Exam in
Neural Networks and Learning Systems
TBMI26 / 732A55

Time:	2022-08-27, 8-12
Teacher:	Magnus Borga, Phone: 013-28 67 77
Allowed additional material:	Calculator, English dictionary

Read the instructions before answering the questions!

The exam consists of two parts:

Part 1 Consists of 10 one-point and 5 two-point questions. The questions test general knowledge and basic understanding of central concepts in the course. The answers should be short and given on the blank space after each question or in the indicated figure.

Part 2 Consists of 4 five-point questions. These questions test deeper understanding and the ability to apply the knowledge to solve problems. All assumptions and calculations made should be presented. Reasonable simplifications may be done in the calculations. **All calculations and answers on part 2 should be on separate papers! Do not answer more than one question on each paper!**

The maximum sum of points is 20 on each part. To pass the exam (grade 3/C) at least 13 points are required on part 1. For grade 4/B, an additional 10 points on part 2 are required and for grade 5/A, 15 points are required on part 2, in addition to pass part 1.

The result will be reported at 2022-09-16 at the latest. The exams will then be available at studerandeexpeditionen at IMT.

GOOD LUCK!

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 1

One-point questions

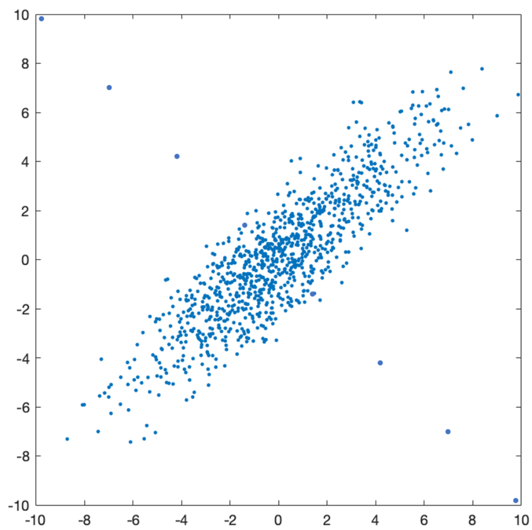
1. Machine learning is often divided into the categories Supervised, Unsupervised and Reinforcement learning. Categorize the following three learning methods accordingly:
   - Q-learning
   - Support vector machines
   - Principal component analysis

2. Describe or draw a situation when the so-called *slack* variables are required in Support Vector Machines.

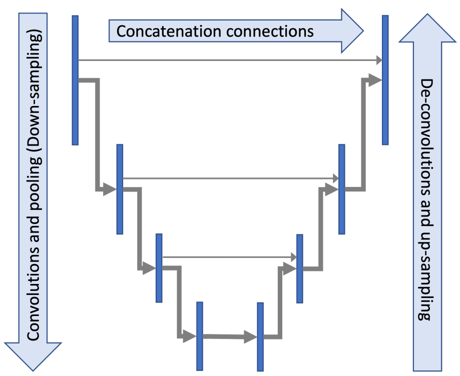
3. You have manually labeled training data for training an AdaBoost classifier. You happen to give one training example the wrong class. Why could this be a particular problem with the AdaBoost classifier?

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

4. Draw the first principal component of the distribution in the figure below!



5. What is the network architecture below called?



6. What does "k" stand for in "k-NN"?

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

7. What is the purpose of data augmentation?

8. How can you notice if a supervised machine learning algorithm has overtrained?

9. Assume we have a non-linear mapping $\varphi(\mathbf{x})$ to a high-dimensional feature space Φ and ϕ is a matrix whose columns i are the feature vectors φ_i in Φ . Define the kernel matrix $\mathbf{K}$ (whose components $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$) using the matrix ϕ.

