STATISTICAL MACHINE LEARNING (WS2018)
TEST (90 MIN / 22P)

Assignment 1. (5p) We are given a set $\mathcal{H} = \{h_i \colon \mathcal{X} \to \{1,\ldots,100\} \mid i = 1,\ldots,1000\}$ containing 1000 strategies each predicting the human age $y \in \{1,\ldots,100\}$ from a facial image $x \in \mathcal{X}$. The quality of a single strategy is measured by the expected absolute deviation between the predicted age and the true age

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

where the expectation is computed w.r.t. an unknown distribution $p(x,y)$. The empirical estimate of $R^{\mathrm{MAE}}(h)$ reads

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

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

What is the minimal number of the training examples m which guarantees that $R^{\mathrm{MAE}}(h)$ is in the interval $[R_{\mathcal{T}^l}(h) - 1, R_{\mathcal{T}^l}(h) + 1]$ for every $h \in \mathcal{H}$ with probability at least 95%?

Assignment 2. (3p) Let the input observation be a vector $\boldsymbol{x} \in \mathbb{R}^d$. Let us consider a feature map $\boldsymbol{\phi}_q \colon \mathbb{R}^d \to \mathbb{R}^n$, $n = d^q$, whose entries are all possible q-th degree ordered products of the entries of $\boldsymbol{x}$. For example, if $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ and $q = 2$ then

$$\boldsymbol{\phi}_q(\boldsymbol{x}) = \begin{pmatrix} x_1x_1 \\ x_2x_1 \\ x_3x_1 \\ x_1x_2 \\ x_2x_2 \\ x_3x_2 \\ x_1x_3 \\ x_2x_3 \\ x_3x_3 \end{pmatrix}$$

Show that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ we can computed the dot product between $\boldsymbol{\phi}_q(\boldsymbol{x})$ and $\boldsymbol{\phi}_q(\boldsymbol{x}')$ as

$$\langle \boldsymbol{\phi}_q(\boldsymbol{x}), \boldsymbol{\phi}_q(\boldsymbol{x}') \rangle = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^q \ ,$$

that is, as the dot product of the original vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ powered to q.

Assignment 3. (4p) Let us consider the space of all linear classifiers mapping $\boldsymbol{x} \in \mathbb{R}^d$ to $\{-1, +1\}$, that is

$$\mathcal{H} = \left\{ h(\boldsymbol{x}; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \mid (\boldsymbol{w}, b) \in (\mathbb{R}^d \times \mathbb{R}) \right\} .$$

Show that the VC dimension of $\mathcal{H}$ is at least $d + 1$.

Assignment 4. (5p) The Radial Basis Function (RBF) neuron has a forward message

$$\rho(\boldsymbol{x}) = e^{-\beta \|\boldsymbol{x} - \boldsymbol{c}\|_2} ,$$

where β and $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ are the trainable parameters and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ the input vector. Let us have a network which combines a layer of RBF neurons and a single linear output neuron :

$$h(\boldsymbol{x}) = \sum_{i=1}^{K} w_i \rho_i(\boldsymbol{x}) + b,$$

where $\boldsymbol{w}$ and b represent the weight and the bias of the output neuron.

(a) Sketch the network.

(b) Give the parameter functions of the RBF layer, i.e., $\dfrac{\partial \rho_j(\boldsymbol{x})}{\partial c_i}$ and $\dfrac{\partial \rho_j(\boldsymbol{x})}{\partial \beta}$ for $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, K\}$.

(c) Consider the squared loss:

$$\ell(y, h(\boldsymbol{x})) = [y - h(\boldsymbol{x})]^2 ,$$

where y is the target value. Use the backpropagation to compute the gradient $\dfrac{\partial \ell}{\partial c_i}$.

Assignment 5. (5p) The probability density function of a Laplace distribution is given by

$$p(x) = \frac{1}{2} e^{-|x-\mu|} ,$$

where μ is a parameter. 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 μ. The task is to estimate it by the maximum likelihood estimator.

(a) Sketch the graph of the probability density.

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

(c) Prove that the ML estimator is given by the median of the training data.