Sampling and overfitting

Formation IA biodiversité

Paul Tresson

May 15, 2025

UMR AMAP

Introduction

What do we want when modelling ?

- Understand things

What do we want when modelling ?

- Understand things
- **Predict things**

What do we want when modelling ?

"All models are wrong, but some are useful"

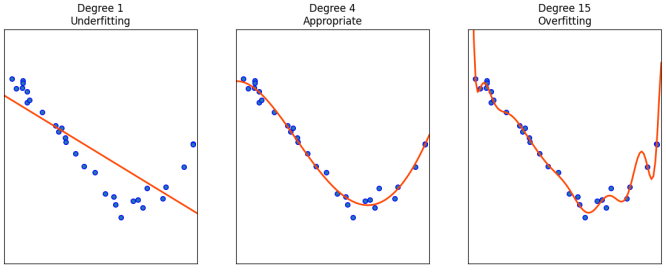
George E. P. Box

What do we want when modelling ?

- **Robustness**: Useful when mistakes
- **Generalization**: Useful applied elsewhere

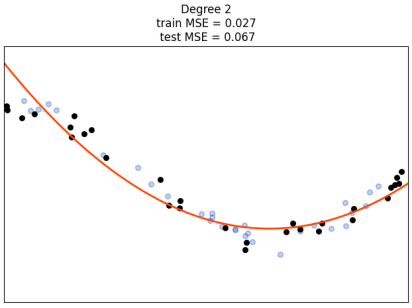
Overfitting

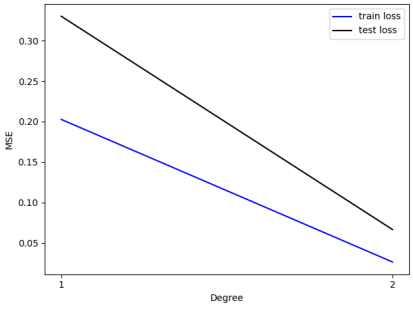
What is overfitting



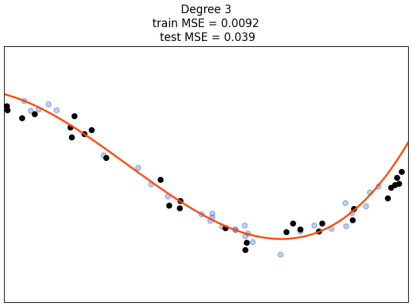
adapted from scikit-learn docs

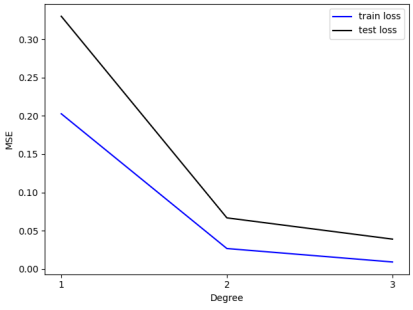
Common tools and intuitions - Train/Test loss



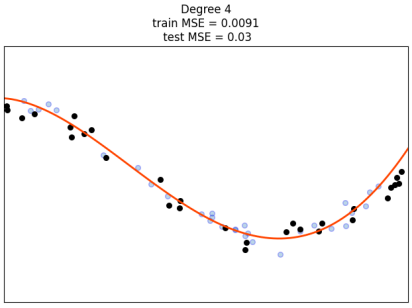


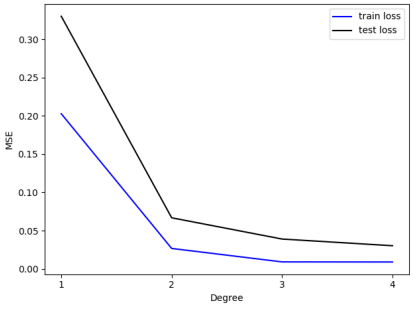
Common tools and intuitions - Train/Test loss



