Exam in
Neural Networks and Learning Systems
TBMI26 / 732A55

Time:	2019-03-18, 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 2019-04-03 at the latest. The exams will then be available at "studerandeexpeditionen" at IMT.

GOOD LUCK!

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 1

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

1. Which of the following methods are supervised learning methods:

   - Mixture of Gaussians
   - Back-propagation
   - k-NN
   - LDA (Linear Discriminant Analysis)
   - SVN (Support Vector Machines)
   - PCA (Principal Component Analysis)

2. Explain the purpose with the so-called *slack variables* in Support Vector Machines.

3. Which of these functions can be used in the hidden layers of a back-prop network?

   - $y = s$
   - $y = \tanh(s)$
   - $y = \frac{s}{\|s\|}$
   - $y = e^{(-s^2)}$

4. What is described by the first eigenvalue of the data covariance matrix?

5. What problem can be illustrated by the multi-armed bandit?

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

6. Write the general definition of a kernel function $k(\mathbf{x}, \mathbf{y})$

7. "k-means clustering" and "mixture of gaussians" are two examples of a general optimisation method. What is that method called?

8. What do you get if you combine Bagging and decision trees?

9. What is described by the "empirical risk"?

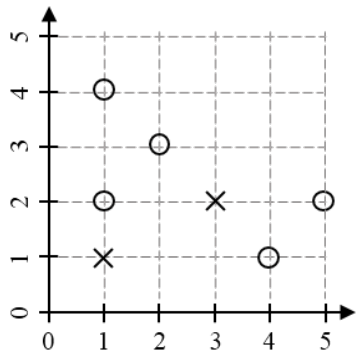
10. What is the purpose of the hidden layers in a multi-layer perceptron calssifier?

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 2

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

11. Perform one iteration of the k-means algorithm on the data plotted below. Circles (o) are input data and crosses (x) are the current prototype vectors (1p). How many more iterations must be made before the algorithm converges? Motivate your answer (1p)



12. Besides the training data, two other data sets are often used in machine learning. What are they called and what is the purpose of each of these datasets?

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

13. Show algebraically how principal component analysis can be "kernelized", i.e. reformulated as an eigenvalue problem on a kernel matrix.

14. In deep neural networks, the "ReLU" activation function is often used in the hidden layers. Draw that function and explain the advantage with this function compared to the classic sigmoid function in deep neural networks.

15. Draw a decision tree that implements the following discriminant function:

$$y = \begin{cases} 1, \text{ for } |x| < 1 \\ -1, \text{ for } |x| \geq 1 \end{cases}$$

AID:	Exam Date: 2019-03-18
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. To train a two-dimensional dataset with N samples we want to use a two layer neural network (one hidden and one output layer) according to figure 1, i.e., two hidden neurons and two output neurons.

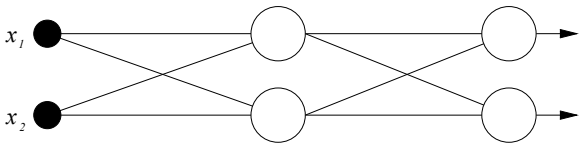


Figure 1: The network for this assignment.

We have two classes coded as $d_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $d_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ respectively. We want to use the activation function $\sigma(s) = \frac{1}{1+e^{-s}}$ for the hidden layer and a linear activation function for the output layer. We also want to train the network using gradient descent batch-learning and with the standard quadratic error:

$$\varepsilon = \frac{1}{N} \sum_{\mu=1}^{N} \sum_{i=1}^{C} (d_i^\mu - y_i^\mu)^2.$$

a) Draw a complete network where you define all signals, indices and weights that are missing in the figure and/or you want to use for your calculations below. (1p)

b) Derive the update rule for all weights in the complete network. Present your answer in component form, not matrix form. Do not forget to declare all indices. (4p)

Hint: The derivative of this sigmoid function is: $\frac{\delta\sigma}{\delta s} = \sigma(s)(1 - \sigma(s))$

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

17. The convolution of a 1D feature map channel $f_i(x)$ and a kernel $h_{ij}(x)$ is defined

$$g_{ij}(x) = (f_i * h_{ij})(x) = \sum_{\alpha=-\infty}^{\infty} f_i(\alpha) h_{ij}(x-\alpha).$$

We now look at the example of a two-channel input feature map ($i = 0, 1$) and a three-channel output feature map ($j = 0, 1, 2$), i.e., we use six kernels in total.

a) 2p Perform the six convolutions below. All values outside the feature map channels f_i are equal to zero. In the arrays f_i and h_{ij}, the respective number written in bold face is at position $x = 0$. Note that g_{ij} is only a part of the convolution result (*'same'*).

