimate λ by the maximum likelihood estimator.

(a) Show that the log-likelihood of the training data is a concave function of λ.

(b) Derive the formula for the optimal estimate of λ.

Assignment 5. (7p) 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 $k = 1, \ldots, n$ is random with some categorical distribution $\boldsymbol{\pi} \in \mathbb{R}^n_+$, $\sum_k \pi_k = 1$. 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

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

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 $\alpha^{(t)}(k \mid \boldsymbol{x}^\ell) = p(k \mid \boldsymbol{x}^\ell)$ 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 (1).

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

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

Substitute the model (1) and solve the optimisation task.

Assignment 6. (3p) Formally define the average pooling layer for a convolutional neural network. Unlike the max pooling layer, which outputs the maximal input of the receptive field, the average pooling returns the arithmetic mean of the inputs. Give both forward and backward messages.