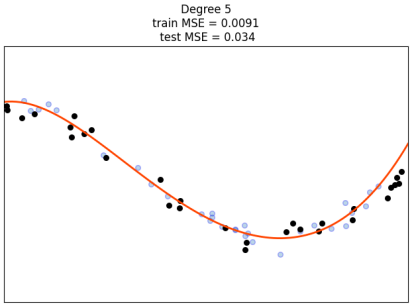


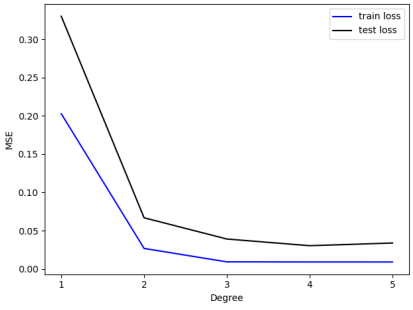
Common tools and intuitions - Train/Test loss



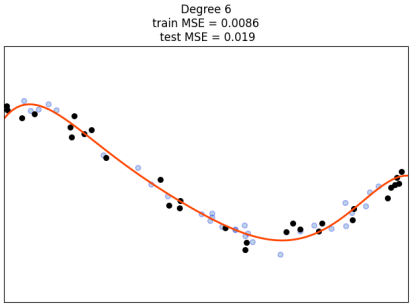


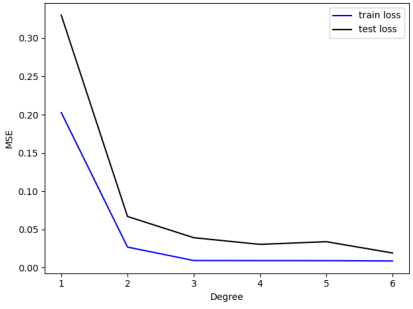
Common tools and intuitions - Train/Test loss



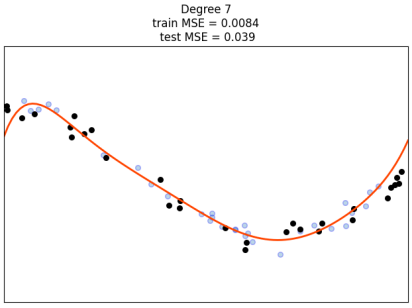


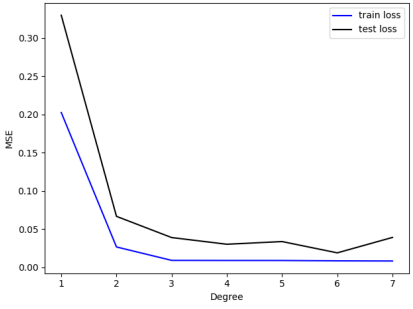
Common tools and intuitions - Train/Test loss



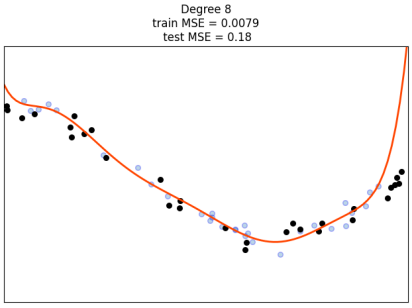


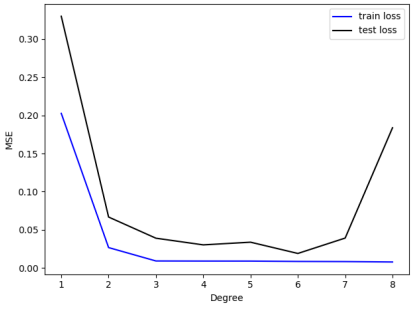
Common tools and intuitions - Train/Test loss



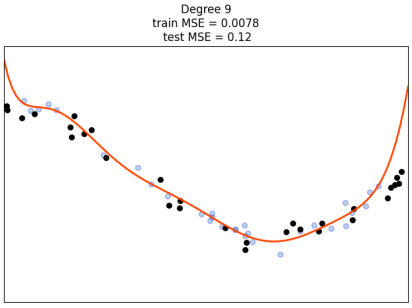


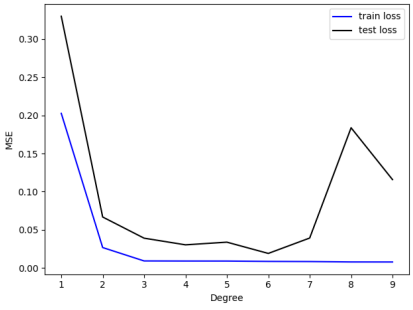
Common tools and intuitions - Train/Test loss



