Machine Learning – February 12, 2019 - A

Time limit: **2 hours**.

Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam for ML 2018/19, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 1

1. Provide a formal definition of the Reinforcement Learning (RL) problem. Describe formally what are the inputs and the outputs of a RL algorithm.
2. Describe the main steps of a RL algorithm. Provide an abstract pseudo-code of a generic algorithm for RL (e.g., Q-learning).

EXERCISE 2

Describe two different methods to overcome overfitting in Convolutional Neural Networks (CNN).

EXERCISE 3

1. Describe the principle of maximal margin used by SVM classifiers. Illustrate the concept with a geometric example.
2. Draw a linearly separable dataset for binary classification of 2D samples. Draw two solutions (i.e., two separation lines): one corresponding to the maximum margin, the other one can be any other solution.
3. Discuss why the maximum margin solution is preferred for the classification problem.

EXERCISE 4

1. Provide the definition of *Confusion matrix* for a multi-class classification problem.
2. Provide a numerical example of a confusion matrix for a 3-classes classification problem with a balanced data set including 100 samples for each class. Show the confusion matrix in two formats: with absolute values and with the corresponding percentage values.
3. Compute the accuracy of the classifier for the numerical example provided above.

Hint: use simple numerical values, so that you do not need to make complex calculations.

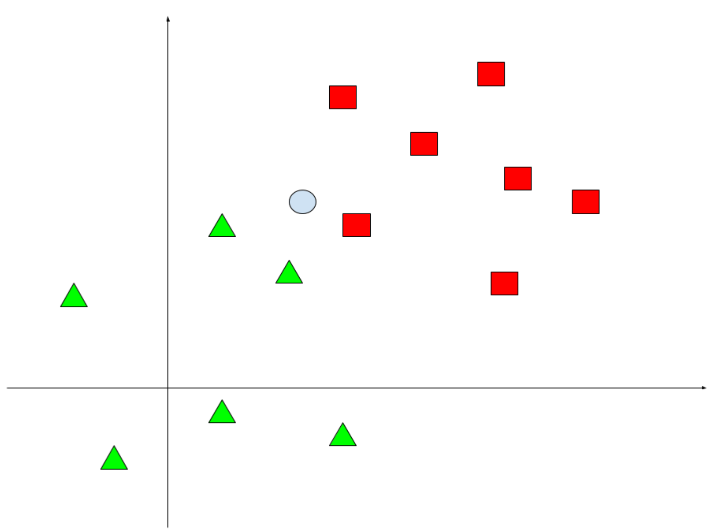
EXERCISE 5

Given an unsupervised dataset $D = \{\mathbf{x}_n\}$

1. Define the Gassian Mixture Model (GMM) and describe the parameters of the model.
2. Draw an example of a 2D data set (i.e., $D \subset \Re^2$) generated by a GMM with $K = 3$, qualitatively showing in the picture also the parameters of the model.
3. Determine the size of the model (i.e., number of independent parameters) for the dataset illustrated above.

EXERCISE 6

1. Describe the K-nearest neighbors (K-NN) algorithm for classification.
2. Given the dataset below for the two classes $\{square, triangle\}$, determine the answers of K-NN for the query point indicated with symbol o for K=1, K=3, and K=5. Motivate your answer, showing (with a graphical drawing) which instances contribute to the solution.



Machine Learning – February 12, 2019 - B

Time limit: **2 hours**.

Last Name First Name Matricola

....

Note: if you are not doing the exam for ML 2018/19, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 1

1. Describe the principle of maximal margin used by SVM classifiers. Illustrate the concept with a geometric example.
2. Draw a linearly separable dataset for binary classification of 2D samples. Draw two solutions (i.e., two separation lines): one corresponding to the maximum margin, the other one can be any other solution.
3. Discuss why the maximum margin solution is preferred for the classification problem.

EXERCISE 2

Given an unsupervised dataset $D = \{\mathbf{x}_n\}$

1. Define the Gassian Mixture Model (GMM) and describe the parameters of the model.
2. Draw an example of a 2D data set (i.e., $D \subset \Re^2$) generated by a GMM with $K = 3$, qualitatively showing in the picture also the parameters of the model.
3. Determine the size of the model (i.e., number of independent parameters) for the dataset illustrated above.

EXERCISE 3

1. Discuss the following statement: *"Accuracy is not always a good performance metric for classification"*.
2. Provide a numerical example to motivate your answer.

EXERCISE 4

Describe the typical stages of a single Convolutional layer in Convolutional Neural Networks (CNN).

EXERCISE 5

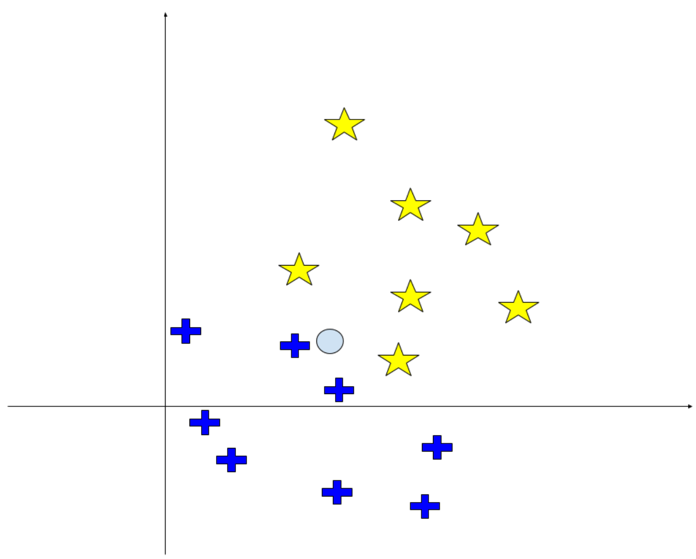
1. Describe the difference between exploitation and exploration in reinforcement learning and discuss why it is important to properly balance between the two strategies.

