Module overview

Contents

What you will learn

This module gives an intuitive introduction to the very **fundamental concepts** of overfitting and underfitting in machine learning.

Machine learning models can never make perfect predictions: the test error is never exactly zero. This failure comes from a **fundamental trade-off** between **modeling flexibility** and the **limited size of the training dataset**.

The first presentation will define those problems and characterize how and why they arise.

Then we will present a methodology to quantify those problems by **contrasting the train error with the test error** for various choice of the model family, model parameters. More importantly, we will emphasize the **impact of the size of the training set on this trade-off**.

Finally we will relate overfitting and underfitting to the concepts of statistical variance and bias.

Before getting started

The required technical skills to carry on this module are:

- skills acquired during the “The Predictive Modeling Pipeline” module with basic usage of scikit-learn.

Objectives and time schedule

The objective in the module are the following:

- understand the concept of overfitting and underfitting;
- understand the concept of generalization;
- understand the general cross-validation framework used to evaluate a model.

The estimated time to go through this module is about 3 hours.