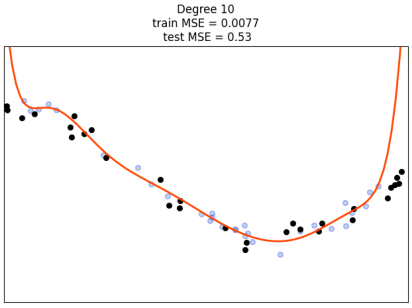


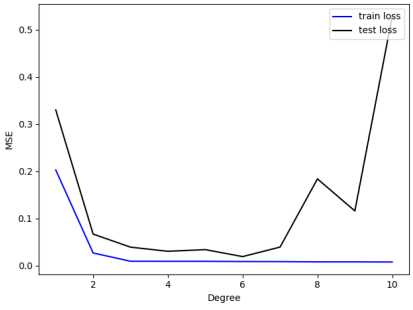
Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss





Common tools and intuitions - Train/Test loss

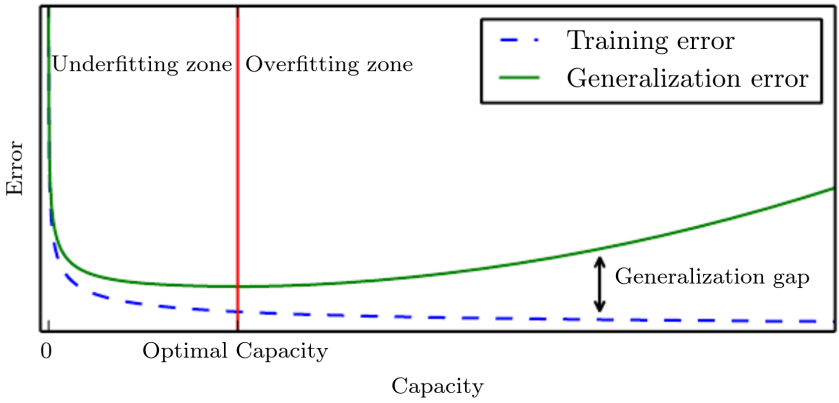


Figure from Goodfellow et al., 2016

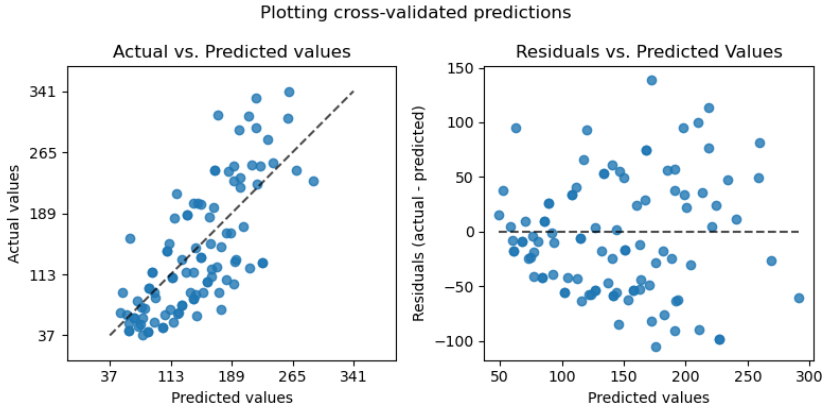
Common tools and intuitions - AIC/BIC

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 ??

$\approx 2.5B$?

