# Machine learning for data science I Exam June 2021

This is a closed book exam.

Write clearly and justify your answers.

Incorrect statements in essay-type questions will incur negative points.

Time limit: 90 min.

**(1) [5 points]** *(essay-type question)* We have the following learning algorithms:

- **OLS**: Ordinary least squares.
- **LR**: Logistic regression.
- **kNN**: k-nearest neighbors with L1 distance.
- **LD**: Linear discriminant analysis
- **BAG**: An ensemble of 100 bagged unprunned decision trees.

If you think some key detail about a learner is missing, you may assume which parameter was used or how it was implemented. However, the assumption must be sensible - something that could reasonably used as a default in some implementation (do not assume behavior that makes the learner clearly behave poorly or is not even feasible to implement). And, whatever you assume, applies to all the below questions.

Order the above learners from best to worst with respect to the following dimensions:

**a.** The time complexity of learning (consider both the number of observations and the number of input variables).

**b.** Time complexity of making a prediction (consider both the number of observations and the number of input variables).

**c.** Interpretability/ease of understanding the model and its predictions.

**d.** Difficulty of implementation if you had to implement it from scratch without third-party libraries.

**e.** Memory (space) complexity of storing the learned model.

Justify your answers.

**(2) [5 points]** *(essay-type question)* List 3 fundamentally different techniques for model agnostic explanation of prediction models (that is, black-box techniques) in a standard regression setting (one or more features, a single target variable). For each of them briefly summarize in which case you would prefer it over the others (or why you would never prefer to use it). For each of them determine the time complexity of applying it (how it depends on the properties of the model, the number of observations, and the number of features).

**(3) [5 points]** *(essay-type question)* Joe F. Random, a machine learning practitioner, occasionally also thinks about problems from a theoretical perspective, because that helps inform his practical decisions. He is contemplating the following problem: We have a classification problem with $X \geq 100$ observations, each with a single target categorical variable and $Y \geq 1$ features. We want to estimate the generalization error of several models, some of which might be very susceptible to overfitting, which we want to detect. Option

A is to estimate it using the average of 20 repetitions of 4-fold cross-validation. Option B is 4 repetitions of 20-fold cross-validation.

Explain to Joe the advantages and disadvantages of each of the two approaches in general. In particular, explain how how these advantages/disadvantages depend on the values of X and Y. Use appropriate terminology regarding estimators and their properties.

**(4) [5 points]** Some machine learning-related mathematics and algorithms:

**a.** Derive that the mean is the optimal estimator under quadratic loss.

**b.** Define a Mercer kernel. Show that the sum of two Mercer kernels is a Mercer kernel.

**c.** Write the Stochastic Neighbor Embedding algorithm (in pseudo-code). Be precise in the objective that we are trying to optimize.

**d.** Write the Random Forests algorithm (in pseudo-code).

**e.** Write the likelihood of the Categorical-Generalized logit GLM (also known as Softmax regression or multinomial logistic regression) and explain why is one of the categories typically set to be the reference category (coefficients set to 0).

**(5) [bonus 3 points]** Some Bayesian statistics:

**a.** Show that ordinary least squares with L2 regularization on the coefficients corresponds to Bayesian linear regression with a Gaussian prior on the coefficients.

**b.** Show that ordinary least squares with L1 regularization on the coefficients corresponds to Bayesian linear regression with a Laplace prior on the coefficients. Hint: Laplace distribution pdf is $p(x) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$.

**c.** What are the advantages and disadvantages of using Bayesian inference as opposed to using Maximum Likelihood inference or another form of Empirical Risk Minimization.

The End.