Sampling and overfitting

Formation IA biodiversité

Paul Tresson

May 14, 2025

UMR AMAP

Introduction

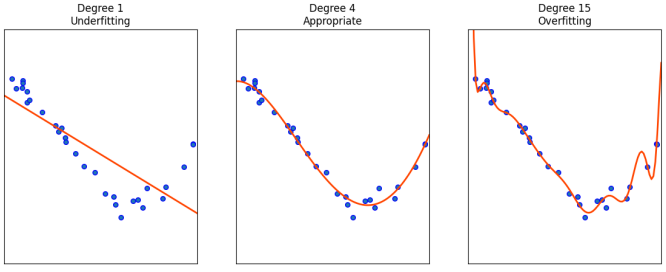
What do we want when modelling ?

What do we want when modelling ?

All models are wrong but some are useful

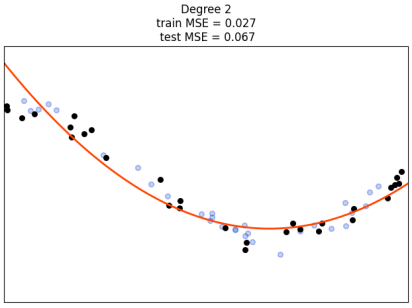
Overfitting

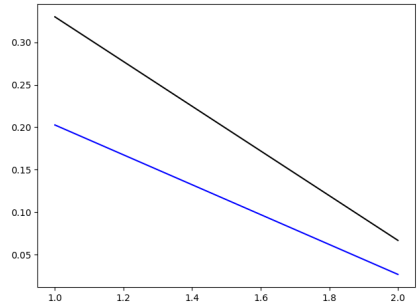
What is overfitting



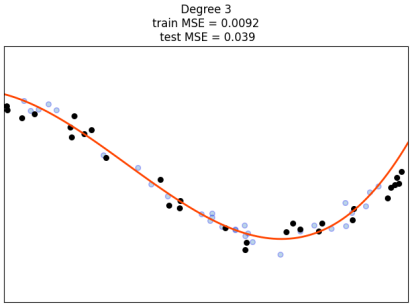
adapted from scikit-learn docs

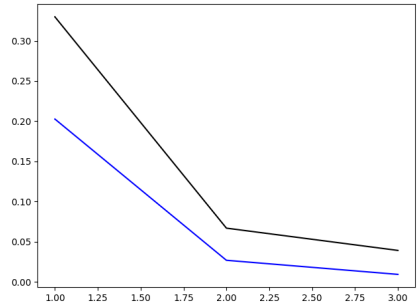
Common tools and intuitions - Train/Test loss



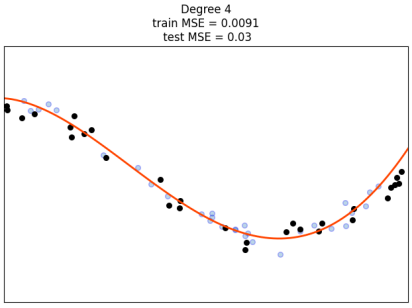


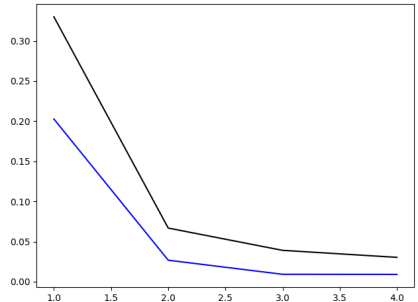
Common tools and intuitions - Train/Test loss



