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%.

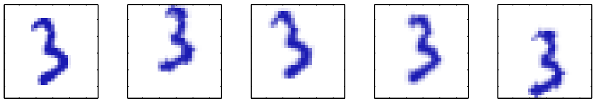
1. If you receive an order for trousers, which is the most probable class the customer who issued the order belongs to? Why?
2. Which is, and how do you compute, the likelihood that an order is for trousers?

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 trainable 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.