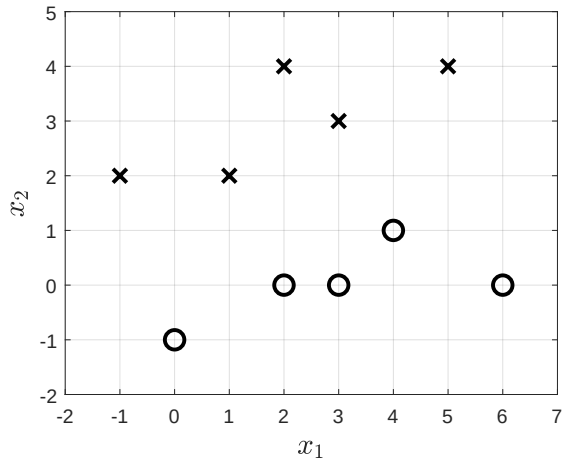
0	1	0	1	$*$	2	**2**	1	$=$	**?**	?	?	?	$i = 0, j = 0$
0	1	0	1	$*$	0	**1**	0	$=$	**?**	?	?	?	$i = 0, j = 1$
0	1	0	1	$*$	0	**2**	0	$=$	**?**	?	?	?	$i = 0, j = 2$
2	0	1	0	$*$	2	**2**	1	$=$	**?**	?	?	?	$i = 1, j = 0$
2	0	1	0	$*$	0	**2**	0	$=$	**?**	?	?	?	$i = 1, j = 1$
2	0	1	0	$*$	0	**1**	0	$=$	**?**	?	?	?	$i = 1, j = 2$
	f_i			$*$		h_{ij}		$=$		g_{ij}			

b) 2p In order to compute the output feature map channels d_j, we sum the g_{ij} over the index i, i.e., we compute $d_j = g_{0j} + g_{1j}$, $j = 0, 1, 2$.
How can d_0 be computed with only one convolution?
For any x, we can compute $(d_1(x), d_2(x))^T$ by a matrix multiplication. What is this matrix?

c) 1p We now assume I input channels, J output channels, and kernels of size K.
How many parameters need to be learned (no bias coefficient)?

AID:	Exam Date: 2019-03-18
Course Code: TBMI26 / 732A55	Exam Code: TEN1

18. The data points in the figure have two features (x_1 and x_2) and belong to either the class "crosses" or the class "circles":



Perform Linear Discriminant Analysis (LDA) on the data to reduce the dimensionality to one dimension that separates the two classes optimally. Draw the reduced data. (5p)

Hint: The inverse of a 2×2-matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $\frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

AID:	Exam Date: 2019-03-18
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.

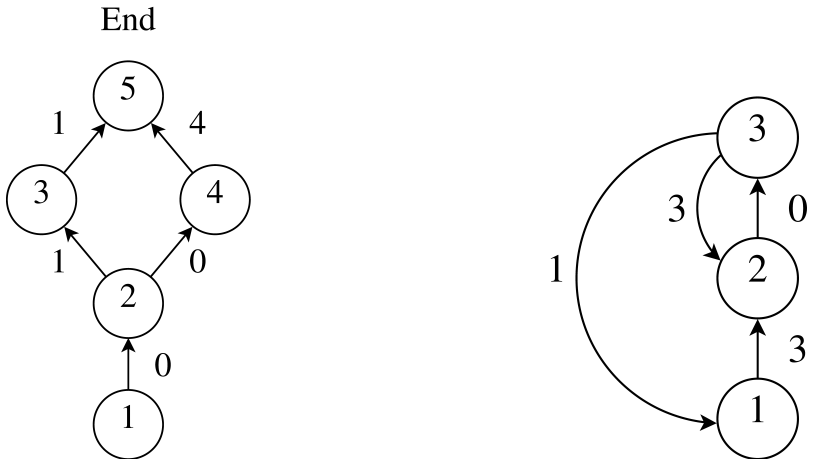


Figure 2: The state models A and B.

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)