2. Describe how exploitation and exploration are generally implemented in RL algorithms. Illustrate this step with an abstract pseudo-code.

EXERCISE 6

1. Describe the K-nearest neighbors (K-NN) algorithm for classification.

2. Given the dataset below for the two classes $\{star, plus\}$, determine the answer of K-NN for the query point indicated with symbol o for K=1, K=3, and K=5. Motivate your answer, showing (with a graphical drawing) which instances contribute to the solution.



Machine Learning – January 18, 2019 - A

Time limit: **2 hours**.

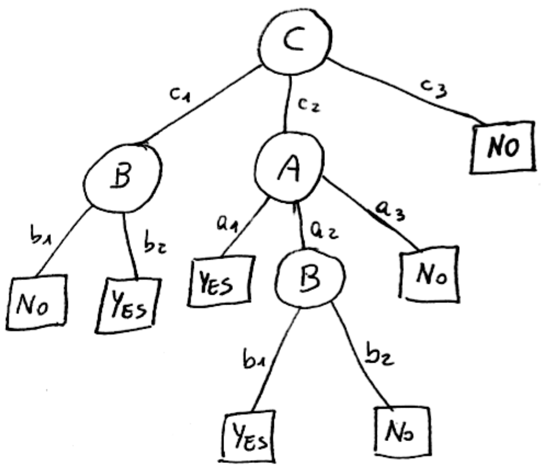
Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam for ML 2018/19, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 1

Given a classification problem for the function $f : A \times B \times C \to \{+, -\}$, with $A = \{a_1, a_2, a_3\}, B = \{b_1, b_2\}, C = \{c_1, c_2, c_3\}$ and the following decision tree T that is the result of a learning algorithm on a given data set:



1. Provide a rule based representation of the tree T.
2. Determine if the tree T is consistent with the following set of samples $S \equiv \{s_1 = \langle a_1, b_1, c_1, No\rangle, s_2 = \langle a_2, b_1, c_2, Yes\rangle, s_3 = \langle a_1, b_2, c_3, Yes\rangle, s_4 = \langle a_2, b_2, c_2, Yes\rangle\}$. Show all the passages needed to get to the answer.

EXERCISE 2

In Bayesian Learning, given a data set D and a hypothesis h, we can express the following relationship between the probability distributions (Bayes theorem):

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In this context:

1. define *Maximum a posteriori* (MAP) hypotheses and *Maximum likelihood* (ML) hypotheses.
2. define the concept of *Bayes Optimal Classifier*
3. discuss about practical applcability of the *Bayes Optimal Classifier*

EXERCISE 3

1. Briefly describe the goal of linear regression and define the corresponding model.
2. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ with $\mathbf{x}_n$ the input values and t_n the corresponding target values, explain how the parameters of the model can be estimated either in a batch or in a sequential mode.

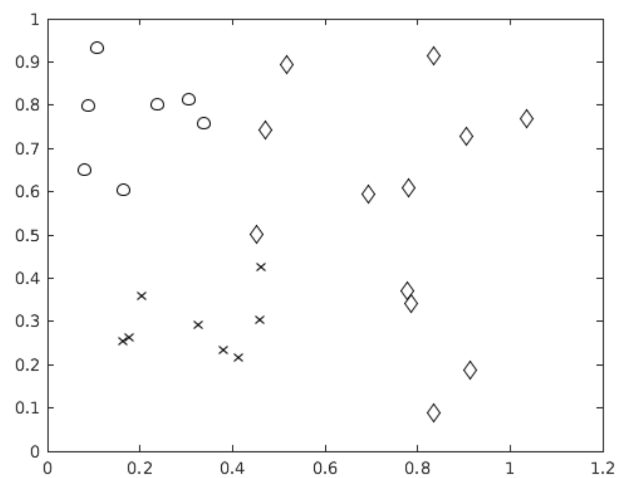
EXERCISE 4

Consider a regression problem for the target function $f : \Re^8 \to \Re^4$. Design a solution based on Artificial Neural Network for this problem: draw a layout of a suitable ANN for this problem and discuss the choices.

1. Determine the size of the ANN model (i.e., the number of unknown parameters).
2. Is Backpropagation algorithm affected by local minima? If so, how can we avoid or attenuate it?
3. Is Backpropagation algorithm affected by overfitting? If so, how can we avoid or attenuate it?

EXERCISE 5

Consider the data shown in the figure below:



Considering classification based on support vector machines (SVMs):

1. Explain if the data are separable and motivate your answer (only 'yes' or 'no' are not acceptable answers).
2. Explain what type of kernel function you would use in this case.
3. Describe what are the possible solutions for applying SVMs for classification of multiple classes.

EXERCISE 6

1. Describe the perceptron model for classification and its training rule.
2. Draw a graphical representation of a 2D data set for binary classification and provide a qualitative graphical example of a possible evolution of perceptron training (4 images showing a possible temporal evolution of the solution of the algorithm on the sketched data set).

Machine Learning – January 18, 2019 - B

Time limit: **2 hours**.