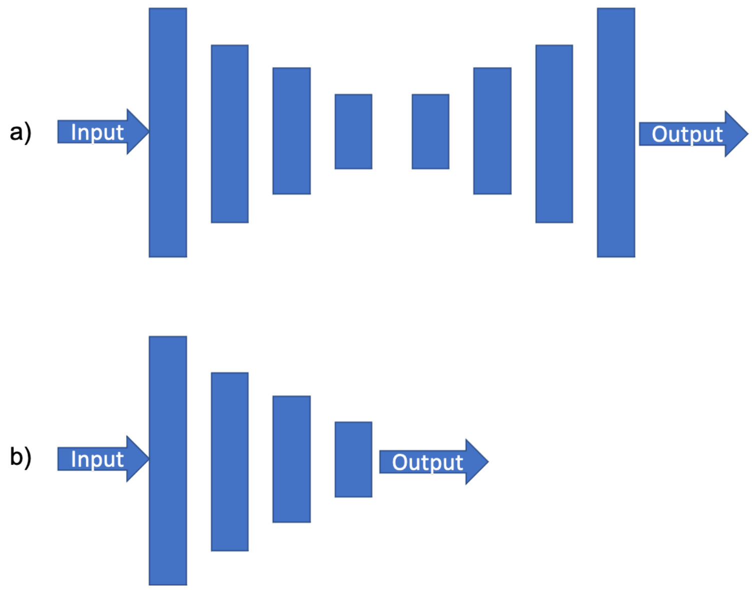
10. What is the difference between the "value function" (V) and the Q-function in Reinforcement learning?

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Two-point questions

11. 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.

12. Consider the two different network architectures below. The height of each box indicates the size of each layer. Which one could be used for an image classification task and which one could be used for image segmentation? Motivate your answer!



13. Consider a polynomial kernel function $\kappa(\mathbf{x}_1, \mathbf{x}_2) = 1 + (\mathbf{x}_1^T \mathbf{x}_2)^2$. What is the distance between two feature vectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

in the new feature space defined by this kernel function?

14. Write a 3-by-3 convolution kernel that detects horizontal lines in an image. The kernel should not be sensitive to the mean intensity of the image neighborhood.

15. You have 900 labeled training samples and want to evaluate how well different supervised classification algorithms perform for this data. Explain/sketch how you would do this with *3-fold cross-validation.*

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

Part 2

1. a) Draw a figure of the following network detailing the input, outputs and operation of each layer.
   Network description: two output neurons with linear activation functions and a hidden layer with two neurons using *tanh* as activation functions.
   For this network, use the figure to derive the weight update rules for online training with the square error as cost function. If needed you can assume that the input consists of two features. Both component form and matrix form are accepted. Make sure to explain all variables and indexes you introduce. (3p)

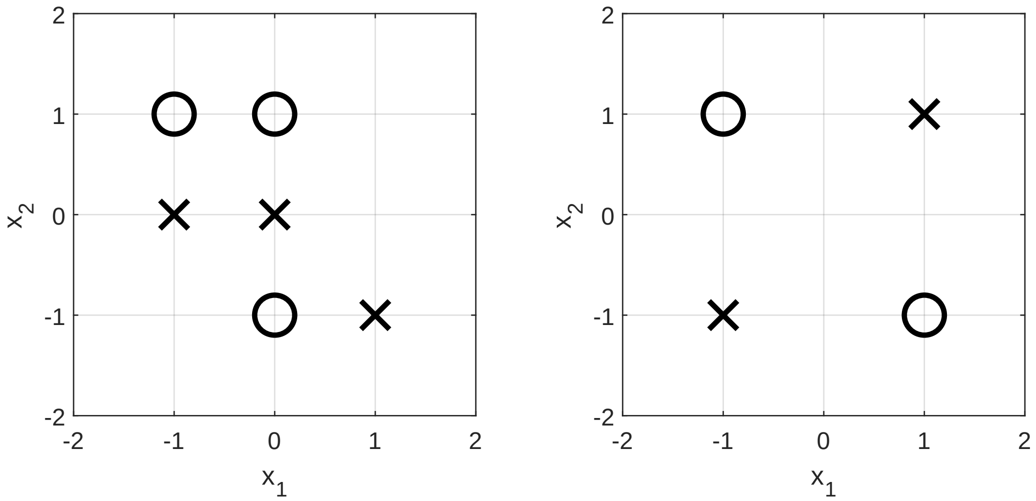
   b) Of what use is the activation function in the hidden layer? (1p)

   c) Explain in a few sentences what is the effects of adding the weight bias on the decision boundary learned by the model. (1p)

