Machine Learning – A – February 11, 2020

Time limit: **2 hours**.

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

Note: if you are not doing the regular exam for ML 2019/20, 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.

.....

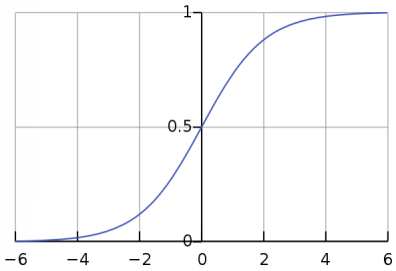
EXERCISE A1

1. Explain the difference between regression and classification.
2. Provide a mathematical formulation of linear regression.
3. Provide an example of a linear regression model that overfits a dataset of your choice, and discuss how this can be mitigated.

EXERCISE A2

1. Define mathematically the problem solved by logistic regression
2. Consider the following dataset and the sigmoid function:

x_1	x_2	x_3	t
0	0	1	1
1	2	3	1
4	4	1	0



Which one among the following solutions fits the data better? Why?

$$\vec{w}_1^T = (1, 0, -1) \qquad \vec{w}_2^T = (-1, -1, 2)$$

A plot of the sigmoid function is reported above. You do not need to compute explicit values of the model.

EXERCISE B1

1. Give a short explanation of the *kernel trick/kernel substitution.* What is the necessary condition for applying the kernel trick?
2. Provide an example of its application. In detail:
   - draw a suitable dataset for binary classification in 2D;
   - discuss which kernel you would use for this dataset;
   - show graphically a possible solution of such a kernel-based model.

EXERCISE B2

Consider the structure of a recurrent neural network (RNN):

1. Design a generic RNN model (or give the relative formula).
2. Explain the concept of 'unfolding' (or 'unrolling') an RNN.
3. For what type of input would you use an RNN? Describe a specific use case of your choice providing details both for the input and output of the RNN.

EXERCISE C1

1. Describe the difference between supervised learning and reinforcement learning with a formal definition of the two problems.
2. Describe the full observability property of Markov Decision Processes and its relation with non-deterministic outcomes of actions.

EXERCISE C2

1. Describe the K-means algorithm in a formal way (i.e., with precise mathematical formulas and equations), including: input and output of the algorithm, its main steps, and the termination condition.
2. Draw a suitable 2-D data set for K-means.
3. Simulate the execution of K-means in such 2-D data, showing at least three steps of the algorithm and the final output.

Machine Learning – A – January 20, 2020

Time limit: **2 hours**.

Last Name First Name Matricola

....

Note: if you are not doing the regular exam for ML 2019/20, 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 A1

Assume the following data about an online shop have been collected:

- Customers are: 25% young men (class YM); 45% young women (YW); 30% neither of the above (O).
- Young men buy: Shoes 30%; Trousers 50%; Shirts 20%.
- Young women buy: Shoes 50%; Trousers 30%; Shirts 20%.
- Other customers buy: Shoes 30%; Trousers 30%; Shirts 40%.

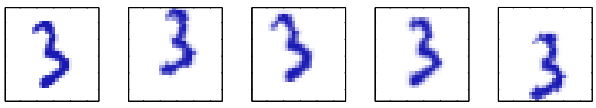
1. If you receive an order for trousers, which is the most probable class the customer who issued the order belongs to? Why?
2. Which is, and how do you compute, the likelihood that an order is for trousers?

EXERCISE A2

1. Explain when a dataset is *linearly separable*
2. Draw an example of a linearly separable dataset in a 2D setting, with two classes $C = \{+, -\}$
3. Draw an example of a non linearly separable dataset in a 2D setting, with two classes $C = \{+, -\}$
4. For each dataset shown above, draw a possible solution based on SVM and explain how it can be obtained.

EXERCISE B1

Consider the following data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ where the intrinsic dimensions are described in terms of a 2D translation and rotation (3 parameters) and the set of principal components $\mathbf{u}_1, \ldots, \mathbf{u}_M$ recovered from this data.