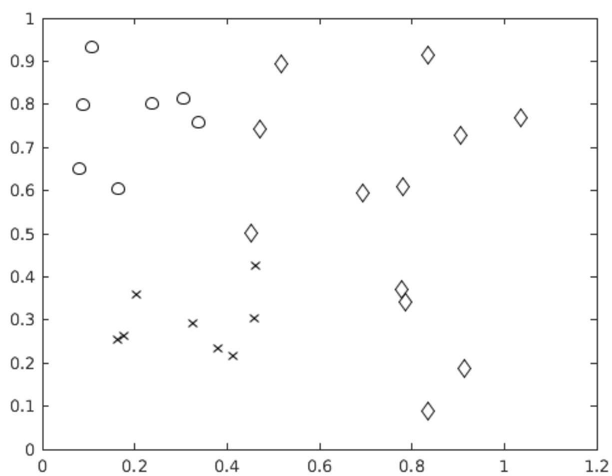
Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam for ML 2018/19, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 1

Consider the data shown in the figure below:



Considering classification based on support vector machines (SVMs):

1. Explain if the data are separable and motivate your answer (only 'yes' or 'no' are not acceptable answers).
2. Explain what type of kernel function you would use in this case.
3. Describe what are the possible solutions for applying SVMs for classification of multiple classes.

EXERCISE 2

1. Describe the perceptron model for classification and its training rule.
2. Draw a graphical representation of a 2D data set for binary classification and provide a qualitative graphical example of a possible evolution of perceptron training (4 images showing a possible temporal evolution of the solution of the algorithm on the sketched data set).

EXERCISE 3

Consider a regression problem for the target function $f : \Re^6 \to \Re^6$. Design a solution based on Artificial Neural Network for this problem: draw a layout of a suitable ANN for this problem and discuss the choices.

1. Determine the size of the ANN model (i.e., the number of unknown parameters).
2. Is Backpropagation algorithm affected by local minima? If so, how can we avoid or attenuate it?
3. Is Backpropagation algorithm affected by overfitting? If so, how can we avoid or attenuate it?

EXERCISE 4

1. Briefly describe the goal of linear regression and define the corresponding model.
2. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ with $\mathbf{x}_n$ the input values and t_n the corresponding target values, explain how the parameters of the model can be estimated either in a batch or in a sequential mode.

EXERCISE 5

In Bayesian Learning, given a data set D and a hypothesis h, we can express the following relationship between the probability distributions (Bayes theorem):

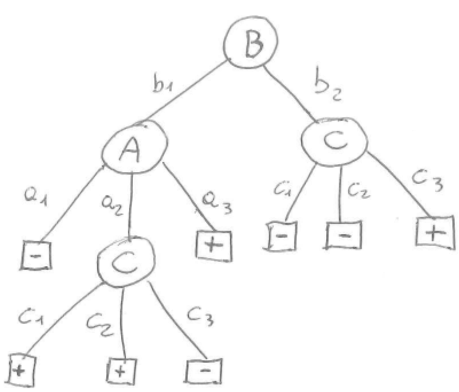
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In this context:

1. define *Maximum a posteriori* (MAP) hypotheses and *Maximum likelihood* (ML) hypotheses.
2. define the concept of *Naive Bayes Classifier*
3. discuss about practical applcability of the *Naive Bayes Classifier*

EXERCISE 6

Given a classification problem for the function $f : A \times B \times C \to \{+, -\}$, with $A = \{a_1, a_2, a_3\}, B = \{b_1, b_2\}, C = \{c_1, c_2, c_3\}$ and the following decision tree T that is the result of a learning algorithm on a given data set:



1. Provide a rule based representation of the tree T.
2. Determine if the tree T is consistent with the following set of samples $S \equiv \{s_1 = \langle a_1, b_1, c_1, -\rangle, s_2 = \langle a_2, b_1, c_2, +\rangle, s_3 = \langle a_1, b_2, c_3, +\rangle, s_4 = \langle a_2, b_2, c_2, +\rangle\}$. Show all the passages needed to get to the answer.

Machine Learning – February 12, 2018

Time limit: **2 hours**.

Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam ML 2017/18, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 0 (points $[0, 1]$ mulitplied to the overall score of the test)

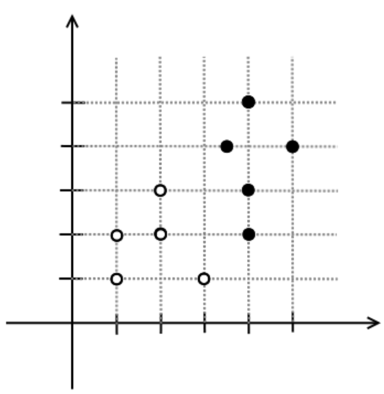
1. Write your name and matricola code in each paper you deliver.
2. Write all the answers of exercises **A** on one sheet marked as **A**, and all the answers of exercises **B** on another sheet marked as **B**. Do not mix answers of exercises **A** and **B** on the same sheet.
3. Do not use text books, slides, notes, mobile phone, laptop, etc.

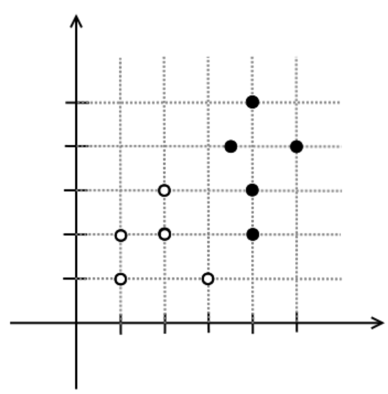
EXERCISE A1

Machine learning problems can be categorized in supervised and unsupervised. Explain the difference between them providing a precise formal definition (not only explanatory text) in terms of input and output of the two categories of problems.