Root Causes

Too many parameters

Root Causes

Too many parameters

Too little training data

Root Causes

Too many parameters

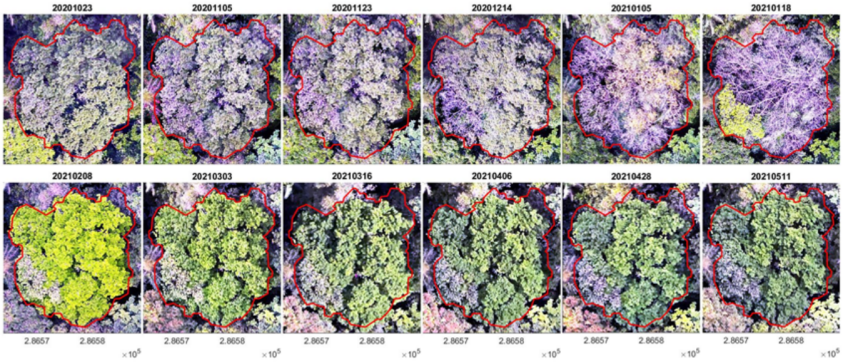
Too little training data

(bad) training data

Illustrated examples in Ecology

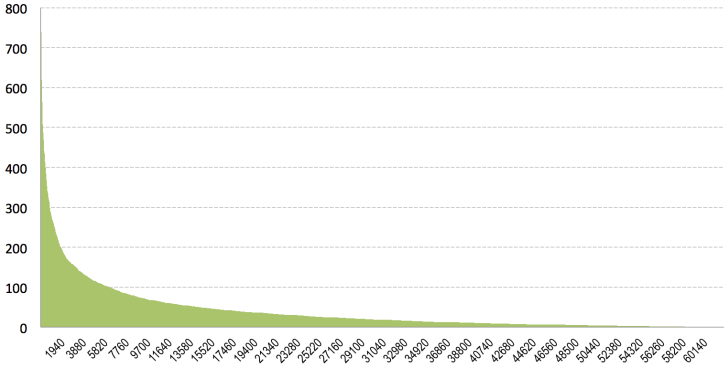
Contraints in ecology

Data from the real world is noisy,



Contraints in ecology

Data from the real world is noisy, unbalanced,

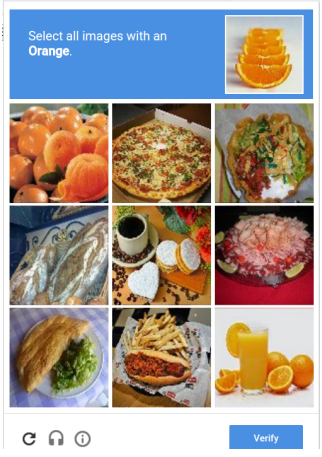


Contraints in ecology

Data from the real world is noisy, unbalanced, hard to collect,

Contraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

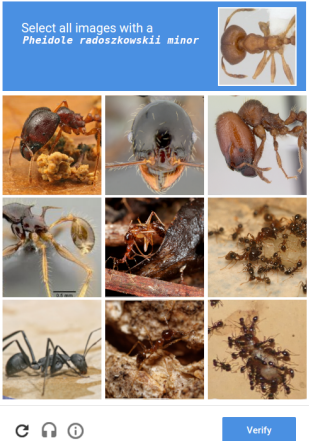


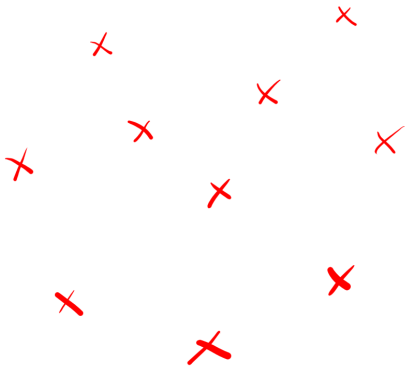
Contraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

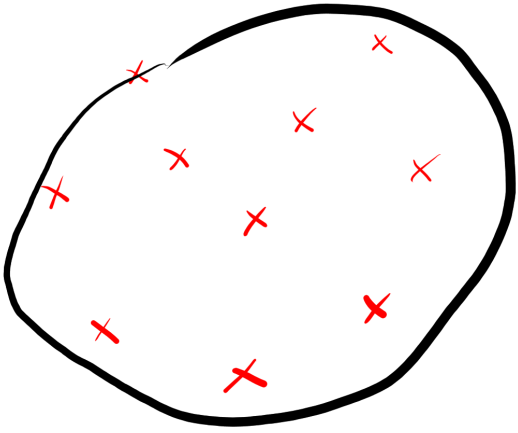
Contraints in ecology

Data from the real world is noisy, unbalanced, hard to collect, hard to interpret.