Hint: $\dfrac{\partial}{\partial x}\tanh(x) = 1 - \tanh^2(x)$

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

2. You have the following data:



(a) Dataset A (b) Dataset B

a) Perform the first two iterations of AdaBoost on Dataset A using brute force optimization and 'decision stumps' as weak classifiers. Calculate the weight update for both iterations. (4p)

b) Will AdaBoost with 'decision stump' classifiers be able to solve Dataset B? Motivate your answer. (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 $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.

Hint: Using exact values in the calculations (i.e. no approximations with calculator) will give very clean answers.

3. Assume we have the data

$$\mathbf{X} = \begin{pmatrix} -3 & -2 & 0 & 1 & 1 & 3 \\ 3 & -2 & 1 & 0 & 2 & 2 \end{pmatrix} \tag{1}$$

with six samples of two feature dimensions.

a) We think the data volume is too large. Show how the data volume can be reduced to one dimension using principal component analysis. State the resulting data volume after the dimensionality reduction. (3p)

b) How much of the information (total variance) in the data is accounted for by the first principle direction? (1p)

c) Show how the original signal can be approximately reconstructed given the results from a). (1p)

Hint: The eigenvalues of a 2×2-matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ are:

$$\lambda_1 = \frac{T}{2} + \sqrt{\frac{T^2}{4} - D} \text{ and } \lambda_2 = \frac{T}{2} - \sqrt{\frac{T^2}{4} - D},$$

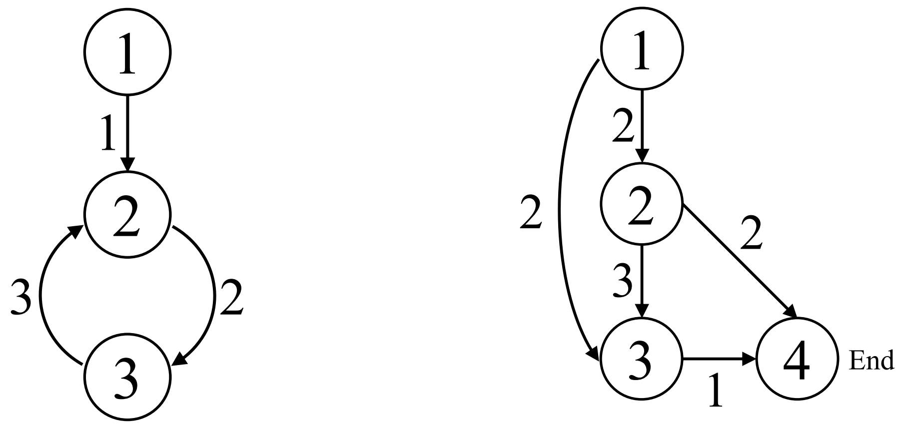
with trace $T = a + d$ and determinant $D = ad - bc$. The corresponding eigenvectors are

$$\mathbf{e}_1 = \begin{pmatrix} \lambda_1 - d \\ c \end{pmatrix} \text{ and } \mathbf{e}_2 = \begin{pmatrix} \lambda_2 - d \\ c \end{pmatrix}$$

Hint: Using exact values in the calculations (i.e. no approximations with calculator) will give very clean answers.

AID:	Exam Date: 2022-08-27
Course Code: TBMI26 / 732A55	Exam Code: TEN1

4. 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.



Figur 2: The state models A (left) and B (right).

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)