EXERCISE A2

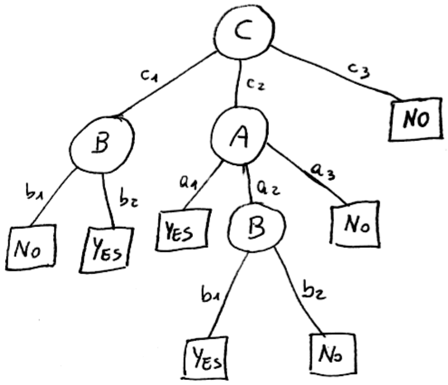
Consider the following data set for binary classification, where the two classes are represented with white and black circles. Draw in each of the diagrams a possible solution for a method based on Perceptron with very small learning rate and a method based on SVM. Describe the difference between the two solutions and explain how these are obtained with the two methods. Discuss which solution would you prefer and why.





EXERCISE A3

Given a classification problem for the function $f : A \times B \times C \rightarrow \{YES, NO\}$, with $A = \{a_1, a_2, a_3\}, B = \{b_1, b_2\}, C = \{c_1, c_2, c_3\}$ and the following decision tree T that is the result of training on a given data set:



1. Provide a rule based representation of the tree T.
2. Determine if the tree T is consistent with the following set of samples
   $S \equiv \{s_1 = \langle a_1, b_1, c_1, NO \rangle, s_2 = \langle a_2, b_1, c_2, YES \rangle,$
   $s_3 = \langle a_1, b_2, c_3, NO \rangle, s_4 = \langle a_3, b_2, c_1, YES \rangle\}.$
   Motivate your answer.

EXERCISE B1

1. Provide the main features about boosting.
2. Write the error function whose minimization leads to a formulation equivalent to the AdaBoost algorithm.

EXERCISE B2

Consider the problem of finding a function which describes how the salary of a person (in hundreds of euros) depens on his/her age (in years), the months in higher education and average grades in higher education. A dataset in the form $\mathcal{D} = \{(\mathbf{x}_1^T, t_1), \ldots, (\mathbf{x}_N^T, t_N)\}$ is provided, with $\mathbf{x} \in \mathbb{R}^3$ denoting the input values and t the target values (salary).

Assuming that one tries to identify this function with a deep feed-forward network:

1. Explain how the problem is formalized by writing the parametric form of the function to be learned highlighting the parameters $\boldsymbol{\theta}$.
2. Explain what are suitable choices for the activation functions of the hidden and output units of the network.
3. Explain what is a suitable choice for the loss function used for training the network and write the corresponding mathematical expression.

EXERCISE B3

Given input values $\mathbf{x}_i$ and the corresponding target values t_i with $i = 1, \ldots, N$, the solution of regularized linear regression can be written as:

$$y(\mathbf{x}) = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x},$$

with $\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{t}$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and λ the regularization weight.

Considering a kernel function $k(\mathbf{x}, \mathbf{x}')$:

1. Provide a definition of the Gram matrix.
2. Explain how a kernelized version for regression can be obtained based on the equations provided above.

Machine Learning – January 19, 2018

Time limit: **2 hours**.

Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam ML 2017/18, write below name of exam, CFU, and academic year (when you were supposed to attend the course)

....

EXERCISE 0 (points $[0, 1]$ mulitplied to the overall score of the test)

1. Write your name and matricola code in each paper you deliver.
2. Write all the answers of exercises **A** on one sheet marked as **A**, and all the answers of exercises **B** on another sheet marked as **B**. Do not mix answers of exercises **A** and **B** on the same sheet.
3. Do not use text books, slides, notes, mobile phone, laptop, etc.

EXERCISE A1

Consider a CNN with the following structure for its first two layers:

conv1 5×5 kernel and 64 feature maps with padding 2 and stride 1

relu1 acting on 'conv1'

pool1 2×2 max pooling with stride 2 acting on 'relu1'

conv2 3×3 kernel and 128 feature maps with padding 0 and stride 2

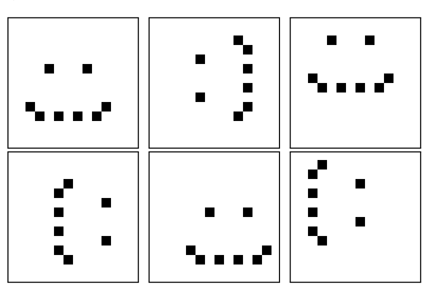
relu2 acting on 'conv2'

pool2 2×2 max pooling with stride 4 acting on 'relu2'

1. For input images of dimension $1242 \times 378 \times 3$ compute the dimensions of the volume on the output of each layer and explain how it is computed.
2. Describe what is the number of parameters of each layer.

EXERCISE A2

Consider the binary (black & white) images below defined on a 12×12 grid:



1. Explain what is the dimensionality of the data space and what is the intrinsic dimensionality of the given data.

2. Suppose you apply PCA on the data $\mathbf{x}_1, \ldots, \mathbf{x}_6$ and find that the data can be fully described using M principal components, namely $\mathbf{u}_1, \ldots, \mathbf{u}_M$. Describe how the original data can be written in the space defined by these M principal components.

3. Is M going to be equal to the number of intrinsic dimensions? Explain.

EXERCISE A3

Consider the following energy-like function defining Support Vector Machine regression:

$$J(\mathbf{w}, C) = C \sum_{i=1}^{N} L_\epsilon(t_i, y_i) + \frac{1}{2}\|\mathbf{w}\|^2,$$

with y_i, t_i target and predicted values, respectively, and $L_\epsilon(t,y) = \begin{cases} 0 & \text{if } |t-y| < \epsilon \\ |t-y| - \epsilon & otherwise \end{cases}$ the ϵ-insensitive error function.

1. Plot the ϵ-insensitive error function and explain what is the difficulty in minimizing J.

2. To overcome this difficulty slack variables ξ^+ and ξ^- are introduced. Explain (qualitatively) the role of the slack variables.