- How can these points be expressed in the basis defined by the principal components? Provide the relative formula.
- Is PCA able to recover a 3 dimensional space that fully describes the data (apart from noise)? Explain your answer.

EXERCISE B2

Consider the following Convolutional Neural Network acting on images of dimension $32 \times 32 \times 3$:

conv1	5×5 kernel and 16 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 32 feature maps with padding 0 and stride 1 acting on 'pool1'
relu2	acting on 'conv2'
pool2	2×2 max pooling with stride 2 acting on 'relu2'
conv3	5×3 kernel and 64 feature maps with padding 0 and stride 2 acting on 'pool2'
relu3	acting on 'conv3'
fc1	with 200 units acting on (flattened) 'relu3'
fc2	with 10 units acting on 'fc1'
output	softmax acting on 'fc2'

1. Compute the number of parameters for each layer of the network.
2. What is a suitable loss function to train the network defined above?

EXERCISE C1

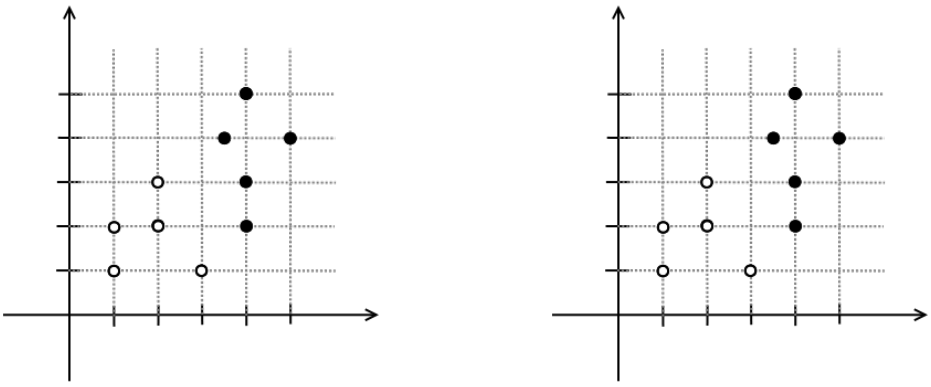
Consider the dataset $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ where each tuple $(\mathbf{x}_n, t_n)$ corresponds to an input value $\mathbf{x}_i \in \mathbb{R}^3$ and the corresponding target value $t_i \in \mathbb{R}$.

1. Provide the definition of a linear regression model (in its most general form) with parameters $\mathbf{w}$ that can be used for estimating a non-linear function y such that $t \approx y(\mathbf{x}, \mathbf{w})$.
2. Provide a suitable loss function and sketch an algorithm for estimating the parameters of the model.

EXERCISE C2

Consider the following data set for binary classification (white vs black circles).

1. Draw in each of the diagrams below a possible solution for a method based on Perceptron with very small learning rate and a possible solution for a method based on SVM.
2. Describe the difference between the two solutions and briefly explain how these are obtained with the two methods.
3. Discuss which solution would you prefer and why.



Machine Learning – B – February 11, 2020

Time limit: **2 hours**.

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

Note: if you are not doing the regular exam for ML 2019/20, 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 A1

Assume you are given the following dataset, representing the samples of a function f:

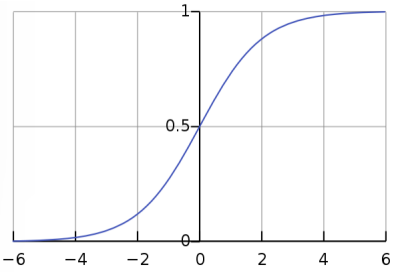
x_1	x_2	x_3	f
0.6	3	1	4.6
1	2	3	2.1
4	4	1	10

1. Which technique would you use to estimate f?
2. Provide a mathematical formulation of the problem solved by the chosen technique.
3. Provide an example of solution using a simple dataset of your choice. (Show the solution only, you don't have to illustrate the steps followed to obtain it).

