Exam in
Neural Networks and Learning Systems
TBMI26 / 732A55

Time:	2018-03-12, 14-18
Teacher:	Magnus Borga, Phone: 013-286777
Allowed additional material:	Calculator, English dictionary

Read the instructions before answering the questions!

The exam consists of three parts:

Part 1 Consists of ten questions. The questions test general knowledge and understanding of central concepts in the course. The answers should be short and given on the blank space after each question. Any calculations do **not** have to be presented. Maximum one point per question.

Part 2 Consists of five questions. These questions can require a more detailed knowledge. Also here, the answers should be short and given on the blank space after each question. Only requested calculations have to be presented. Maximum two points per question.

Part 3 Consists of four questions. All assumptions and calculations made should be presented. Reasonable simplifications may be done in the calculations. **All calculations and answers on part 3 should be on separate papers! Do not answer more than one question on each paper!** Each question gives maximum five points.

The maximum sum of points is 40 and to ~~pass the exam (grade 3) normally 18 points are required. There is no requirement of a certain number of points in the different parts of the exam.~~ The answers may be given in English or Swedish. **Write clearly using block letters! (Do not use cursive writing.) Answers that are difficult to read, will be dismissed.**

The result will be reported at 2018-03-26 at the latest. The exams will then be available at IMT.

GOOD LUCK!

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 1

(Please use the space under each question for your answer in this part.)

1. Machine learning is often divided into the categories Supervised, Unsupervised and Reinforcement learning. Categorize the following three learning methods accordingly:

   - Q-leaerning
   - k-means
   - kNN (k nearest neighbors)

2. Mention a classifier that uses the *maximum margin* principle.

3. What is the basic requirement on an activation function for being used in a multilayer perceptron with back propagation?

4. You have a high-dimensional data set where many of the features are correlated. Suggest a proper pre-processing before feeding that data into a classifier.

5. Explain (briefly) the exploration-exploitation dilemma in reinforcement learning.

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

6. We have 10 samples in a 20-dimensional space that we want to analyse with a kernel method. How large is the kernel matrix?

7. What is being optimized by Linear Discriminant Analysis (LDA)?

8. Mention a method that can speed up gradient search.

9. Explain, or draw an example of, when a sigmoid function in the output layer of a linear classifier trained by minimizing the means sqaure error can improve the accuracy.

10. What is the benefit with a convolutional layer compared to a fully connected layer in a multi-layer neural network?

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 2

(Please use the space under each question for your answer in this part.)

11. K-means and Mixture of Gaussians are two examples of one general algorithm. What is the name of that algorithm and how does it differ from gradient-search methods?

12. When training a supervised learning system, the available data are usually divided in three groups that are used for different purposes. Name these groups and explain how they are used!

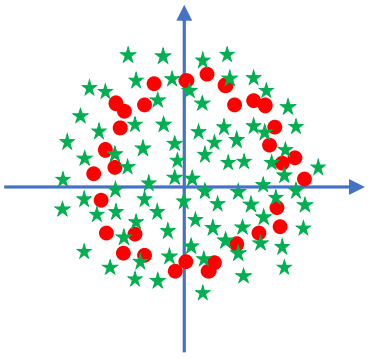
AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

13. The optimal plane separating two linearly separable classes is in a Support Vector Machine found by optimizing the cost function

$$\min \|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$$
$$\text{subject to the constraint } y_i\left(\mathbf{w}^T\mathbf{x}_i + w_0\right) \geq 1 \text{ for all } i,$$

where $\{\mathbf{x}_i, y_i\}$ are the training examples and $y_i = \pm 1$. Show how a *kernelized* version of the optimization problem is derived using the kernel trick!

14. Linus is going to build a classifier to diagnose a medical disease based on two measurements, x_1 and x_2. He has 32 patients with the disease and he has recruited 68 healthy controls. The data looks like the figure below in the two-dimensional feature space. Since he has not taken a course in neural networks, he is trying to solve the problem with a linear classifier. What is the expected accuracy (approximately) he will achieve after training the linear classifier? Draw an example of such a solution.



AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

15. The sigmoid function is a classical choice for a 3-layer network. For deep networks, convergence becomes extremely slow. Motivate why and suggest a solution.

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 3

(N.B. Write all answers in this part on separate sheets of papers! Don't answer more that one question on each sheet!)