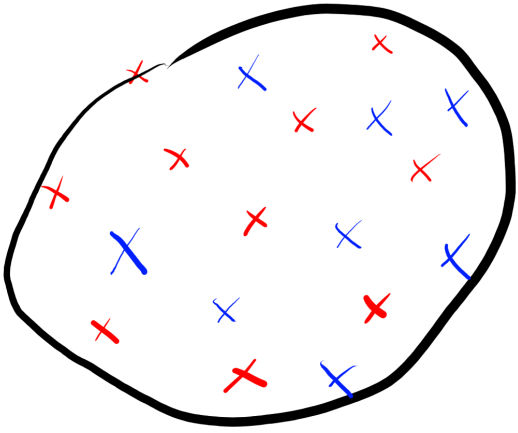




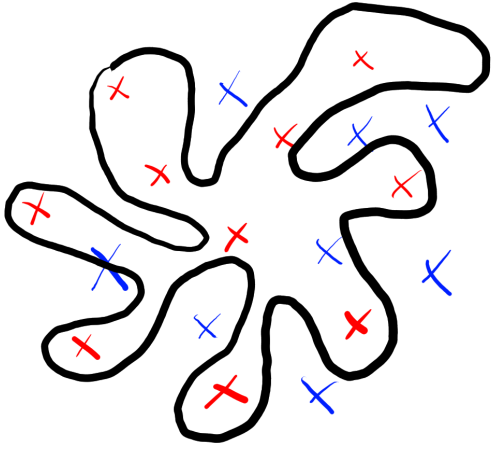
Train set



A good fitted model



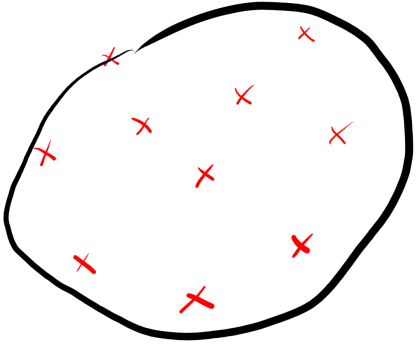
Test set



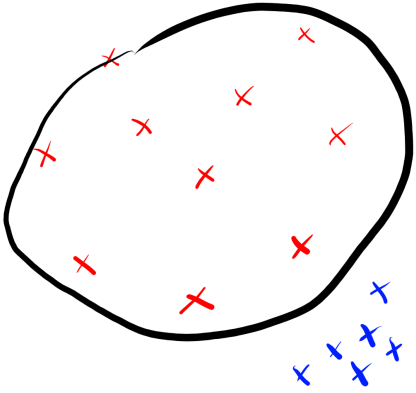
An overfitted model

Biases in the train set

Biases in the train set



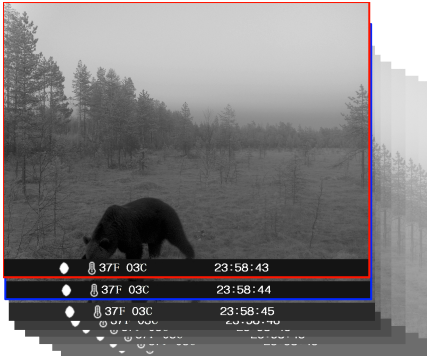
Biases in the train set

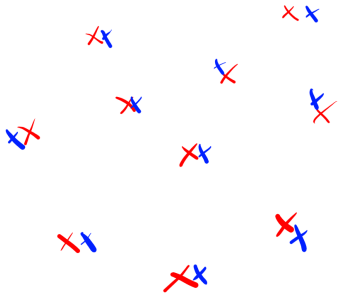


Biases in the train set - autocorrelation

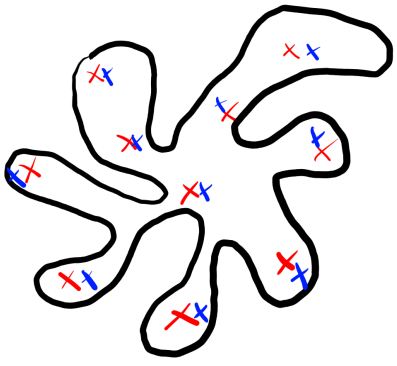
Biases in the train set - autocorrelation