EXERCISE A2

1. Define mathematically the problem solved by logistic regression
2. Consider the following dataset and the sigmoid function:

x_1	x_2	x_3	t
0	0	1	0
1	2	3	0
4	4	1	1



Which one among the following solutions fits the data better? Why?

$$\vec{w}_1^T = (2, 0, -2) \qquad \vec{w}_2^T = (-2, -2, 4)$$

A plot of the sigmoid function is reported above. You do not need to compute explicit values of the model.

EXERCISE B1

1. Explain what properties a kernel function should typically satisfy.
2. Indicate which of the following kernel functions are not valid explaining why:
   (a) $k(\mathbf{x}, \mathbf{x}') = 1$;
   (b) $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + \gamma)^4$;
   (c) $k(\mathbf{x}, \mathbf{x}') = \sum_i [\sin(\mathbf{x}_i) - \sin(\mathbf{x}'_i)]$;
   (d) $k(\mathbf{x}, \mathbf{x}') = \sum_i -\log(\mathbf{x}_i) \log\left(\frac{\mathbf{x}'_i}{\mathbf{x}_i}\right)$, with $\mathbf{x}_i, \mathbf{x}'_i > 0$ for all i;
   (e) $k(\mathbf{x}, \mathbf{x}') = 1 - \frac{|\mathbf{x}^T\mathbf{x}'|}{\|\mathbf{x}\|\|\mathbf{x}'\|}$;

EXERCISE B2

Consider the problem of finding a function which describes how the salary of a person (in hundreds of euros) depends 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 $t \in \mathbb{R}$ the target values (salary). Assuming that one tries to estimate 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.
4. Assuming that the gradients of the loss with respect to the parameters are available, describe an algorithm for training the parameters of the network. What are the hyper-parameters of the training algorithm (if any)?

EXERCISE C1

1. Describe the Markov Decision Process (MDP) model used in reinforcement learning, provide its mathematical formulation, and explain the elements of the model.
2. Describe the Q-learning algorithm, referring to the mathematical formulation of the MDP given above.

EXERCISE C2

1. Describe the K-means algorithm in a formal way (i.e., with precise mathematical formulas and equations), including: input and output of the algorithm, its main steps, and the termination condition.
2. Draw a suitable 2-D data set for K-means.
3. Simulate the execution of K-means in such 2-D data, showing at least three steps of the algorithm and the final output.

Machine Learning – B – January 20, 2020

Time limit: **2 hours**.

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

Note: if you are not doing the regular exam for ML 2019/20, 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 A1

Assume the following data about an online shop have been collected:

- Customers are: 45% young men (class YM); 30% young women (YW); 25% neither of the above (O).
- Young men buy: Shoes 20%; Trousers 30%; Shirts 50%.
- Young women buy: Shoes 20%; Trousers 50%; Shirts 30%.
- Other customers buy: Shoes 30%; Trousers 30%; Shirts 40%.

1. If you receive an order for shoes, which is the most probable class the customer who issued the order belongs to? Why?
2. Which is, and how do you compute, the likelihood that an order is for shoes?

EXERCISE A2

1. Explain when a dataset is *linearly separable*
2. Illustrate the error function minimized by the Least Squares method
3. Show an example, in a 2D dataset for binary classification, of application of Least Squares
4. Draw a 2D dataset for binary classification, describe a problem Least Squares suffers from and discuss one plausible approach to solve it.

EXERCISE B1

Consider the set of principal components $\mathbf{u}_1, \ldots, \mathbf{u}_D$ recovered from the (mean subtracted) data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and the variance of this data along each component $\lambda_1, \ldots, \lambda_D$.