EXERCISE B1

Briefly describe a linear classification method and discuss its performance in presence of outliers. Use a graphical example to illustrate the concept.

EXERCISE B2

In Bayesian Learning, given a data set D and a hypothesis h, we can express the following relationship between the probability distributions (Bayes theorem):

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In this context:

1. define *Maximum a posteriori* (MAP) hypotheses and *Maximum likelihood* (ML) hypotheses.

2. define the concept of *Bayes Optimal Classifier*

3. discuss about practical applcability of the *Bayes Optimal Classifier*

EXERCISE B3

Describe the Markov property of Markovian models representing dynamic systems. Describe the difference between a Markov Decision Process (MDP) and a Hidden Markov Model (HMM). Draw and explain the graphical models of MDP and HMM.

Machine Learning – March 23, 2018

Time limit: **1.5 hours**.

Last Name	First Name	Matricola
....	....	....

For AIML students, choose at least 3 questions from the 6 below.

EXERCISE A1

1. Provide a formal definition of overfitting.
2. Discuss the problem of overfitting in learning with Decision Trees and illustrate possible solutions to it.

EXERCISE A2

1. Describe the *Naive Bayes Classifier* and highlight the approximation made with respect to the Bayes Optimal Classifier.
2. Provide design and implementation choices for solving the following problem through *Naive Bayes Classifier*:

   Classification of scientific papers in categories according to their main subject. The categories to be considered are: ML (Machine Learning), KR (Knowledge Representation), PL (Planning). Data available for each scientific paper are: title, authors, abstract and publication site (name of the journal and/or of the conference).

EXERCISE A3

1. Define with a precise formal definition the unsupervised learning problem.
2. Provide a full example of unsupervised learning problem (i.e., a specific invented data set), possibly in a graphical form.
3. Describe a solution to the defined problem based on K-Means, providing examples of execution of some steps of the algorithm and a reasonable solution.

EXERCISE B1

1. Provide the main steps of classification based on K-nearest neighbors (K-NN).
2. Draw an example in 2D demonstrating the application of the 3-NN algorithm for the classification of 3 points given a dataset consisting of points from 4 different classes.

Notes: You can choose how the points of the 4 classes are distributed. Use a different symbol for each class (e.g. use (*,x,+,-) for the classes and (o) for the points to be classified).

EXERCISE B2

1. Describe the role of the following notions related to parameter estimation of an artificial neural network:
   - backpropagation
   - forward and backward pass
   - Stochastic Gradient Descent
2. Provide the main steps of the backpropagation algorithm.

EXERCISE B3

1. Briefly describe the goal of linear regression and define the corresponding model.
2. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1^T, t_1)^T, \ldots, (\mathbf{x}_N^T, t_N)^T\}$ with $\mathbf{x}_n$ the input values and t_n the corresponding target values, explain how the parameters of the model can be estimated either in a batch or in a sequential mode.

Machine Learning 2017/2018

13/12/2017

Question 1. (5 points) The following data have been collected and we want to learn the general concept *Acceptable*, by using Decision Tree Learning.

House	Furniture	Nr rooms	New kitchen	Acceptable
1	No	3	Yes	Yes
2	Yes	3	No	No
3	No	4	No	Yes
4	No	3	No	No
5	Yes	4	No	Yes

1. Formalize the learning problem.
2. Describe how variables are chosen when building a Decision Tree.
3. Execute the ID3 algorithm on the data set above and generate the corresponding output tree.

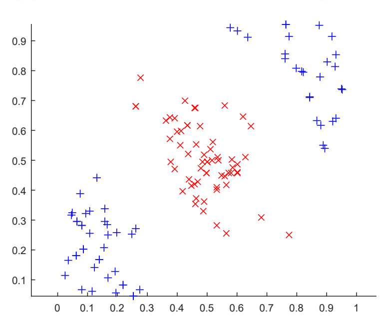
Note: point 3 can be answered even if point 2 is not properly addressed, by using any invented method (or invented numbers) for the selection of the variables.

Question 2. (5 points)

1. What is a maximum likelihood (ML) hypothesis?
2. Comment the following statement: in a classification problem, the class returned by the ML hypothesis on a new instance x is always the most probable class.

Question 3. (5 points) Consider classification of the data shown in the figure below by using SVMs.

1. Explain which kernel function you would choose to obtain perfect separability.
2. How SVMs classification can be applied if the data are note perfectly separable?



Question 4. (5 points) Let $x \in \mathcal{X} = \{(1.8, 2.1)^T, (3.2, 1.4)^T, (2.5, 4.1)^T, (1.6, 7.7)^T, (3.1, 9.1)^T\}$ and $t \in T = \{0, 1, 1, 0, 1\}$. Consider designing an artificial neural network for learning the function $t = f(x, \theta)$.

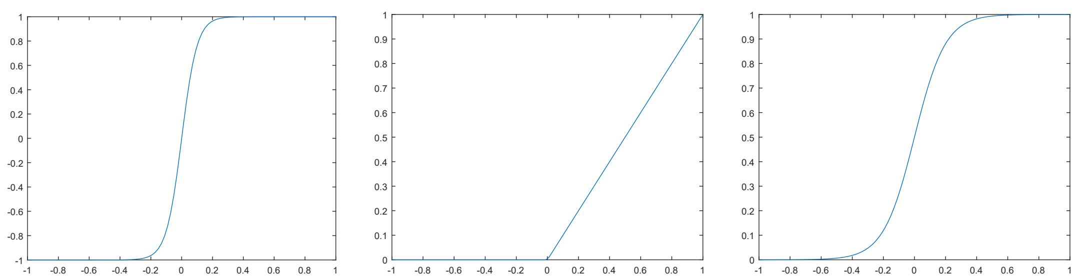
1. Explain what is a suitable activation function for the output layer of the network.

