Final Exam, Machine Learning (MIRI) June 4th, 2020

1. Instructions

- Please upload your answer through the *Practicals* section at the racó **before 12pm**, Barcelona local time.
- You may upload a scanned or photographed image of your hand-written answer (if *legible*), or a text file.
- If during the exam you have any questions, you may ask me privately through Google Chat – just search my name in the search box `marta.arias.v@upc.edu`; I will be available throughout the duration of the exam
- You **cannot use** any outside help like talking/messaging other people, or searching the internet, etc. Your answer should be solely based on your understanding of the material and work during the course.
- The exam is **closed-book** so you may not look through scripts, class notes or similar resources
- No calculator or any other computation device is needed
- By uploading your solution you are implicitly adhering to the commitment of academic integrity unless you explicitely state otherwise in your exam answer
- **Please select and answer eight (8) questions from the list available in the next page; all questions weigh equally**
- Good luck!

Question 1. In the lab for random forest, you saw the following call to build a random forest for classification (the task is to predict the type of email: spam/nonspam). Please explain what this code does and what its paramters mean. In the code, `learn` contains the indices of rows of the dataset that correspond to the training set.

```
model.rf <- randomForest(type ~ ., data=spam3[learn,],
    ntree=150, proximity=FALSE,
    sampsize=c(nonspam=800, spam=500), strata=spam3[learn,]$type)
```

Question 2. Explain the difference between *training* error, *validation* error, *test* error, and *generalization* error.

Question 3. Explain what the Bayes error rate is and how it relates to the generalization error of any classifier.

Question 4. It is said that generative algorithms for supervised learning learn the joint distribution $p(x, y)$ where y is the target and x corresponds to a vector of explanatory variables, and discriminative algorithms learn $p(y|x)$. Please explain what this means.

Question 5. Please explain the difference between a parameter of a model and a hyper-parameter. You may use an example if you want.

Question 6. Please explain why different runs of the routine `nnet` for training a multi-layer perceptron may give you different solutions.

Question 7. Please explain the potential danger of not having any type of regularization in a modelling task and the danger of having too much of it.

Question 8. Please explain the relation between the bias/variance tradeoff and the k of the k-nearest neighbor algorithm.

Question 9. What is the main objective of the resampling techniques that we have seen during the course (e.g. *cross-validation*)?

Question 10. Can you think of a situation where the EM algorithm for clustering is preferable to k-means?

Question 11. What is the main purpose of the *backpropagation* algorithm in the context of neural networks?