Exam

Instructions:

- Please write your matriculation number on the top of all pages.
- Please write your solutions in the bold boxes on the assignment sheets.
- Only pen and paper are allowed.
- **Good Luck!**

Name	
Matriculation number	
Field of study	

Exercise 1: Basics in Machine Learning [10 points]

a) Mark whether the following statements are true or false. *Note:* For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
K-nearest neighbors algorithms is a parametric method.		
Hyperparameters are trained to optimize the results on the training set.		
Generalisation is a central problem of machine learning.		
An overfitted model perfectly matches the evaluation data.		
Regularization typically increases the error on the training set.		
Logistic regression is typically used to predict real values.		
Empirical risk is an average of a loss function on a finite development set.		
Support vector machine can be used only for binary classification tasks.		
Each data item is assigned to exactly one cluster with k-means clustering.		
In active learning, systems actively select queries and request feedback from human.		

Exercise 2: Neural Networks in General [5+4+3+5 points]

a) The task of the following network is to predict $y = \begin{pmatrix} \text{p(positive)} \\ \text{p(negative)} \\ \text{p(neutral)} \end{pmatrix}$ given a tweet in bag-of-words representation as input. The network has two fully-connected hidden layers. The non-linear activation in the two hidden layers is $f = x^2$, the element-wise square of vector x. The output layer uses softmax activation. The first, the second hidden layer and the output layer do not have biases. The network has the following weight matrices:

$$W^1 \in \mathbb{R}^{2\times4} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \end{bmatrix}, W^2 \in \mathbb{R}^{2\times2} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}, W^3 \in \mathbb{R}^{3\times2} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Compute the output y of the network for the tweet ‘this exam is fair’ Represent the tweet as a 4-dimensional bag-of-words vector given the following 4-word vocabulary
V = $\{v_1$: exam, v_2 : good, v_3 : bad, v_4 : fair$\}$.

Compute the cross-entropy loss (using $\log_{10}$) for the output y, given the correct label $\hat{y} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.

Always round to 1 decimal points.

b) Given the following image. Fill each of the boxes with two missing equations.

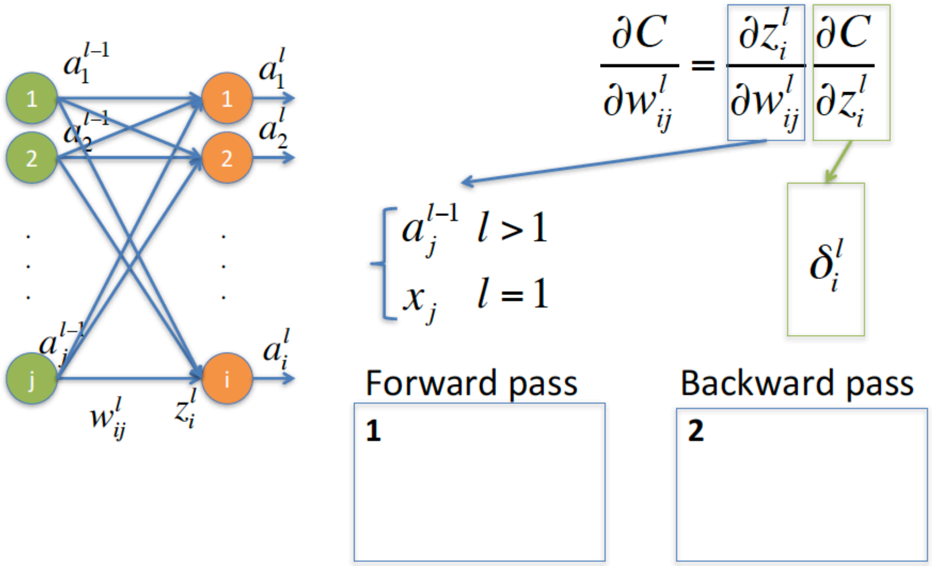


Abbildung 1: Backpropagation algorithms

c) Why does parameter initialization have a strong impact on the final results? Describe one possibility how to initialize the parameters.

d) Mark whether the following statements are true or false. *Note:* For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
Feed-forward neural networks can be used for emotion recognition tasks.		
The number of layers is a hyper-parameter of feed-forward neural networks.		
The number of neurons is a parameter of feed-forward neural networks.		
For multiclass multilabel classification tasks, sigmoid is a commonly used function for the output layer.		
Mini-batch gradient descent methods can be used to train neural networks.		