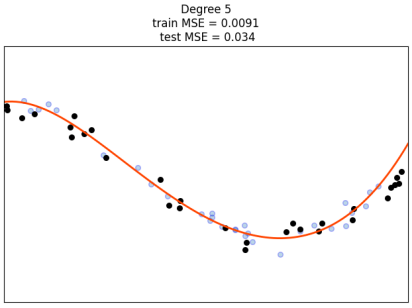


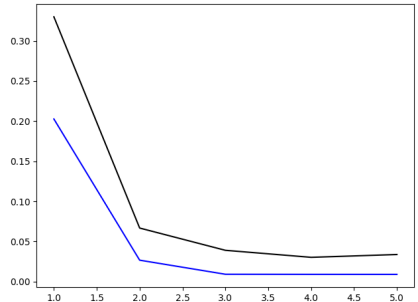
Common tools and intuitions - Train/Test loss



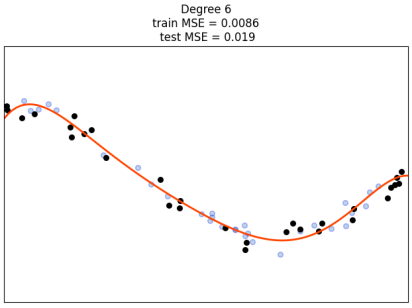


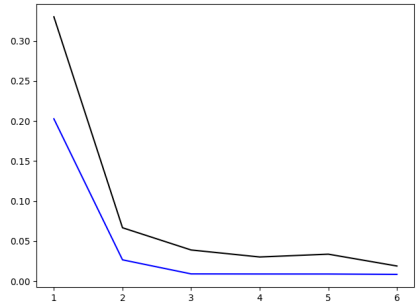
Common tools and intuitions - Train/Test loss



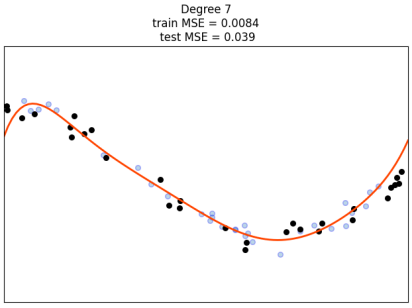


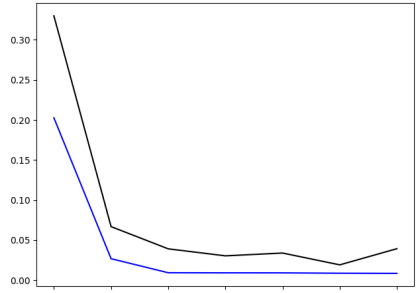
Common tools and intuitions - Train/Test loss



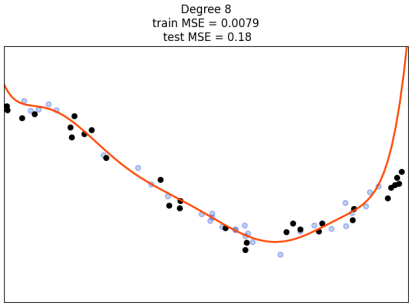


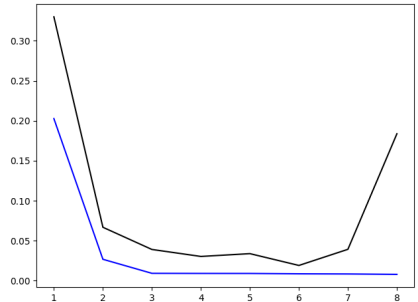
Common tools and intuitions - Train/Test loss



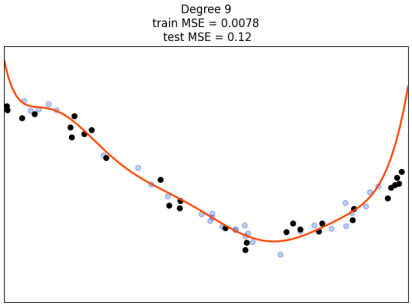


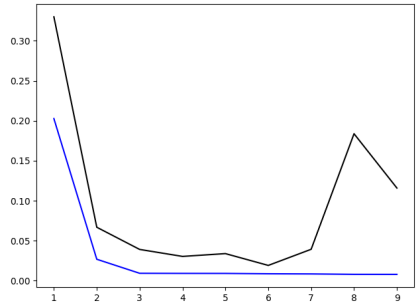
Common tools and intuitions - Train/Test loss