Biases in the train set - autocorrelation

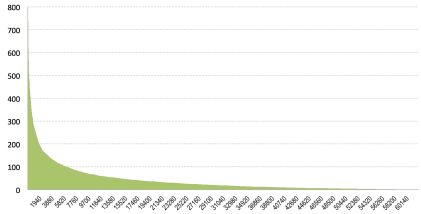




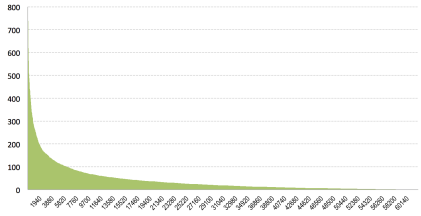
Biases in the train set - autocorrelation

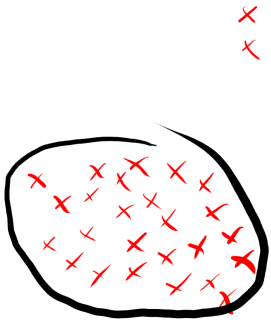


Unbalanced data

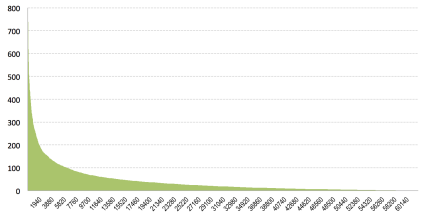


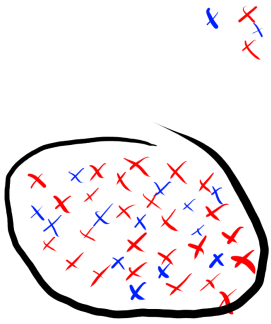
Unbalanced data



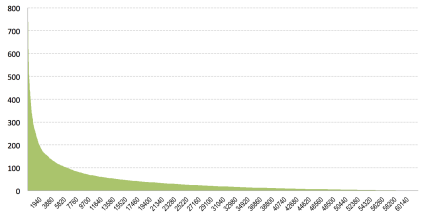


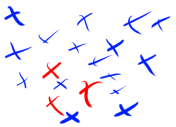
Unbalanced data

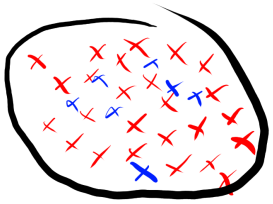




Unbalanced data





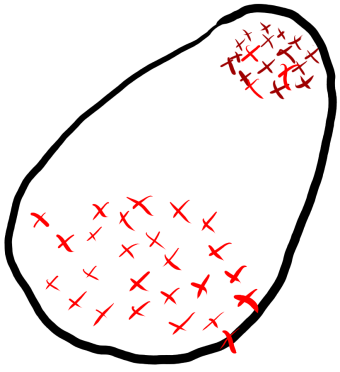


Deal with unbalanced data

- Oversample ?

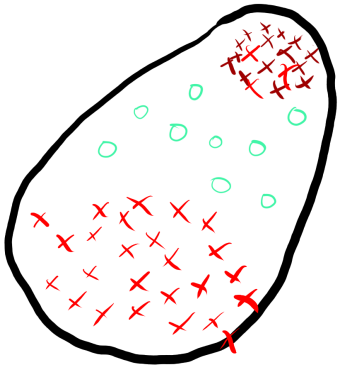
Deal with unbalanced data

- Oversample ?



Deal with unbalanced data

- Oversample ?

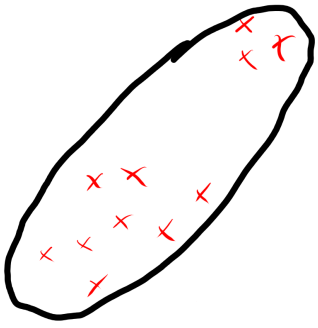


Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?

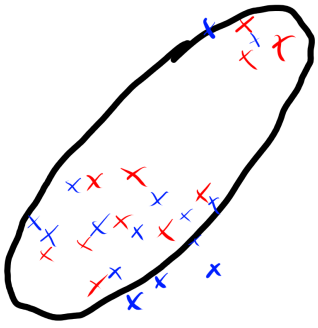
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



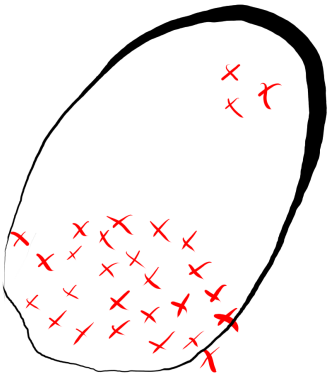
Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?