Exercise 3: Convolutional Neural Networks [2+3+5 points]

a) Convolutional neural networks have been proposed for image processing. What is the intuition of using them for language and how does the input look like in the case of language?

b) Compute the output of a convolutional layer with the following input and filter matrices (stride = 2×2) as well as the output of a subsequent max pooling layer with the given pooling window (stride = 1×1). Note that the non-linear activation is $f = x^2$ and we do not apply padding.

Input matrix:

$$\begin{pmatrix} 2 & 2 & 1 & 0 \\ -1 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 0 & 1 & 2 & -2 \end{pmatrix}$$

Filter matrix:

$$\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}$$

Pooling window: 2×2

c) Mark whether the following statements are true or false. *Note:* For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
The filter weights are trainable parameters of a CNN.		
A convolutional layer with 10 filters (with bias) of size 3x3 has 100 trainable parameters.		
Zero padding is only necessary when processing several input matrices at a time (in a batch or mini-batch).		
A typical convolutional layer for language spans the whole sentence.		
The average pooling layer can be used to downsample a matrix.		

Exercise 4: Recurrent Neural Networks [5+4+3+5 points]

a) Compute the output of the softmax layer of the last hidden state for the input sentence 'deep learning'. Represent each word as one-hot vector given the 3-word vocabulary V = $\{v_1$: machine, v_2 : deep, v_3 : learning$\}$.
The memory (a_0) is initialized as the vector $[-1; 1]$. Use the following weights of the network, there are no biases in any layer. Always round to 1 decimal points. IMPORTANT: In order to simplify the task, replace the *sigmoid* function with $f = x^2$, the element-wise square of vector x.

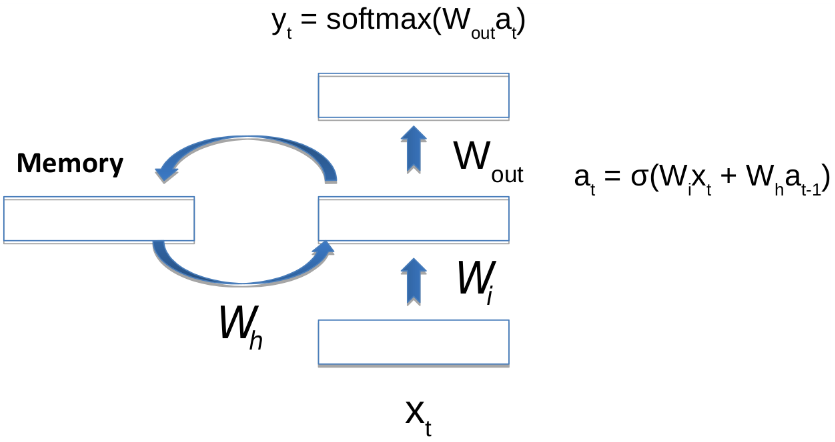

y_t = softmax($W_{out}a_t$)
Memory
W_{out}
$a_t = \sigma(W_i x_t + W_h a_{t-1})$
W_i
W_h
x_t

Abbildung 2: RNN architecture

$$W_{\mathrm{i}} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, W_{\mathrm{h}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, W_{out} = \begin{bmatrix} -1 & 1 \\ 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{1}$$

b) A Long short-term memory (LSTM) cell is defined by the following equations:

$$f_t = sigmoid(W_f[h_{t-1}, x_{t-1}] + b_f) \tag{2}$$

$$i_t = sigmoid(W_i[h_{t-1}, x_{t-1}] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{5}$$

$$o_t = \tanh(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \odot \tanh(C_t) \tag{7}$$

Identify three lines with mistakes in these equations and correct them.

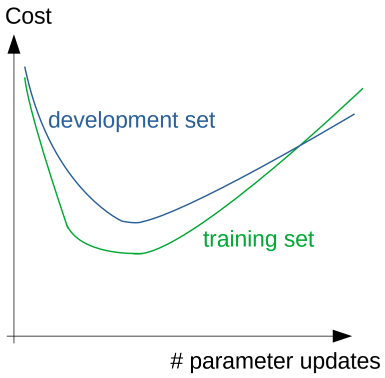
c) Explain the idea of the backpropation through time algorithms.

d) Mark whether the following statements are true or false. *Note:* For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
RNN can be used for language modeling task.		
In an Jordan network, the output of the hidden layer from previous time step is stored in the memory.		
The number of parameters of LSTM unit is four times the number of parameters of original RNN unit.		
In a LSTM unit, the memory and the output are the same.		
RNN suffers from vanishing and exploding gradients.		

