

# Practical work 04 – 10/10/2019

## Model Selection

---

### Objectives

The main objectives of this Practical Work for Week 4 are the following :

- a) Play through an example of overfitting and determine the optimal model complexity by using hyper-parameter tuning based on 5-fold cross validation.
- b) Compose the confusion matrix from the scores obtained from a classification model, compute the different performance measures and learn the characteristics these are capable of highlighting.

### Submission

- **Deadline** : Monday 21 October, 22pm
- **Format** :
  - Exercise 1 (Model Selection)
    - Jupyter notebook.
    - Comments and results (plot with learning curve showing the results for different model complexities) either in the notebook or in a pdf-report.
  - Exercise 2 (Performance Measures) :
    - Jupyter notebook.
    - Comments and results either in the notebook or in a pdf-report.

## Exercise 1 Model Selection

The objective of this exercise is to build a classification systems to predict whether a student gets admitted into a University or not based on their results on two exams<sup>1</sup>.

You have historical data from previous applicants that you can use as a training set. For each training example  $i$ , you have the applicant's scores on two exams ( $x_1^{(i)}, x_2^{(i)}$ ) and the admissions decision  $y^{(i)}$ . Your task is to build a classification model that estimates an applicant's probability of admission based on the scores from those two exams.

In the notebook see `overfitting_stud.ipynb`, you'll find the code to load the data and further instructions.

- a) Construct a dummy predictor that does random predictions. Show numerically that it produces a performance equal to  $1/\text{\#classes}$  - i.e. here equal to 50%.
- b) Construct different (polynomial) models of different complexities (different degree number of parameters). Train these models with the training set and use the trained parameters for doing predictions. Measure the error rate on the training and the test set.  
Remark : Do the programming so that you can easily change the input dataset.
- c) Determine the model best suited for the problem at hand and justify why it is the best model.
- d) For all this use two different versions of the data :
  - i) First version : `scores_train_1.csv` and `scores_test_1.csv` for training and testing, respectively.
  - ii) Second version : `scores_train_2.csv` and `scores_train_2.csv` for training and testing, respectively.

## Exercise 2 Bayes and unbalanced data sets

For this exercise, we will observe the *Cervical Cancer data set*. The data are available on Moodle in file `kag_risk_factors_cervical_cancer.csv` and more information on the task can be found on <https://www.kaggle.com/loveall/cervical-cancer-risk-classification>. Open the notebook `cervical_simplified_stud.ipynb` and follow the instructions. Some code needs to be completed in order to compute prior probabilities and questions need to be answered at the end of the notebook.

---

1. Data source : Andrew Ng - Machine Learning class Stanford

## Exercise 3 Confusion Matrix and Performance Measures

Let's assume we have trained a digit classification system able to categorise images of digits from 0 to 9.






After training, the system has been run against a test set (independent of the training set) including  $N_t = 10'000$  samples. The output of the system is provided by a softmax layer and are given as estimates of the a posteriori probabilities  $P(C_k|\mathbf{x})$  for  $k = 0, 1, 2, \dots, 9$ .

In the file `confusion_a.csv`, you find the output of a first system A with the a posteriori probabilities  $P(C_k|\mathbf{x})$  in the first 10 columns and with the ground truth  $y$  in the last column.

- Write a function to take classification decisions on such outputs.
- What is the overall error rate of the system?
- Compute and report the confusion matrix of the system A.
- What are the worst and best classes in terms of precision and sensitivity (recall)?
- In file `confusion_b.csv` you find the output of a second system B. Which of the systems, A or B, performs better in terms of error rate and F1?

You can use the jupyter notebook `confusion_matrix_stud.ipynb`.

## Exercise 4 Optional : Review Questions

- Explain the terms bias and variance. What are the factors that make the bias larger or how can it be made smaller? What factors lead to a large variance or how can it be reduced? 
- Why is the training error an increasing function of the split ratio (fraction of samples used for training)? Why is the validation error a decreasing function of the split ratio? 
- Describe in words how you construct a confusion matrix. How can you compute the accuracy from it? How does this relate to the error rate? 
- Describe the terms class accuracy, recall and precision. Describe typical situations where you would like to obtain a high recall or a high precision, respectively. 
- What is the minimum overall accuracy of a 2-class system relying only on priors and that is built on a training set that includes 5 times more samples in class A than in class B? 
- Let's assume a trash recycling factory using an automated system to detect PET versus non-recyclable plastic bottles. The system uses a camera taking images from a running belt where the bottles are traveling. Today in the news, Migros announced that they would double the quantity of PET bottles on sale in their shops. What do you need to do with your system? (You can make assumptions on the content of the initial training set you used to build the system)