- Give the name of an algorithm that can be used to obtain the principal components and the corresponding variances.
- Quantify the exact approximation error when only the first $M < D$ principal components are used for describing the data.
- Provide the formula describing how the data points are expressed in the basis defined by the first M principal components.

EXERCISE B2

Consider the following Convolutional Neural Network acting on images of dimension $56 \times 56 \times 3$:

conv1	7×7 kernel and 16 feature maps with padding 3 and stride 1
relu1	acting on 'conv1'
pool1	2×2 max pooling with stride 2 acting on 'relu1'
conv2	5×5 kernel and 32 feature maps with padding 2 and stride 3
relu2	acting on 'conv2'
pool2	2×2 max pooling with stride 2 acting on 'relu2'
conv3	1×1 kernel and 32 feature maps with padding 0 and stride 1
relu3	acting on 'conv3'
fc1	with 100 units acting on (flattened) 'relu3'
relu4	acting on 'fc1'
fc2	with 50 units acting on 'relu4'
relu5	acting on 'fc2'
fc3	with 2 units acting on 'relu5'
output	identity ('fc3')

1. Compute the number of parameters for each layer of the network.
2. What is a suitable loss function to train the network defined above?

EXERCISE C1

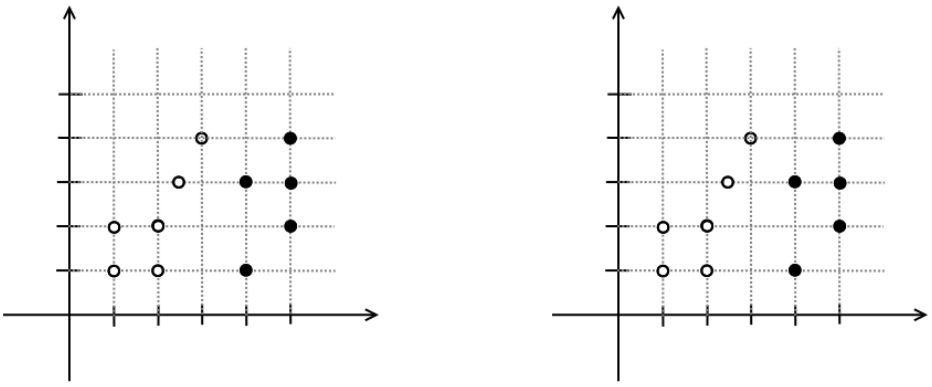
Consider the dataset $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ where each tuple $(\mathbf{x}_n, t_n)$ corresponds to an input value $\mathbf{x}_i \in \mathbb{R}^3$ and the corresponding target value $t_i \in \mathbb{R}$.

1. Provide the definition of a linear regression model (in its most general form) with parameters $\mathbf{w}$ that can be used for estimating a non-linear function y such that $t \approx y(\mathbf{x}, \mathbf{w})$.
2. Discuss possible causes of overfitting for this problem and how to avoid/attenuate them.

EXERCISE C2

Consider the following data set for binary classification (white vs black circles).

1. Draw in each of the diagrams below a possible solution for a method based on Perceptron with very small learning rate and a possible solution for a method based on SVM.
2. Describe the difference between the two solutions and briefly explain how these are obtained with the two methods.
3. Discuss which solution would you prefer and why.



Machine Learning – Exam - December 16, 2019

Time limit: **2 hours**.

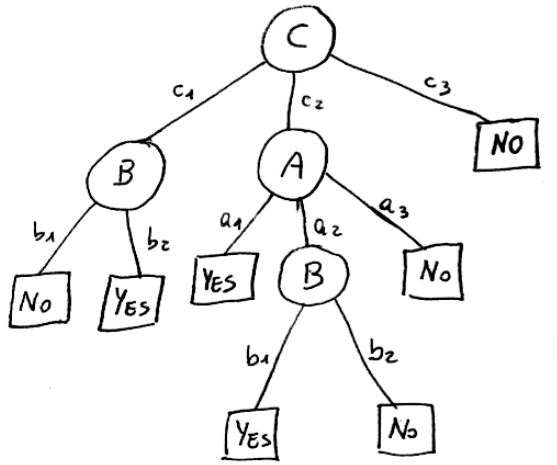
Last Name ..... First Name ..... Matricola .....