2. For the selected activation function explain if the output units saturate and how learning the parameters of the network is affected if this is the case.

Question 5. (5 points)

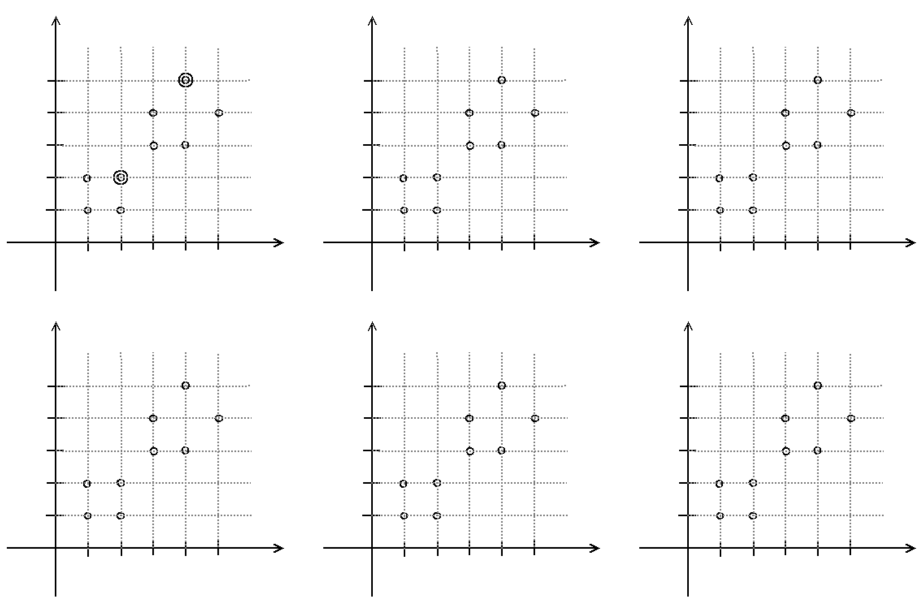
1. Consider an image of $w_{im} \times h_{im}$ elements (pixels) and a convolution kernel of dimension $3 \times 3 \times 16$. What are the possible values of the stride and padding in order to convolve without skipping any pixels, while obtaining a feature map with the same dimensions with the input image.

2. Associate the correct name of the activation function to the plots above and provide the corresponding mathematical definition. The list of names is $\{$*ReLU, Sigmoid, Hyperbolic Tangent*$\}$.



Question 6. (5 points)

Simulate the execution of K-means in this 2-D data set with k=2 and initial centroids indicated by double circles: use one diagram for each step of the algorithm. Describe explicitly how each step is obtained and what is the termination condition of the algorithm. Drawing only the steps is not sufficient.



Machine Learning – Test - Part 1

Time limit: **2 hours**.

Last Name	First Name	Matricola
....	....	....

EXERCISE 1

The following data have been collected and we want to learn the general concept *Acceptable*, by using Decision Tree Learning.

House	Furniture	Nr rooms	New kitchen	Acceptable
1	No	3	Yes	Yes
2	Yes	3	No	No
3	No	4	No	Yes
4	No	3	No	No
5	Yes	4	No	Yes

1. Formalize the learning problem: decribe exactly the target function to learn and the dataset.
2. Describe qualitatively how attributes are chosen when building a Decision Tree.
3. Simulate the execution of ID3 algorithm on the data set above and generate the corresponding output tree.

Note: point 3 can be answered even if point 2 is not properly addressed, by using any invented method (or invented numbers) for the selection of the variables.

EXERCISE 2

1. Provide a formal definition of a maximum likelihood (ML) hypothesis
2. Comment the following statement: *in a classification problem, the class returned by the ML hypothesis on a new instance x is always the most probable class.*

EXERCISE 3

Briefly describe a linear classification method and discuss its performance in presence of outliers. Use a graphical example to illustrate the concept.

EXERCISE 4

Given input values $\mathbf{x}_i$ and the corresponding target values t_i with $i = 1, \ldots, N$, the solution of regularized linear regression can be written as:

$$y(\mathbf{x}) = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x},$$

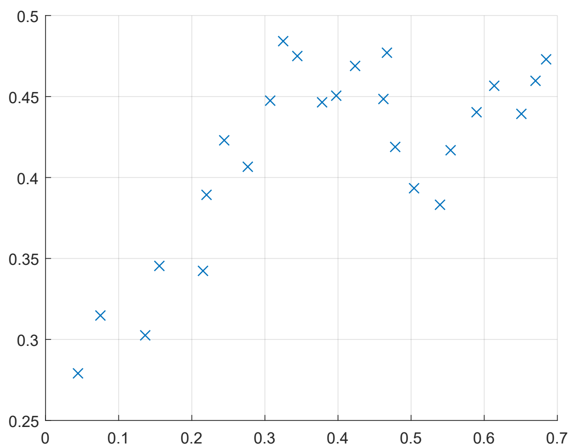
with $\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{t}$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and λ the regularization weight.

Considering a kernel function $k(\mathbf{x}, \mathbf{x}')$:

1. Provide a definition of the Gram matrix.
2. Explain how a kernelized version for regression can be obtained based on the equations provided above.

EXERCISE 5

Consider the learning problem of estimating the function $f : \Re \mapsto \Re$ with dataset $D = \{(x_i, y_i)\}$ plotted in the figure below:



1. Describe how to perform regression based on these data using a method of your choice. Specifically, provide a mathematical formulation of the model, highlighting the model parameters.
2. Considering the method you have chosen describe a way to reduce overfitting.
3. Draw a plausible plot of the learned model based on your choices.