Exercise 5: Tricks [2+3+5 points]

a) When training a neural network and monitoring its loss on the training and development data, the following curves are observed.



Do the curves behave as expected? Explain two strategies that can be applied to the training in order to receive a good model in the end.

b) Given a multi-layer perceptron with 20 hidden layers with 200 neurons each and use sigmoid activation. Which problem might occur during training, how can you detect it and how could you solve it?

c) Mark whether the following statements are true or false. *Note:* For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
Gradient clipping helps in the case of exploding gradients.		
Given a training set with 100 examples, stochastic gradient descent performs one update step per epoch.		
Dropout is a regularization technique.		
A fully-connected layer with ReLU activation and residual connection has the form $\mathrm{relu}(Wx + b) + x$		
Highway Network automatically removes unnecessary layers.		

Exercise 6: Deep Reinforcement Learning [4+4+5 points]

Given is a Markov Decision Process with state space $S = (s_1, s_2, s_3, s_4)$, action space $A = (a_1, a_2, a_3)$ with three discrete actions. The state transition and reward functions are deterministic. Assume the states to be one-hot encodings of their indices (meaning $s_1 = [1, 0, 0, 0]^T$, $s_2 = [0, 1, 0, 0]^T$, …).

- Figure 3 shows one possible architecture for an action-value function $Q(s, a)$ with $s_t \in S$ and $a_1, a_2, a_3 \in A$. Sketch a different architecture with the same number of layers and without changing the layer types. Describe their differences and compare them by listing advantages/disadvantages. (*hint*: consider different inputs / outputs).

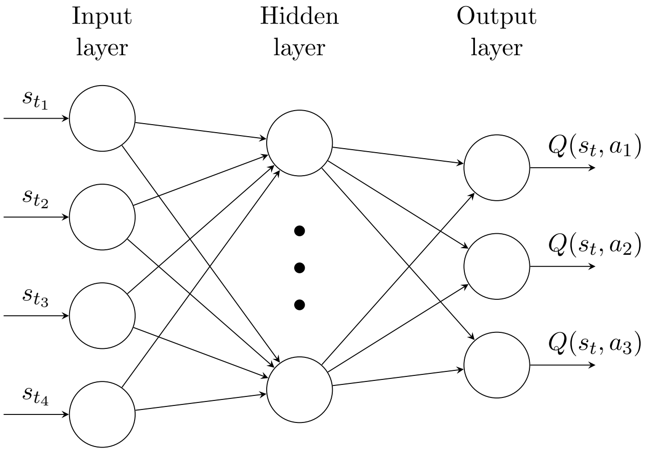


Abbildung 3: Example architecture for an action-value function network.

- From now on, assume the activation function after the hidden layer to be a Rectified Linear Unit (ReLU). Thus, the action-value function for the example architecture in figure 3 is given by

  $$Q(\mathbf{s}, a_i, \theta) = [ReLU(\theta \mathbf{s})]_i \tag{8}$$

  with parameters $\theta = \begin{pmatrix} -1 & -4 & 0 & 2 \\ 3 & 3 & 1 & 1 \\ 1 & 5 & 1 & 1 \end{pmatrix}$.

  (*Note: the index i on the right hand side of equation 8 refers to the i-th row of the vector*)
  Calculate the next action taken by the agent if it acts greedily w.r.t Q and starts in state s_2.

- Mark whether the following statements are true or false. Note: For each correct answer you will get one point and for each wrong answer you will get one MINUS point. The minimum amount of points for this task is 0.

Statement	**True**	**False**
Reinforcement Learning algorithms are trained on a static dataset.		
A value function is specific to a policy.		
Temporal Difference learning is a model-free approach.		
Deep-Q-Networks learning algorithm uses Cross-Entropy loss.		
Deep-Q-Networks learning algorithm draws mini-batches randomly from a buffer.		

Exercise 7: General Questions [3+4+3 points]

a) Describe the problem setting of meta supervised learning. What is the main difference compared to supervised learning?

b) Explain the basic ideas of generative adversarial networks.

c) Among eleven ethical principles identified in existing AI guidelines, name three of them and describe them shortly.