Note: if you are not doing the regular exam for ML 2019/20, 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 $f : A \times B \times C \rightarrow \{+, -\}$, 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. 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, with the last image showing a possible final solution).

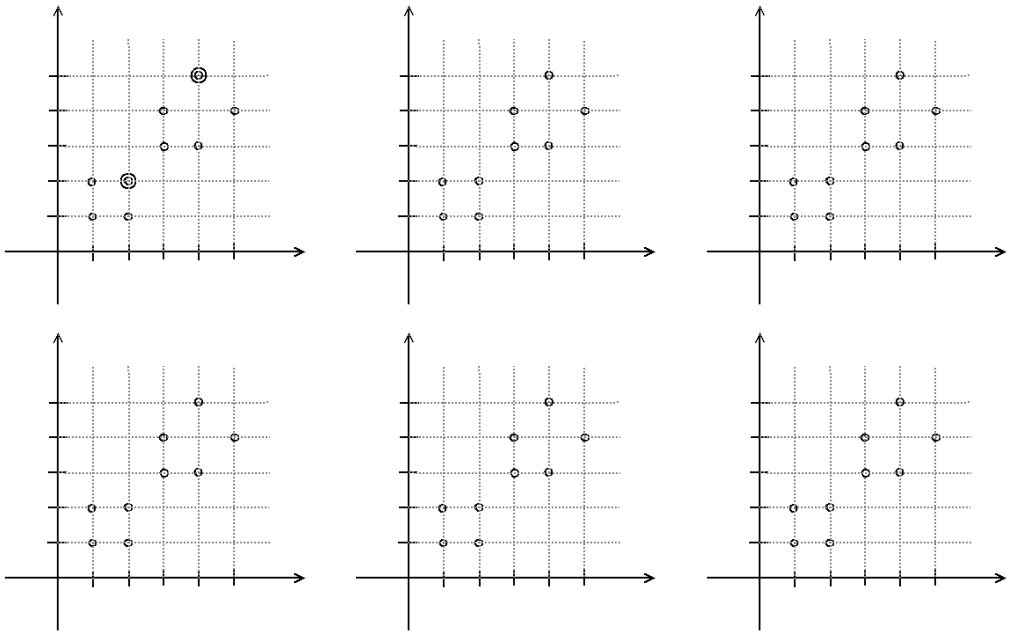
EXERCISE 4

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, respectively.

1. Provide the dimensions of the weight matrices W_1 and W_2
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 5

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 6

1. Briefly describe what is the architecture of an autoencoder and its purpose.
2. Draw an example of autoencoder.

Machine Learning – Test - November 4, 2019

Time limit: **2 hours**.

Last Name First Name Matricola

....

Note: if you are not doing the regular exam for ML 2019/20, 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

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:

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

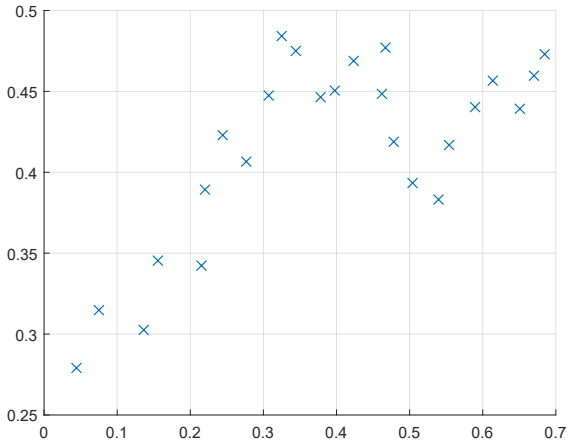
with $\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{t}$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and λ the regularization weight.

