STATISTICAL MACHINE LEARNING (WS2022)
EXAM T1 (90MIN / 26P)

Assignment 1 (6p). Assume we are training a Convolution Neural Network (CNN) classifier $h\colon \mathcal{X} \to \mathcal{Y}$ to predict a digit $y \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ from an image $x \in \mathcal{X}$. We train the CNN by the Stochastic Gradient Descent (SGD) algorithm using 100 epochs. After each epoch we save the current weights so that at the end of training we have a set $\mathcal{H} = \{h_t\colon \mathcal{X} \to \mathcal{Y} \mid i = 1, \ldots, 100\}$ containing 100 different CNN classifiers. The goal is to select the best CNN out of $\mathcal{H}$ with small classification error

$$R(h) = \mathbb{E}_{(x,y)\sim p}([\![y \neq h(x)]\!])$$

where the expectation is w.r.t. an unknown distribution $p(x, y)$ generating the data. Ideally, we would like to find the best hypothesis $h_{\mathcal{H}} \in \operatorname{Arg\,min}_{h\in\mathcal{H}} R(h)$. We select the model $\hat{h} \in \operatorname{Arg\,min}_{h\in\mathcal{H}} R_{\mathcal{V}^m}(h)$ by minimizing the validation error

$$R_{\mathcal{V}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} [\![y^j \neq h(x^j)]\!]$$

computed on the validation set $\mathcal{V}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ i.i.d. drawn from $p(x, y)$. We evaluate the performance of the selected model $\hat{h}$ by the test error

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

computed on the test set $\mathcal{S}^n = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, n\}$ i.i.d. drawn from $p(x, y)$.

a) What is the minimal size m of the validation set $\mathcal{V}^m$ to have a guarantee that

$$R(\hat{h}) \leq R(h_{\mathcal{H}}) + 0.05$$

holds with probability 95% at least?

b) Assume that the test set has $n = 20\,000$ examples. Compute the minimal ε such that

$$R(\hat{h}) \in (R_{\mathcal{S}^n}(\hat{h}) - \varepsilon, R_{\mathcal{S}^n}(\hat{h}) + \varepsilon)$$

holds with probability 95% at least.

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 with margin-rescaling loss 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 (4p). The Poisson distribution for a random variable X with values $x = 0, 1, 2 \ldots$ is given by

$$p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

where $\lambda > 0$ is a parameter.

a) Show that it is an exponential family. Give its base measure, sufficient statistic, natural parameter (as a function of λ) and the cumulant function.

b) Given an i.i.d. training set $\mathcal{T}^m = \{x_i \in \mathbb{N}_0 \mid i = 1, 2, \ldots, m\}$, we want to estimate the parameter λ. Derive the maximum likelihood estimator for this distribution family.

Assignment 4 (5p). In a factory there are two machines that look identical, machine A and machine B. One machine works fine, and produces signals x which are Gaussian distributed with mean μ and variance σ_0^2. The other machine is broken, and produces signals x which are Gaussian distributed with mean μ and variance $\sigma_1^2 > \sigma_0^2$.

A worker is sent to determine whether machine A is broken and machine B is fine, or, conversely, machine A is fine and machine B is broken. The worker takes n_A measurements from the machine A, and n_B from the machine B.

The dataset from the machine A is $\mathcal{X}^A = \{x_1^A, x_2^A, \cdots, x_{n_A}^A\}$ and the dataset from the machine B is $\mathcal{X}^B = \{x_1^B, x_2^B, \cdots, x_{n_B}^B\}$. All measurements are statistically independent. We know that one machine is fine and the other is broken. The parameters μ, σ_0 and σ_1 are known.

Propose a classifier that uses the datasets $\mathcal{X}^A$, and $\mathcal{X}^B$ to determine which machine is broken, assuming 0/1 loss. Show that the mean square deviations $V_A = \frac{1}{n_A}\sum_i (x_i^A - \mu)^2$ and $V_B = \frac{1}{n_B}\sum_j (x_j^B - \mu)^2$ are the only statistics needed for the decision.

Assignment 5 (4p). Consider the Max Pooling layer which reduces the dimensionality of a two dimensional input. The forward message of max pooling is

$$f_{kl}(\boldsymbol{x}) = \max_{(i,j)\in\Omega(k,l)} x_{ij}, \tag{1}$$

where (i, j) denotes the coordinates of the input, (k, l) denotes the coordinates of the output and $\Omega(k, l)$ is the set of input coordinates covered by the receptive field of the output (k, l) as shown in Figure 1.

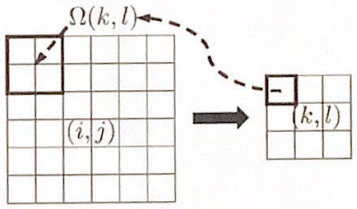

FIGURE 1. Max pooling layer with filter size $F = 2$ and stride $S = 2$.

Derive the backward message $\frac{\partial f_{kl}}{\partial x_{ij}}$ of the layer.

Assignment 6 (3p). Consider the squared log-cosh loss:

$$\ell\big(y, h(x)\big) = \log(\cosh(h(x) - y)),$$

where y is the target and $h(x)$ the output of the regressor for input x. The cosh function is defined as $\cosh(x) = \frac{e^x + e^{-x}}{2}$. Give the pseudo code for the corresponding Gradient Boosting Machine using this loss (including the gradient).