# Assignment 1 2020

*Alipio Jorge*

*26/02/2020*

## Objectives

The bias-variance trade off is an important characteristic of machine learning processes that has to be considered when making decisions about a learning task. Poor decisions can lead to poor a generalization ability of the models and a failure of the learning task. It is important to know how methods and models behave with different parameter values. Expressive models have a high bias and low variance. Less expressive models have low bias and high variance. The aim of this first assignment is to perform a short study on the effect of expressivenes of models using different classification approaches.

## Dataset

Use the `Pima Indians Diabetes` and / or the `Breast Cancer`dataset in the package `mlbench`. For out-of-sample assessment, split the datasets into a 80%/20% split. Both datasets are binary. Other datasets can be used instead. Check with me.

## Guidelines

- Study the effect of expressivenes using methods like: kNN, linear regression, LDA, QDA, Logistic Regression, CART and others you would like to include.
  - Plot measure train and test errors to illustrate differences in expressiveness and overfitting. Dependeing on the methods, you should use different expressiveness/complexity parameters (as $k$ in `knn`)
  - Draw decision boundaries for subsets of two attributes.
  - Draw decision boundaries considering the two first components in PCA using for example `prcomp` in R. You may adapt the `knn.plot` from classes. However, take into account that you must *translate* between the original data sets and the 2D PCA reduction.
- How does each method place itself in the spectrum of expressiveness?
- Which parameters of these methods can be tuned in order to increase or decrease expressiveness?
- Draw conclusions

## Suggested structure

- Introduction
- Question 1
- Question 2
- ....
- Conclusions
- References

## To submit:

- a fully operational R Markdown document or a Jupyter notebook not longer than about 6 A4 pages when printed (informative concision is also valued), and its rendering in HTML or pdf.
  - The document should have a clear narrative interleaved with plots and tables.
  - The objectives for each experiment and plot should be clear so that the reader understands why it is worth to read a particular part.
  - The conclusion should be a short high level account of what was observed.
  - It is not necessary to describe the methods. It is more important to point out the differences in the methods.
  - It is not necessary to describe the data.

## Evaluation

- This assignment is worth 3.5 values out of twenty
- Components
  - Report 40%
    - Narrative 20%
    - Writing style 20%
  - Technical 60%
    - Correctness 30%
    - Coverage 10%
    - Plots 10%
    - Added value 10% (Out of the box)
- Students can be interviewed about their assignment.

## Groups

Assignments are submitted by groups of of 3 or 4 students. Different elements may have different grades. Other group sizes will not be considered.

It is advisable that the students from the same group perform overlapping work and only after that exchange ideas with each other. Group work is important for learning from other people.

## Submissions

Formal deadline is 18th March 2019, to be submitted in moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

## Ethical principles

When submitting, students commit themselves to follow string ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole of the work. All the materials used and consulted must be credited in the work.