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 – Test - December 16, 2019

Time limit: **2 hours**.

Last Name First Name Matricola

....

Note: if you are not doing the regular exam for ML 2019/20, 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

A car driver in Rome has to move from one side of the Tiber river to the other very often every day. 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, but due to traffic conditions, it is not guaranteed that the shortest path is also the quickest way. Moreover, traffic conditions are unpredictable, fully observable and (quasi-)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.

EXERCISE 2

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.

EXERCISE 3

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, respectively.

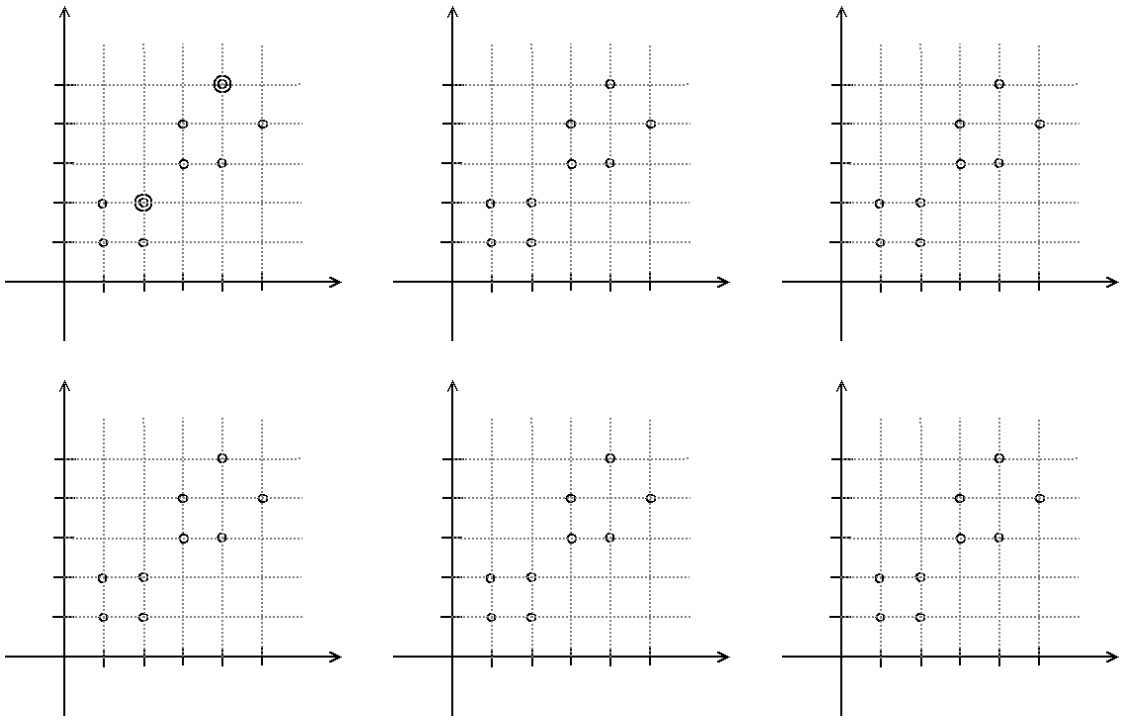
1. Provide the dimensions of the weight matrices W_1 and W_2
2. Provide the formula explicitly stating how the values of $\mathbf{y}$ are computed given an input vector $\mathbf{x}$ in terms of the weight matrices and the activation functions (you can ignore the bias terms).

EXERCISE 4

1. Briefly describe what is the architecture of an autoencoder and its purpose.
2. Draw an example of autoencoder.

EXERCISE 5

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 6

1. Describe an ensemble method (at your choice) for combining multiple learners. Describe precisely all the elements of the method.
2. Assume you have 4 binary classifiers for images with medium-good classification accuracy, describe the application of the method illustrated in the previous point to these 4 classifiers.