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

Assignment 1 (6p). Let the observation $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^n$ and the hidden state $y \in \mathcal{Y} = \{+1, -1\}$ be generated from a multivariate normal distribution

$$p(\boldsymbol{x}, y) = p(y)\, \frac{1}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{C}_y)^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^T \boldsymbol{C}_y^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y)} \tag{1}$$

where $\boldsymbol{\mu}_y \in \mathbb{R}^n$, $y \in \mathcal{Y}$, are mean vectors, $\boldsymbol{C}_y \in \mathbb{R}^{n\times n}$, $y \in \mathcal{Y}$, are covariance matrices and $p(y)$ is a prior probability. Assume that the model parameters are unknown and we want to learn a strategy $h \in \mathcal{X} \to \mathcal{Y}$ which minimizes the probability of misclassification. To this end we use a learning algorithm $A\colon \cup_{m=1}^{\infty}(\mathcal{X}\times\mathcal{Y})^m \to \mathcal{H}$ which based on samples generated from the distribution (1) returns a strategy h from the class $\mathcal{H} = \{h(\boldsymbol{x}) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b) \mid \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$ containing all linear classifiers.

a) What is the approximation error in case that $\boldsymbol{C}_+ = \boldsymbol{C}_-$?

b) Is the approximation error going to increase or decrease if $\boldsymbol{C}_+ \neq \boldsymbol{C}_-$?

c) Assume the algorithm finds a linear classifier which has the minimal classification error on the training examples. Is the algorithm statistically consistent?

Explain your answers.

Assignment 2 (6p). Assume we are training a Convolution Neural Network (CNN) based 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 CNN classifiers. The goal is to select the best CNN out of $\mathcal{H}$ that has the minimal classification error

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

where the expectation is w.r.t. an unknown distribution $p(x, y)$ generating the data. Because $p(x, y)$ is unknown, we approximate $R(h)$ by the empirical risk

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

computed from a validation set $\mathcal{V}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ containing m examples i.i.d. drawn from $p(x, y)$.

a) Define a method based on the Empirical Risk Minimization which uses $\mathcal{V}^m$ to select the best CNN out of a finite hypothesis class $\mathcal{H}$.

b) What is the minimal number of validation examples m we need to collect in order to have a guarantee that $R(h)$ is in the interval $(R_{\mathcal{V}^m}(h) - 0.01, R_{\mathcal{V}^m}(h) + 0.01)$ for every $h \in \mathcal{H}$ with probability at least 95%?

Assignment 3 (6p). Consider the following mixture model for sequences $s = (s_1, \dots, s_n)$ of discrete states $s_i \in K$

$$p(s) = \pi p_1(s) + (1 - \pi) p_2(s),$$

where $p_{1/2}$ are homogeneous Markov models

$$p_{1/2}(s) = p_{1/2}(s_1) \prod_{i=2}^{n} p_{1/2}(s_i \mid s_{i-1}).$$

Neither the mixture coefficient π nor the parameters of the two Markov models are known. You are given an i.i.d. training set $\mathcal{T}^m = \{s^j \in K^n \mid j = 1, \dots, m\}$ of sequences. Explain how to learn all parameters of the mixture by an EM algorithm. Recall that the parameters of a homogeneous Markov model are the probabilities for the first state $p(s_1 = k)$ and the matrix of position independent transition probabilities $p(s_i = k \mid s_{i-1} = k')$.

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

$$f_{kl}(\boldsymbol{x}) = \max_{(i,j) \in \Omega(k,l)} x_{ij}, \tag{2}$$

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.

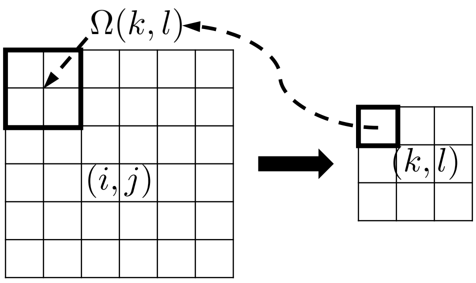

FIGURE 1. Max pooling layer with filter size $F = 2$ and stride $S = 2$.

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

Assignment 5 (5p). Consider the squared logarithmic loss:

$$\ell\big(y, h(x)\big) = \Big(\log\big(1 + y\big) - \log\big(1 + h(x)\big)\Big)^2,$$

where y is the target and $h(x)$ the output of the regressor for input x. Give the pseudo code for the corresponding Gradient Boosting Machine using this loss (including the gradient) and discuss differences to the squared loss GBM.