Final exam – Introduction to Data Analysis – MOSIG 1
Duration: 1 hour

E. Gaussier, O. Goga, P. Loiseau

May 12, 2020

Exam's rules: Due to the special circumstances, the exam is made available for a duration of 24h (starting Tuesday May 12, 2pm) to prevent network issues or similar issues. This means that you have 24h starting from May 12, 2pm to download the PDF file and start the exam. However, the **duration of the exam is one hour**: from the moment you get this PDF file in front of you, **you must spend only at most one hour working on it** and writing down the answers. Working on the exam only requires access to that PDF file and not to any internet connection.

Any document is allowed during the exam.

At the end of the exam, you should send us (to all three teachers) a PDF with your answers. You can use any means to produce that PDF as long as the **answers are very clearly readable**: using latex or word or any other text processing system; writing on a piece of paper and sending us a picture, etc. The time you take after writing the last word to produce a PDF and send it to us does not count towards the 1 hour.

The exam contains 4 questions of approximately equal weight.

Question 1

In this question, we face a binary classification problem. Each data example has p features $x_1, \cdots, x_p$ and the true label is denoted y. We denote by $x^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})$ and $y^{(i)}$ the features and label of the data example i. We assume that there exists a function f^* such that for all i:

$$y^{(i)} = f^*(x^{(i)}) + \epsilon_i,$$

where ϵ_i is a centered random variable (i.e., noise) of variance σ^2 (we assume the ϵ_i's are i.i.d.).

Suppose that we are looking to fit a function $\hat{f}$ from a training dataset within the class $\mathcal{H}$ of all linear functions from any product of monomials of degree at most d (i.e., any product $x_i^k \cdot x_j^m$ for any $i, j \in \{1, \cdots, n\}$ and any $k, m \leq d$) to $\{0, 1\}$. Parameter d represents the complexity of the model.

1. Write down the expression of the true risk and of the empirical risk.

2. Write down the decomposition of the true risk (or generalization error) into approximation and estimation error. Give a lower bound on this risk.

3. How would you propose to do the training for d large?

4. Suppose that x_1 is a sensitive feature. Write down the definition of the independence fairness criterion and of the equal opportunity fairness criterion in terms of equality of the appropriate conditional probabilities.

Question 2

We consider a multilayer perceptron with 1 hidden layer of 3 neurons (in addition to the input layer and the output layer). The input has 2 features x_1 and x_2 and the output has 1 dimension y. We denote by h_1, h_2, h_3 the neurons in the hidden layer. We denote by W the weights of the first linear layer (input to hidden) and by w the weights of the second linear layer (hidden to output).

1. What is the dimension of W and of w?

2. Assume that there is no non-linear function. Give the expression of y as a function of x_1 and x_2. Is this architecture capable of learning the OR function? the AND function? and the XOR function?

3. Now assume that we use the ReLU activation function after the first linear layer. Repeat the previous question.

Question 3

Let us consider the following dataset with 5 examples divided into two classes ('+' and '-').

Exple	Feature #1	Feature #2	Feature #3	class
1	0.3	0.7	5	+
2	0.2	0.8	6	+
3	0.25	0.75	8	+
4	0.7	0.3	7	-
5	0.75	0.25	6	-

1. Describe the k-NN algorithm in 3 lines (max).

2. What will the 1-NN algorithm give on the example $(0.2, 0.7, 7)$?

3. Is the result obtained questionable? If yes why and how to solve the associated problem?

Question 4

Describe the k-Means algorithm in 5 lines (max).