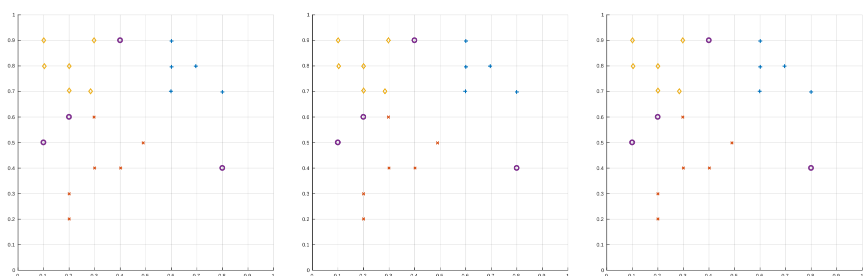
EXERCISE 6

1. Provide the main steps of classification based on K-nearest neighbors (K-NN).
2. Draw an example for a 4-classes classification problem in 2D. Use symbols (*,x,+,-) for the four classes. Graphically show the application of the K-NN algorithm with $K = 3$ for the classification of 3 different query points.

Machine Learning 2017/2018

Example 2

Question 1. (5 points) Graphically show how classification of the unknown instances, depicted as circles, is performed based on the K-nearest neighbors algorithm for $K = 1$, 3 and 5. Textually describe how K-NN works, just designing the result is not sufficient.



Question 2. (5 points) Briefly describe the main parts that form a single convolutional layer and their function.

Question 3. (5 points) Describe how Principle Components are identified based on the principle of variance maximization.

Question 4. (5 points)

A car driver in Rome has to move from one side of the Tiber river to the other every morning. There are three possible alternative paths passing to three different bridges and the paths are known. The driver wants to minimize the time to reach the target location, and due to traffic conditions, it is not guaranteed that the shortest path is also the quickest way. Moreover, traffic conditions are unpredictable and non-deterministic, but stationary.

1. Describe a complete model for this problem based on MDP, specifying all its elements.

2. Describe how to solve the problem based on Reinforcement Learning and determine the exact training rule to use to learn the best behavior.
3. Discuss the strategy for balancing exploration and exploitation in this case.

Question 5. (5 points)

Confusion matrices are used to summarize the result of a classification algorithm.

1. Provide a formal definition of a confusion matrix.
2. Provide an example (with invented numbers) of a confusion matrix for a classification problem for the target function $f : \{0, 1\}^4 \rightarrow \{Low, Medium, High\}$ and a data set containing 150 samples (of which 70 for Low, 30 for Medium and 50 for High).
3. Describe how to compute classification accuracy, given a confusion matrix.

Question 6. (5 points)

Consider a data set D for scoring different schools with the following real-valued attributes: staff salaries per pupil (x_1), teacher's test score (x_2), parents' education (x_3), school grade (y).

For this problem, an expert of the domain proposes to use the following model.

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1 x_2 + \theta_5 x_3^2$$

1. Define an error function for this model.
2. Discuss if a linear model for regression can be used in this case.
3. Describe an iterative approach to solve the problem.

Machine Learning – Test - December 20, 2018

Time limit: **2 hours**.

Last Name	First Name	Matricola
....	....	....

Note: if you are not doing the exam for ML 2018/19, write below name of exam, CFU, and academic year (when you were supposed to attend the course). Please specify also if you are an Erasmus student.

....

EXERCISE 1

Given input values $\mathbf{x}_i$ and the corresponding target values t_i with $i = 1, \ldots, N$, the solution of regularized linear regression can be written as:

$$y(\mathbf{x}) = \sum_i^N \alpha_i \mathbf{x}_i^T \mathbf{x},$$

with $\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{t}$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and λ the regularization weight.

Considering a kernel function $k(\mathbf{x}, \mathbf{x}')$:

1. Provide a definition of the Gram matrix.
2. Explain how a kernelized version for regression can be obtained based on the equations provided above.

EXERCISE 2

Consider a two-layers ANN which receives in input vectors $\mathbf{x}$ of dimension 128 and produces output vectors $\mathbf{y}$ of dimension 10. The hidden layer of the ANN is composed of 50 units which use the ReLU activation function. The output units use a linear activation function.

