First and Last Name: **ID #:**

Course on
Deep Learning

Academic Year 2022/2023

Exmaple of Exam

Instructions

- Write *Name*, *Surname* and *ID #* on **each** sheet (**odd** pages only).
- Write the answer in the white space below the question; It is not possible to attach additional sheets, so try to be clear and not verbose.
- In case of errors, please clearly indicate which part of the answer must be considered rated; cancel the irrelevant parts.
- Make sure no sheet is missing at the time of delivery.

First Part

Question 1

For which label distribution and with which loss is it reasonable to adopt the sigmoid activation function for the output layer, according to the maximum likelihood principle? (select the **true** answer)

[a] Gaussian distribution / Mean Squared Error

[b] Gaussian distribution / Cross-Entropy

[c] Multinoulli distribution / Cross-Entropy

[d] Bernoulli distribution / Cross-Entropy

[f] None of the above.

Question 2

Consider a CNN layer with 10 filters of size 4x4, a stride of 1 and input images of size 8x8. How many parameters are we required to train for such a layer? (do not consider the bias terms) Please answer with the exact number of the parameters (no formulas)

Answer:

Question 3

Which one of the following is an advantage of using deep neural networks over linear models? (select the correct answer)

[a] A Deep neural network has the same expressive power of linear models

[b] A Neural Network always performs better than linear models

[c] A deep neural network has better generalization than a linear model

[d] The functions we want to learn are always a composition of simpler functions, so deep neural networks are always more suited for learning problems

[e] None of the above

Question 4

The main feature of a Denoising Autoencoder is: (select the correct answer)

[a] The use of an architecture with a hidden layer with a number of units that is much lower that the dimension of the input space.

[b] The use of an architecture with a hidden layer of linear units.

[c] The use of data that has been preprocessed to remove noise.

[d] The use of an architecture with a first recurrent layer of sigmoidal units to reduce the noise in input.

[e] The use of input data corrupted by noise.

Question 5

Training of a Restricted Boltzmann Machine is performed thanks to: (select the correct answer)

[a] The standard Back-propagation algorithm.

[b] Gradient descent plus ancestral sampling.

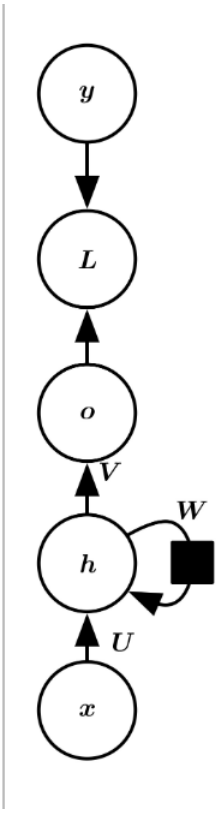
[c] Gradient ascend plus ancestral sampling.

[d] A multi-phase algorithm based only on Gibbs sampling.

[e] Gradient ascent plus Gibbs sampling.

Question 6

Suppose to have a IO-isomorphic prediction task and a RNN with the following Recurrent Network:



where $\boldsymbol{y}$ is the target, $\boldsymbol{L}$ the loss function, $\boldsymbol{o}$ the RNN output, $h\mathbf{h}$ the hidden state, $\boldsymbol{x}$ the input at time t, and the black square represents the time-shift operator q^{-1}. $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{V}$ are weights matrices. Suppose to use back-propagation through time with mini-batch equal to 1 for training. Given an input sequences composed of 4 items, how many terms should be *summed* up to compute the gradient of the loss with respect to $\boldsymbol{W}$? [do not consider the contribution of $\mathbf{h}^{(0)}$ which is the zero vector].

Answer:

Question 7

Explain in detail what is the role of Monte Carlo Chains in the training of a stochastic neural network. Give an example of a neural network model where Monte Carlo Chains are used.

Question 8

In the context of sequential transductions, give the definition of causality and discuss how this concept is implemented in Recurrent Neural Networks (RNN). Are all RNN architectures causal ?

Question 9

Why is it convenient to use more that one hidden layer in neural networks? In other words, what is the advantage of a multi-layer neural network over a single-hidden-layer neural network?.

Question 10

What are the main problems we face when optimising deep neural networks? Foe each one, explain why it is a problem for optimisation algorithms. Where applicable, explain how it is possible to avoid such problems.

Second Part

Question 1

Given the Neural Network described by the following equations:

$$\begin{aligned} \mathbf{h}^{(1)} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \mathbf{a}^{(1)} &= ReLU(\mathbf{h}^{(1)}) \\ h^{(2)} &= (\mathbf{w}^{(2)})^\top \mathbf{a}^{(1)} + b^{(2)} \\ y &= \sigma(h^{(2)}) \\ J &= \frac{1}{2}(t-y)^2 \end{aligned}$$

with $\mathbf{x} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$, $\mathbf{W}^{(1)} = \begin{bmatrix} 0.5 & 0.75 \\ 1 & -1 \end{bmatrix}$, $\mathbf{b}^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{w}^{(2)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $b^{(2)} = -1$ and $t = 1$, compute the values of $\frac{\partial J}{\partial \mathbf{W}^{(1)}}$ and $\frac{\partial J}{\partial \mathbf{w}^{(2)}}$. (Hint: $\sigma(0) = 0.5$)

Question 2

In the context of Restricted Boltzmann Networks, write all the steps to prove the following result:

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i=1}^{n_v} \sigma\left((2\boldsymbol{v} - 1) \odot (\boldsymbol{b} + \boldsymbol{W}\boldsymbol{h})\right)_i$$