Deal with unbalanced data

- Oversample ?
- Undersample/saturate ?
- Adapt loss ?



Deal with lack of data

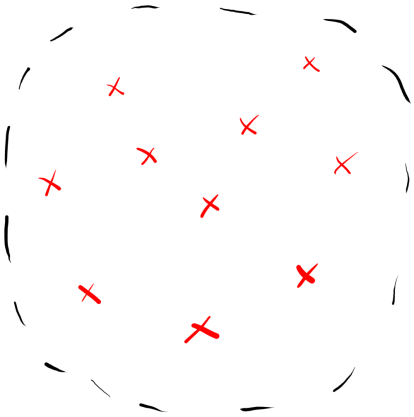
- Data augmentation

Deal with lack of data

- Data augmentation

Deal with lack of data

- Data augmentation
- Pretrained model



Deal with lack of data

- Data augmentation
- Pretrained model
- **... collect more data**

Need to be very carefull on how to evaluate

How to sample and evaluate ?

Random split ?

“random split training validation 80/20”

Random split ?

"random split training validation 80/20"

> For the uncurated dataset, we randomly sample 142 million images

Oquab et al., 2023

Random split ?

"random split training validation 80/20"

> For the uncurated dataset, we randomly sample 142 million images

Oquab et al., 2023

Works for huge DL papers, maybe not for you

Cross-validation

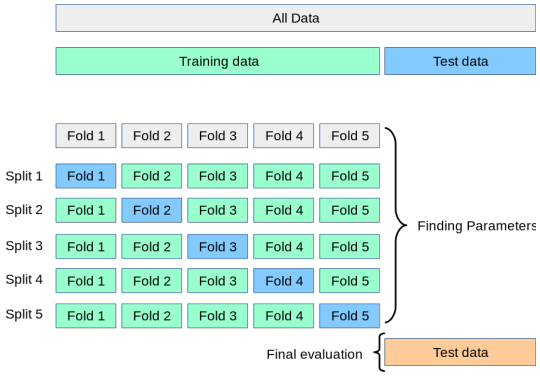


Figure from scikit-learn docs

Cross-validation

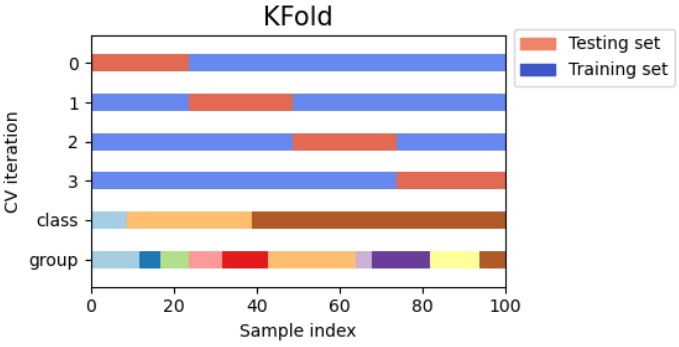


Figure from scikit-learn docs

Cross-validation

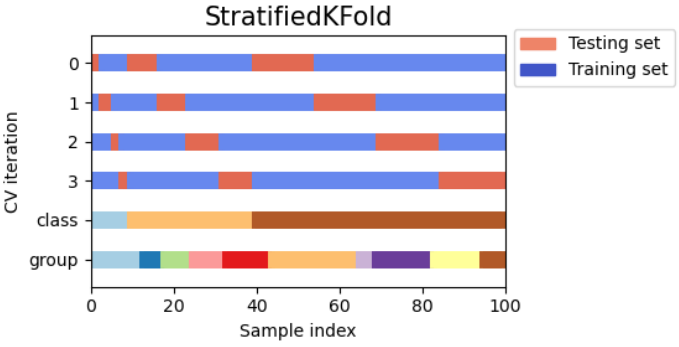


Figure from scikit-learn docs

Case study : spatial cross-validation

Case study : Aging models ?

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.

Oquab, Maxime et al. (2023). **"Dinov2: Learning robust visual features without supervision"**. In: *arXiv preprint arXiv:2304.07193*.