# Machine Learning - Lesson n°2

Chloé-Agathe Azencott

September 20, 2016

Scribes : Manon REVEL et Nathan VERMEERSCH

## 1 Supervised Learning - Reminder

From a training set $\mathcal{D} = \{x^{\mathrm{i}}, y^{\mathrm{i}}\}$ ($x^{\mathrm{i}}$ are the inputs objects and $y^{\mathrm{i}}$ the expected outputs), one wants to deduce a function. The output of the function may be a regression ($y^{\mathrm{i}}$ is a real number) that models the behavior of a phenomenon or a classification ($y^{\mathrm{i}}$ is 0 or 1) that predicts class label of the input objects. To achieve this, the model has to be generalized from the presented data to unseen situations in a "reasonable" way. Many models may fit with a set of data, however their complexity and accuracy are variable.

The point of this section is to find a way to evaluate and to select models.

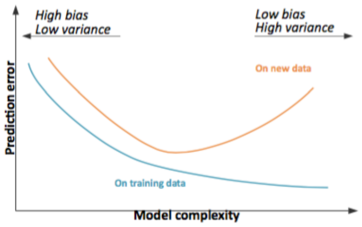
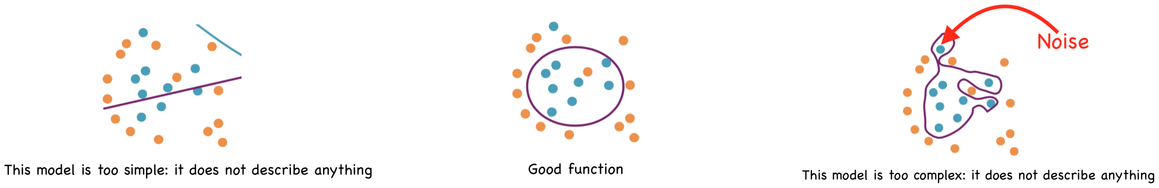


Figure 1: Generalization error relatively to the model complexity

Even if, intuitively, one could have thought that increasing the model complexity would approach more accurately a data set, Figure 1 shows that a complex model that fits with a training set will be unable to describe a new set. Overfitting occurs: we model noise rather than intended inputs (Figure 2c). Inversely, a set described with too a simple model will be unable to model a phenomenon and underfitting occurs (Figure 2a).


This model is too simple: it does not describe anything
Good function
Noise
This model is too complex: it does not describe anything

(a) *Underfitting* (b) *Accuratemodel* (c) *Overfitting*

Figure 2: Three simple graphs

Therefore, models have to be evaluated in order to find the one that fits the best (Figure 2b).

Contrary to models previously studied in other fields like partial differential equations or mechanics, machine learning problems are ill-posed problems, that is to say they are no well-posed problems. Well-posed problems have the following properties: a unique solution exists, and it depends continuously on the initial condition. In machine learning, data gives information about the solution, but is not sufficient to lead one to a unique solution. As a result, one has to find some criteria to choose the best model. In the case of supervised learning, that means making assumptions about $H$, the hypothesis space. The inductive bias is the set of assumptions that the learner makes to generalize beyond the training data. The model selection focuses on choosing the right inductive bias. We can evaluate this choice thanks to the empirical errors $E$ that says how our label defers from our prediction.

**Transitory question :** Empirical errors allows us to elaborate how good a label is on a training set. But what if we take a new set ?

# 2 Model Evaluation

**Validation sets**

One should not be satisfied with tests on the training set. Indeed, they can be misleading, as the graph in Figure1 shows. For instance, a very complex model might seem to work for the training set, but could yet be totally inefficient for other pieces of data. That is why a validation set, other than the training set, is needed.

Once we have chosen parameters, to tune them, we separate the data sets. At first sight, it seems logical to simply divide the available data in two parts: one is the training set, the other the validation set. However, data should be used in the most efficient way in machine learning, as data can be scarce or difficult to obtain. Thus, there exists several methods to make the most of available data. We divide the data set in training set (used to see if the parameters fit) and a validation test (used to tune the parameters). We can also create a test set (used to assess the performance). To compare optimal and suboptimal parameters, we have all interests to divide it more. NB: Model selection pick the best model, model assessment estimates its prediction error on new data.

Important concept

In this chapter, model selection methods will be used and two processes of sample re-use will be explained: cross-validation and bootstrap.

**Transitory question :** How much data should go in each of the training, validation and test sets ?

## 2.1 Cross-Validation

The Cross-Validation is a method based on the re-use of the sample. We cut the training set in k separate folds. We train on (k-1) folds and validate on one.

Important concept

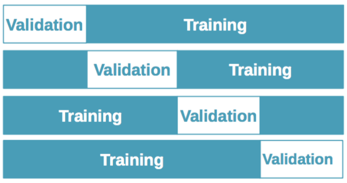


Figure 3: Cross-Validation with $k = 4$

The cross-validation estimate of the prediction error is:

$$CV(f) = \sum_{i=1}^{n} L(y_i, f_{k(i)}(x^{\mathrm{i}}))$$

It estimates the expected prediction error. We do then, k evaluations, with k outcomes and k empirical errors. Notwithstanding, one should pay attention to the value of $k$ when using cross-validation. For instance, if one chooses $k = N$ where $N$ is the number of items in the set of available data, the following problems can be encountered: potential high variance and burdensome computation. As a result, values of $k = 5$ or $k = 10$ are usually chosen.

- **Positive points**

  Low bias estimator : we have a unbiased estimator of the expected error

- **Negative points**

  High variance : we have overlap between the k evaluations, so they are correlated, so dependent. The variance is high because of the sum of the co-variances.

