STATISTICAL MACHINE LEARNING (WS2020)
EXAM FEB 4, 2021 (90 MIN / 30P)

Assignment 1 (6p). Assume you are going to learn a two-class classifier $h\colon \mathcal{X} \to \{+1, -1\}$ from examples with the goal to minimize the expected classification error. The classifier is selected from a hypothesis space $\mathcal{H}$ based on the minimal training error defined as the number of misclassified examples. Consider the following three cases of hypothesis space:

(1) $\mathcal{H}_1 = \{h(x) = \operatorname{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$.
(2) $\mathcal{H}_2 = \{h(x) = \operatorname{sign}(|x - \mu_1| - |x - \mu_2|) \mid \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}\}$.
(3) $\mathcal{H}_3 = \{h(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b) \mid \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$.

a) What is the Vapnik-Chervonenkis dimension of $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$?

b) Assume that in all three cases your algorithm can find a classifier with the minimal training error. In which cases is the algorithm statistically consistent?

Assignment 2 (4p). Given a training set of examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \{+1, -1\}) \mid i = 1, \ldots, m\}$ the SVM algorithm finds parameters of a linear classifier $h(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)$ by solving an unconstrained problem

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{arg\,min}_{(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}} F(\boldsymbol{w}, b)$$

where $F\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a convex function of the parameters.

a) Define the objective function $F(\boldsymbol{w}, b)$ and describe all its components.

b) How is the objective function $F(\boldsymbol{w}, b)$ related to the number of training errors?

Assignment 3 (6p). A non-negative random variable $X \geqslant 0$ has exponential distribution $p(x) = be^{-bx}$, where b is an unknown parameter.

a) Explain how to estimate this parameter from an i.i.d. training set $\mathcal{T}^m = \{x_j \in \mathbb{R}_+ \mid j = 1, \ldots, m\}$ by using the maximum likelihood estimator .

b) The random variable Y is a mixture

$$Y = \lambda X_1 + (1 - \lambda) X_2$$

of two exponentially distributed variables with unknown parameters b_1, b_2 and unknown mixture weight $0 < \lambda < 1$. Explain how to estimate all mixture parameters from an i.i.d. training set $\mathcal{T}^m = \{y_j \in \mathbb{R}_+ \mid j = 1, \ldots, m\}$ by using the EM-algorithm.

Assignment 4 (4p). Consider a homogeneous Markov model for sequences $s = (s_1, \ldots, s_n)$ with elements from a finite set K. Its joint distribution is given by

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

where $p(s_1 = k)$ is the marginal distribution for the first element of the sequence and $p(s_i = k \,|\, s_{i-1} = k')$ is the matrix of transition probabilities. Given a state $k^* \in K$, we want to know its expected number of occurrences in a sequence generated by the model. Give an algorithm for computing this expectation.
Hint: Use the fact that the expected value of a sum of random variables is equal to the sum of their expected values.

Assignment 5 (6p). Consider the following simple neural network having n inputs:

$$\hat{y}(\boldsymbol{x}, \boldsymbol{w}) = \sigma\left(\sum_{i=1}^{n} w_i\, x_i\right),$$

where σ is the logistic sigmoid function:

$$\sigma(s) = \frac{1}{1 + e^{-s}}.$$

The network is trained using Stochastic Gradient Descent where the training set can be described as $\mathcal{T}^m = \{(x^i, y^i) \in (\mathbb{R}^n \times \{0,1\}) \mid i = 1, \ldots, m\}$. The loss function is the binary cross-entropy:

$$\ell(y, \hat{y}) = y \log \hat{y} + (1 - y) \log(1 - \hat{y}).$$

(1) Use the back-propagation algorithm and derive the gradient for a single sample:

$$\nabla \ell(\boldsymbol{w}) = \left(\frac{\partial \ell}{\partial w_1}, \frac{\partial \ell}{\partial w_2}, \ldots, \frac{\partial \ell}{\partial w_n}\right).$$

(2) Reuse the neuron activity computed during the forward pass and simplify the result.

Assignment 6 (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-neighbor 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 neighbor 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:

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