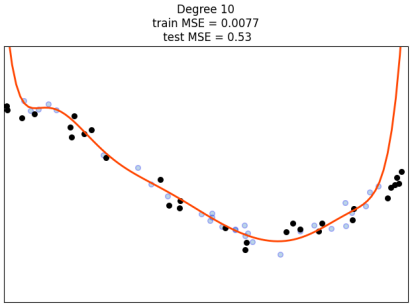


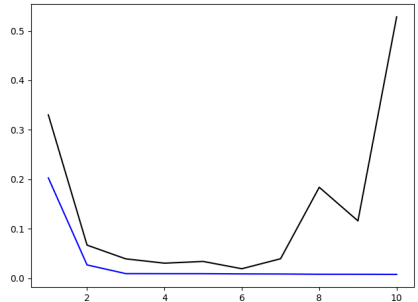
Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss

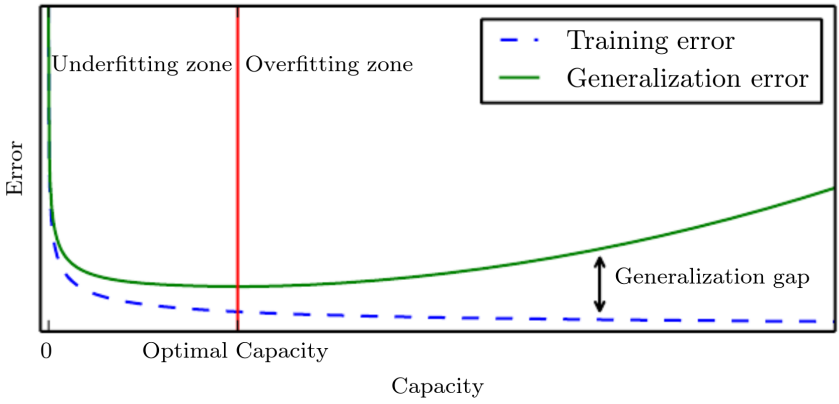


Figure from Goodfellow et al., 2016

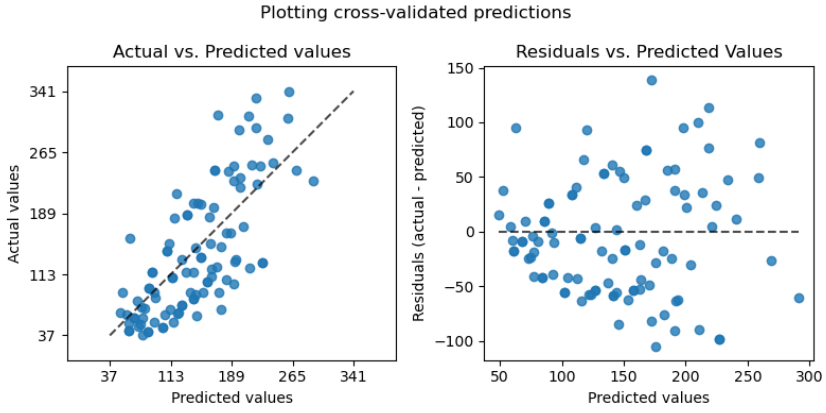
Common tools and intuitions - AIC/BIC

Akaike information criterion (AIC)

Bayesian information criterion (BIC)

Is the model parameter efficient ?

Common tools and intuitions - Biases



from scikit-learn docs

And in Machine(/Deep) Learning ??

How many parameters to have

Shrek learning botany starting from random noise ?

And in Machine(/Deep) Learning ??

$\approx 2.5B$?

Common traps (in Ecology)

How to sample and evaluate ?

Usefull ressources

- `scikit-learn` docs !
-
-

Thanks for you attention !

Let's practice !

References i

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016). ***Deep learning***. Vol. 1. 2. MIT press Cambridge.