DEEP LEARNING (SS2020)
EXAM T0 (90 MIN / 23P)

Assignment 1. (6p) A neural network has been trained for classifying patterns by predicting the posterior class probabilities $p(y \mid x)$, $y \in K$. The relative class frequencies in the training set were $p(y)$. When applying the network, it turned out that the prior class probabilities for real data are different and equal to $p^*(y)$. Explain how to use the network as a predictor without re-training it. We assume the 0/1 loss.

Assignment 2. (4p) We want to train a fully convolutional network for semantic segmentation of color images $x\colon D \to \mathbb{R}^3$, where D denotes the pixel domain of the image. The training data consist of pairs (x, y), where x is a colour image and $y\colon D \to K$ is a partial annotation. The label set K consists of semantic labels and one additional label "unknown". Explain how to train the CNN on such data. The network should output only semantic labels.

Assignment 3. (4p) Let us consider a batch normalization layer applied directly after a linear layer. Let $x(i)$ denote the output of a single node in the linear layer, where i is the index in the mini-batch. The output $y(i)$ of the corresponding node of the batch-norm layer is given by

$$y(i) = \frac{x(i) - u}{s}\gamma + \delta,$$

where u denotes the mini-batch average of x and s^2 denotes the corresponding variance. The parameters of the batch-norm layer are denoted as γ and δ. Give a formula for the Jacobian $\frac{\partial y(i)}{\partial x(j)}$ of the batch-norm layer.

Assignment 4. (5p) We have a training set (x_i, y_i), $i = 1 \ldots n$ and want to train a classifier in the form of a feed-forward neural network that maps observations x to class labels y and has parameters θ.

(1) Define the maximum-likelihood criterion.
(2) Define the cross-entropy loss.
(3) Let p^* be the empirical distribution of the training set. Define the KL divergence between $p^*(y|x)$ and the model predictive distribution. Take the expectation of this KL divergence over $p^*(x)$.
(4) Show that all 3 criteria are equivalent for learning the network parameters.

Assignment 5. (4p) Recall SGD with momentum:

$$\begin{aligned} v_{t+1} &= \mu v_t + g_t \\ \theta_{t+1} &= \theta_t - \varepsilon v_{t+1}, \end{aligned} \tag{1}$$

where g_t is the stochastic gradient at point θ_t. Derive this algorithm by applying a variance reduction technique to stochastic gradient estimates in plain SGD.