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.