STATISTICAL MACHINE LEARNING (WS2020)
EXAM 19.01.20 (90 MIN / 28P)

Assignment 1 (6p). 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 $R(h) = \mathbb{E}_{(x,y_1,\ldots,y_n)\sim p}(\sum_{i=1}^{n} [\![h_i(x) \neq y_i]\!])$ 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

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

where $\mathcal{S}^l = \{(x^i, y_1^i, \ldots, y_n^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ 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. What is the minimal probability that $R(h)$ will be in the interval $(R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1)$?

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 $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability γ at least? Write l as a function of ε, n and γ.

Assignment 2 (4p). Assume we are given a training set of examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \{+1, -1\}) \mid i = 1, \ldots, m\}$ which is known to be linearly separable with respect to a feature map $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$. That is, we can find parameters $(\boldsymbol{w}, b) \in \mathbb{R}^{n+1}$ of a linear classifier $h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle + b)$ which has zero training error. Assume that you cannot evaluate the feature map $\boldsymbol{\phi}(x)$ because it is either unknown or its evaluation is expensive. However, you know how to cheaply evaluate a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, x') = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x') \rangle, \forall x, x' \in \mathcal{X}$. Show how to find in this case parameters of the linear classifier by Perceptron algorithm and how to evaluate the linear classifier.

Assignment 3 (4p). Suppose you have learned a neural network classifier that predicts the posterior class probabilities $p(k \,|\, x)$ for an unknown distribution

$$p(x, k) = p(x \,|\, k)\, p(k) = p(k \,|\, x)\, p(x).$$

The network uses softmax activation in the last layer and has been learned on an i.i.d. training set $\mathcal{T}^m = \{(x^j, k^j) \mid j = 1, \ldots, m\}$. You want to apply the classifier in some application domain with "shifted" class priors $p(k) \to p_a(k)$. The appearance probabilities $p(x \,|\, k)$ remain unchanged. Explain how to utilise the classifier for this domain without retraining it.
Remark: We assume that the shifted class priors $p_a(k)$ are known.

Assignment 4 (6p). We consider the following probabilistic model for real valued sequences $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ of length n. The elements of the leading part, up to some position k are independent and normally distributed with mean μ_1 and variance σ_1^2. The trailing elements are independent and normally distributed with mean μ_2 and variance σ_2^2. The boundary position k between the two parts is itself random and follows a categorical distribution with probabilities $p(k) = \pi_k$.

a) Explain how to estimate the model parameters $\mu_{1,2}$, $\sigma_{1,2}$ and π_k, $k = 1, \ldots, n$ from i.i.d. training data $\mathcal{T}^m = \left\{(\boldsymbol{x}^j, k_j) \;\middle|\; j = 1, \ldots, m\right\}$ by using the maximum likelihood estimator.

b) Assume now that the model parameters are known. Given a sequence x we want to predict the boundary position between the leading and trailing part. We want to use the quadratic loss $\ell(k, k') = (k - k')^2$. Show that the optimal prediction for the boundary is given by

$$k^* = \sum_{k=0}^{n} k\, p(k \mid x).$$

Assignment 5 (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}.

Assignment 6 (5p). Define a neural module (layer) joining a linear layer and an ELU (Exponential Linear Unit) layer. Give the forward, backward and parameter messages. Consider n inputs, K units of the linear layer and K units of the ELU layer each processing the output of the corresponding unit of the preceding linear layer. Each ELU unit applies the non-linearity:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \exp(x) - 1, & \text{if } x \leq 0. \end{cases}$$

- 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.