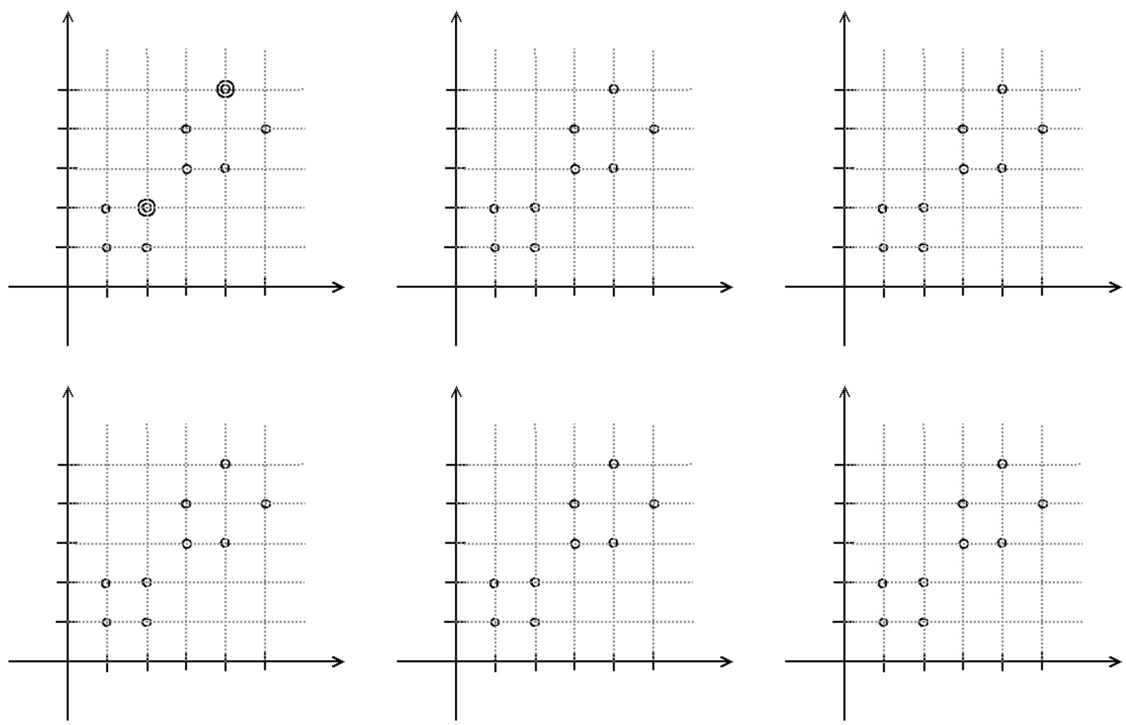
- The weight matrices of the hidden and output layers are denoted W_1 and W_2. Provide the dimensions of the weight matrices W_1 and W_2
- Provide the formula explicitly stating how the values of $\mathbf{y}$ are computed given an input vector $\mathbf{x}$ in terms of the weight matrices and the activation functions (you can ignore the bias terms).

EXERCISE 3

Briefly describe what is the architecture of an autoencoder and its purpose.

EXERCISE 4

Simulate the execution of K-means in this 2-D data set with k=2 and initial centroids indicated by double circles: use one diagram for each step of the algorithm. Describe explicitly how each step is obtained and what is the termination condition of the algorithm. Drawing only the steps is not sufficient.



EXERCISE 5

1. Provide the main steps of classification based on K-nearest neighbors (K-NN).

2. Draw an example in 2D demonstrating the application of the 3-NN algorithm for the classification of 3 points given a dataset consisting of points from 4 different classes.

Notes: You can choose how the points of the 4 classes are distributed. Use a different symbol for each class (e.g. use (*,x,+,-) for the classes and (o) for the points to be classified).

EXERCISE 6

Assume you have 4 image classifiers with medium-good classification accuracy.

1. Describe an ensemble method for achieving higher classification accuracy by combining such classifiers.

2. Are there any specific properties that each classifier has to have to achieve higher accuracy? If the answer is positive, explain which these properties are.

Machine Learning – June 23, 2020

Time limit: **2 hours 15 minutes**.

EXERCISE 1

Consider a setting where the input space I is the set of finite strings over the characters $a, b, c, \cdots, z$. Notice that input strings can be of different length.

Given the following dataset D:

x	**t**
a	1
ab	1
$caza$	4
$ayka$	4
$aabba$	9
aaa	9
$zazaa$	9
$accaaca$	16
$khaaala$	16
$akdfkkatyuakka$	16
$jhxaaksrtkaatyuap$	25

1. Identify the learning problem at hand, in particular the form of the target function, and define a suitable linear model for it.
2. Apply the kernel trick to the model defined above and provide the analytical form of the corresponding error function.
3. Define the solution obtained with your choices for the dataset D.

EXERCISE 2

A secret string $s = b_0b_1b_2b_3$ of 4 bits fulfills the following constraints:

- if $b_0 = 0$, s contains an even number of 0's and 1's;
- if $b_0 = 1$, s contains at least three 1's.

No other prior information about s is available.

1. Define the prior probability distribution $P(s)$ of the hypothesis string s.
2. Assuming $b_0 = 0$, define the conditional probability distribution $P(s|b_0 = 0)$ and indicate all maximum a-posteriori hypotheses.
3. Assuming $b_0 = 1$ and $b_1 = 1$, indicate all maximum likelihood hypotheses and compute the likelihood that $b_2 = 1$.

EXERCISE 3

Consider a dataset D for the binary classification problem $f : \Re^3 \mapsto \{A, B\}$.

1. Describe a probabilistic generative model for such a classification problem, assuming Gaussian distributions.
2. Identify the parameters of the model and determine the size of the model (i.e., the number of independent parameters).

EXERCISE 4

1. Describe the k-armed bandit problem (also known as One-state MDP).
2. Describe the Reinforcement Learning procedure to compute the optimal policy in the k-armed bandit problem with stochastic behavior and unknown functions.

EXERCISE 5

Consider that the output of layer l of a CNN is the set of feature maps M with size $256 \times 256 \times 64$

1. What is the size of the feature maps N when max-pooling with a 2×2 kernel and stride 2 is applied on M?
2. Design a convolutional layer which, when applied on M, produces feature maps with the same size as N. Describe all the relevant parameters of the layer you have designed.
3. What happens if the non-linear activation functions of the hidden layers of the CNN are replaced with linear functions? Is the effective depth of the network affected and how?

EXERCISE 6

Consider N convolutional neural networks trained to classify images of cats and dogs with a corresponding confidence value (output of sigmoid activation function)

- Describe a way to combine the predictions of the CNNs in order to get a single more accurate prediction.
- Assume $N = 3$, class '0' represents dogs, class '1' cats and for a given image the three CNN outputs are: $(0.912, 0.432, 0.444)$. Apply the method described above to classify the image using the predictions of the three CNNs.