16. The convolution of a 2D image $f(x, y)$ and a kernel $h(x, y)$ is defined as

$$g(x, y) = (f * h)(x, y) = \sum_{\alpha=-\infty}^{\infty} \sum_{\beta=-\infty}^{\infty} f(\alpha, \beta) h(x - \alpha, y - \beta).$$

a) Perform the convolution below, i.e. calculate the image C. All values outside the image array A are equal to zero. In the arrays A and B, the respective number written in bold face is at position $(x, y) = (0, 0)$. Note that C is only a part of the convolution result. (2p)

0	1	0	1		0	1	0		**?**	?	?	?
2	0	1	0	*	1	**2**	2	=	?	?	?	?
0	2	0	0		0	2	0		?	?	?	?

image A * kernel B = image C

b) In practice, when implementing convolution, e.g. like `convolve(A,B)`, no part of the kernel can be placed outside the image. Consequently, the resulting 2D array has size 1×2 and is equal to the central part of the image array C above. By extending the image array A to a new image array AA in a suitable way, `convolve(AA,B)` will be equal to C.
Give the image array AA. (1p)

c) A CNN consists of N complex layers. Each complex layer consists of a convolution with a 3×3 kernel and a mean pooling layer with stride 2 in each dimension. How large image in the input layes does a kernel on layer N cover? (2p)

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

17. You have the following data:

$$\mathbf{X} = \begin{bmatrix} -1 & 0 & 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}$$

where $\mathbf{X}$ contains six 2d-samples (one per column), and $\mathbf{Y}$ contains classification labels for the corresponding samples.

a) Perform the first AdaBoost iteration on the data $\mathbf{X}$ using the labels $\mathbf{Y}$. Sketch the classification problem. Use 'decision stumps' as weak classifiers. (2p)

b) Perform the second AdaBoost iteration on the data. Sketch the classification problem. Use 'decision stumps' as weak classifiers. (2p)

c) Classify the data using a strong classifier consisting of the two weak classifiers from a) and b). (1p)

Hint: The standard way of updating the weights in the standard AdaBoost method is $d_{t+1}(i) \propto d_t(i)e^{-\alpha_t y_i h_t(\mathbf{x})}$, where $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$.

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

18. You have the following training samples:

X	=	-1.6	-1.4	-1.2	-0.8	-0.4	0	0.3	0.7	0.9	1.1
D	=	-1	-1	-1	1	1	1	1	-1	-1	-1

Where D is the desired output for the samples in X.

a) Your task is to use a neural network setup in order to separate the data using Cover's theorem. You shall have one node in the output layer. Use *sign* as activation function, both in the hidden layer and the output layer. Sketch the network and assign W (weights in the hidden layer) and V (weights in the output layer) so that the output gives 100 % in accuracy (4p).

b) What would you have to initiate and/or change in order for a computer to solve the task above using backpropagation with gradient descent (1p)?

AID:	Exam Date: 2018-03-12
Course Code: TBMI26 / 732A55	Exam Code: TEN1

19. The figure shows two different deterministic state models and the corresponding rewards. The states are enumerated and the arrows represent actions. The numbers close to the arrows show the corresponding rewards. If the system reaches a state denoted "End" no additional rewards are given, i.e. the V-function is defined as 0 in such a state. An optimal policy is in this context the policy which maximizes the reward.

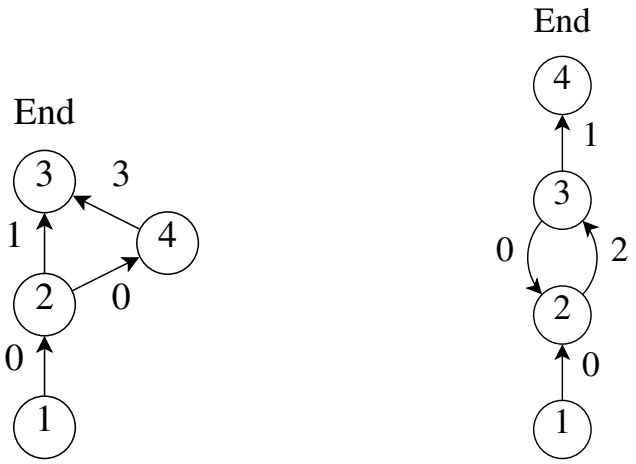


Figure 1: The state models A and B.

a) Calculate the optimal Q- and V-functions for system A as functions of $0 < \gamma < 1$. (2p)

b) Calculate the optimal Q- and V-functions for system B as functions of $0 < \gamma < 1$. (3p)