CS-E4890 DEEP LEARNING exam 15.12.2017 **Kivinen, Laaksonen, Keurulainen**

Allowed equipment: pens, pencils, erasers, a non-programmable calculator

Remember to give course feedback to earn one point!

Question 1 [max. 9 points]: Explain the following concepts or abbreviations briefly (each with 30–50 words, with a mathematical definition and/or with an illustration):

(i) information capacity

(ii) convolutional neural network

(iii) Jacobian matrix

(iv) unfolded computational graph

(v) equivariance vs. invariance

(vi) vanishing and exploding gradient problem

(vii) momentum

(viii) minibatch

(ix) parameter sharing

Question 2 [max. 6 points]: (i) What is the self-information $I(x)$ of event $\mathrm{x} = x$ given its probability $P(x)$? (ii) What is the Shannon entropy $H(\mathrm{x}) = H(P)$ of the probability distribution $P(\mathrm{x})$? (iii) Given another probability distribution $Q(\mathrm{x})$, what is the Kullback-Leibler divergence $D_{\mathrm{KL}}(P||Q)$ between them? (iv) Show that for cross-entropy $H(P,Q)$ it is true that

$$H(P,Q) = H(P) + D_{\mathrm{KL}}(P||Q) = -E_{\mathrm{x}\sim P}[\log Q(x)]$$

(v) Assuming that $P(\mathrm{x})$ is the empirical data distribution and $Q(\mathrm{x})$ is the probability given by a model for input $\mathrm{x} = x$, and that the training data $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ are independent and identically distributed, show that the maximum likelihood estimate

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} Q(\mathcal{X};\theta)$$

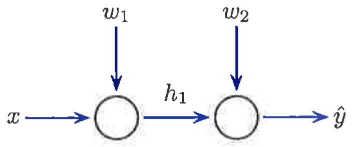
of the model parameter θ is equal to the minimum cross-entropy estimate

$$\theta_{\mathrm{MCE}} = \arg\min_{\theta} -E_{\mathrm{x}\sim P}[\log Q(x;\theta)] = \arg\max_{\theta} E_{\mathrm{x}\sim P}[\log Q(x;\theta)]$$

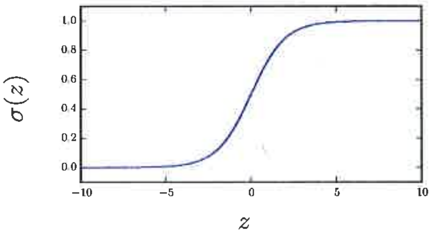
Question 3 [max. 6 points]: When designing a deep fully-connected feedforward neural network, one needs to make decisions about the following issues: (i) the type of the optimization algorithm, (ii) the output cost function, (iii) the type of the output units, and (iv) the type of the hidden units. Describe briefly some central alternatives for each of the decisions and the criteria for selecting the best choice among the alternatives.

Question 4 [max. 3 points]: Assume the very simple neural network with a scalar input, scalar weights, scalar output, and two artificial neurons, as illustrated in the below-left graph; there are no biases in this simplified neural network.

Neural network under study:



Logistic sigmoid function:



The activation function in both neurons is the logistic sigmoid

$$\sigma(z) = \frac{1}{1+e^{-z}} \text{ ; the function is visualized in the above-right graph.}$$

The cost function applied to the network output $\hat{y}$ is

$$J(x; w_1, w_2) = \frac{1}{2}(\hat{y} - y)^2 \, .$$

Derive the analytical expressions of the partial derivatives of the cost function with respect to the parameters w_2 and w_1, i.e., $\frac{\partial J}{\partial w_2}$ and $\frac{\partial J}{\partial w_1}$, using back-propagation.

Given the numerical values of $x = 0.5$, $w_1 = -1$, $w_2 = 1$, and $y = 1$, calculate the numerical values of $\hat{y}$, J, $\frac{\partial J}{\partial w_2}$, and $\frac{\partial J}{\partial w_1}$.

Question 5 [max. 3 points]: Name THREE different regularization methods in deep learning and explain briefly with a couple of sentences the key motivation and idea behind each of them.

Question 6 [max. 6 points]: Let $\mathbf{u} = \{\mathbf{v}, \mathbf{h}\}$ denote the random variables in a neural network having stochastic units, with $\mathbf{v} = \{v_i\}_{i=1}^{I}$ denoting the set of visible (observed data variable) units, $\mathbf{h} = \{h_j\}_{j=1}^{J}$ the set of latent (hidden) variable units. Select TWO of the following model families:

- linear factor model
- variational autoencoder
- restricted Boltzmann machine,

and ONE particular representative model out of each of the selected families. For each of these two models,

- Illustrate the model architecture graphically, drawing nodes for the random variables and the connection weights between them.
- Discuss the functional form of the probability distribution of (i) $\mathbf{v}, \mathbf{h}$, (ii) $\mathbf{h}$, (iii) $\mathbf{v}|\mathbf{h}$, (iv) $\mathbf{v}$, (v) $\mathbf{h}|\mathbf{v}$, and discuss the relative computational complexity of evaluating them.

Finally, discuss similarities with and differences between the two models you had chosen, in terms of network architecture, and in terms of computational complexities of evaluating the distributions mentioned above.