  The computation can become burdensome

With a small k, one trains on a little fraction of the original training data, so the variance is high and it gives an inaccurate model. Generally we take then k=5 or k=10 to have at least 80percent of the original training set in the newly defined one.

## 2.2 Bootstrap

This method consists in building $B$ data sets by picking randomly with replacement from the available data. Typically, $B = 100$. Bootstrap also asks several questions that one should be careful to: leave-one-out bootstrap error (the probability that a particular item belongs to a particular data set is only 63,2%, so a data set can be restrictive), and the size of each data set.

Important concept

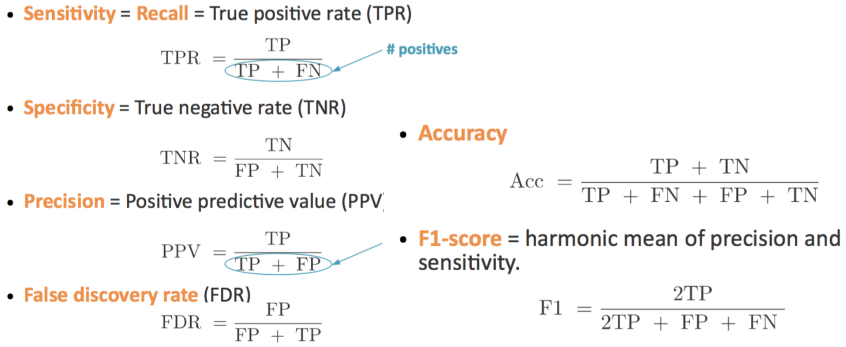
## 2.3 Confusion Matrix

|  |  | True | Class |
|---|---|---|---|
|  |  | -1 | +1 |
| Predicted | -1 | **T**rue **N**egative | **F**alse **N**egative |
| Class | +1 | **F**alse **P**ositive | **T**rue **P**ositive |

False positives can also be named type I errors or false alarms.

False negatives can also be called type II errors or misses.

We can then define a set of tools thanks to which we evaluate a model. Indeed, the proportion of positives that are correctly identified (sensitivity or recall), the proportion of negatives wrongly said positive (a wrong alarm or specificity) and inversely may become relevant rates to consider according to the environment of one's studies : in a clinical settings, we have better to have a very high sensitivity otherwise that means that we have failed to diagnostic patients.



- Sensitivity = Recall = True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

# positives

- Specificity = True negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

- Precision = Positive predictive value (PPV

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- False discovery rate (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

- Accuracy

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$$

- F1-score = harmonic mean of precision and sensitivity.

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Figure 4: Tools for the evaluation of a model

## 2.4 Receiver-Operator Characteristic (ROC curve)

The curves are a graphic mean to know if a model is optimal and to discard suboptimal models. The ROC curve represents the true positive rate in ordinate and the false positive rate in abscissa.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

A good model should have a curve localized in the high left corner. We should maximize the area under the curve.

## 2.5 Precision-Recall (PR curve)

The Precision-Recall curve is a graph with the recall (the sensitivity) in ordinate and the precision (or positive predicated value) in abscissa.

$$TPR = Recall = \frac{TP}{TP + FN}$$

$$Precision = PositivePredictedValue = \frac{FP}{FP + TN}$$

An algorithm with a high precision is an algorithm that points out more relevant results than irrelevant. An algorithm with a high recall is an algorithm that points the most relevant results. In clinical settings, a high recall is appreciated.

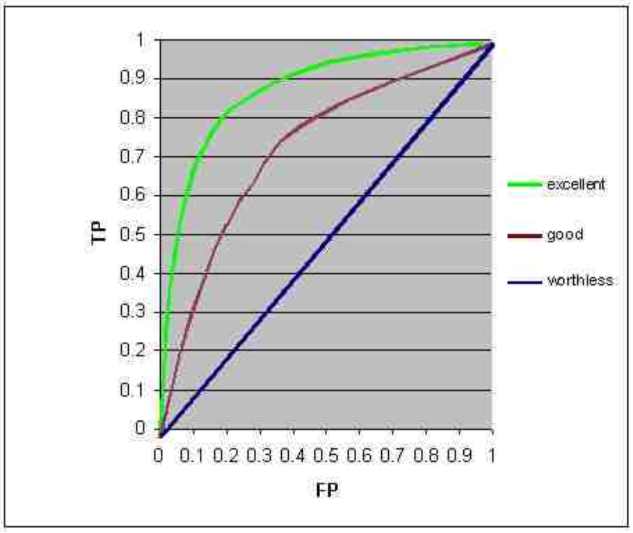


Figure 5: Cross-Validation with k=4

# 3 Model Selection

## 3.1 Penalizing model complexity

As indicated by his name, a penalizing model complexity tool penalize a model that is too much complex with an added contribution in the empirical error. We define then an augmented error E' :

Important concept

$$E' = empiricalerrorE + modelcomplexity\lambda$$

1. Mallow's Cp: It compares the precision and bias of models. It is defined with a criteria to assess the fit of regression models with different numbers of parameters are being compared.

2. Akaike's Information Criterion (AIC)

3. Bayesian Information Criterion (BIC)

## 3.2 Minimum description length

Based on the Shannon's information theory, the MDL method will choose the model that compress the data the best. We calculate the sum of the average code length to transmit the difference between model prediction and true outputs and the average code length to transmit $\theta$ for each model. This is the code length needed to transmit the outputs Y. We then choose the model with the smallest length.

## 3.3 Structural risk minimization

Finally, the structural risk minimization separates the models according to their VC dimension. It fits a nested sequence of the models (their VC dimension increases). The best model is the one with the lower bound on test error.

We have seen some tools for measures of performance for classifiers and regressors.

Important concept

Bibliography : Computer Aided Diagnosis of Lung Ground Glass Opacity Nodules and Large Lung Cancers in CT.