HYPERPARAMETER TUNING

JENS BAETENS

Tunen
Aanpassen
v.d hyperparameters
manueel
automatische
leerbaarheid (brute-force)
efficienter

HYPERPARAMETERS VS PARAMETERS?

HYPERPARAMETERS VS PARAMETERS?

Parameters

- Interne waarden van het model
- Worden geoptimaliseerd bij training

Gewichten / Support vectors

Hyperparameters

- Configuratie van het model
- Vrij te kiezen

→ C / gamma / ~~kernel~~ L_1 / L_2-norm

→ "graad v.d. hogere orde features"

→ type scaling ...

HYPERPARAMETER TUNING

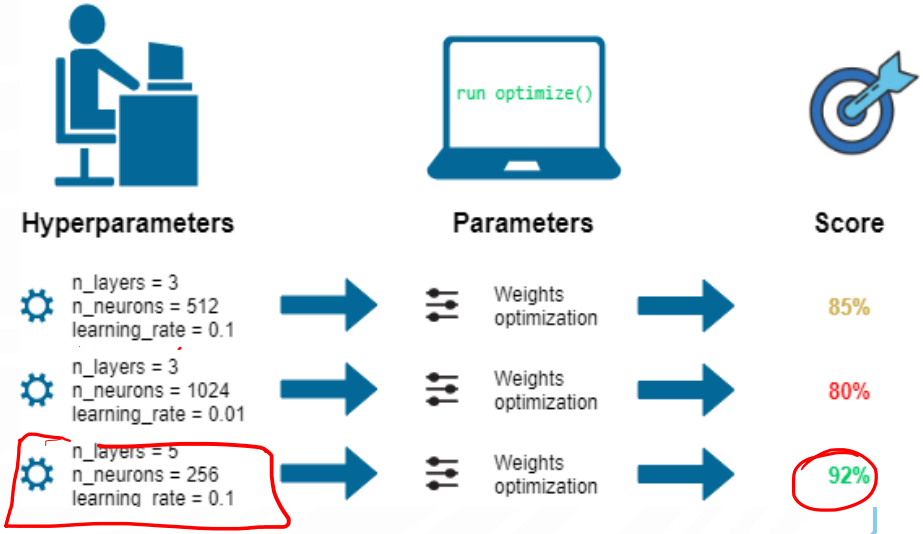
Wat is de beste configuratie?

set hyperparameters

Overloop alle mogelijke combinaties

Kies het model met de hoogste score

acc.
precisie
f1-score
...



WELKE DATA?

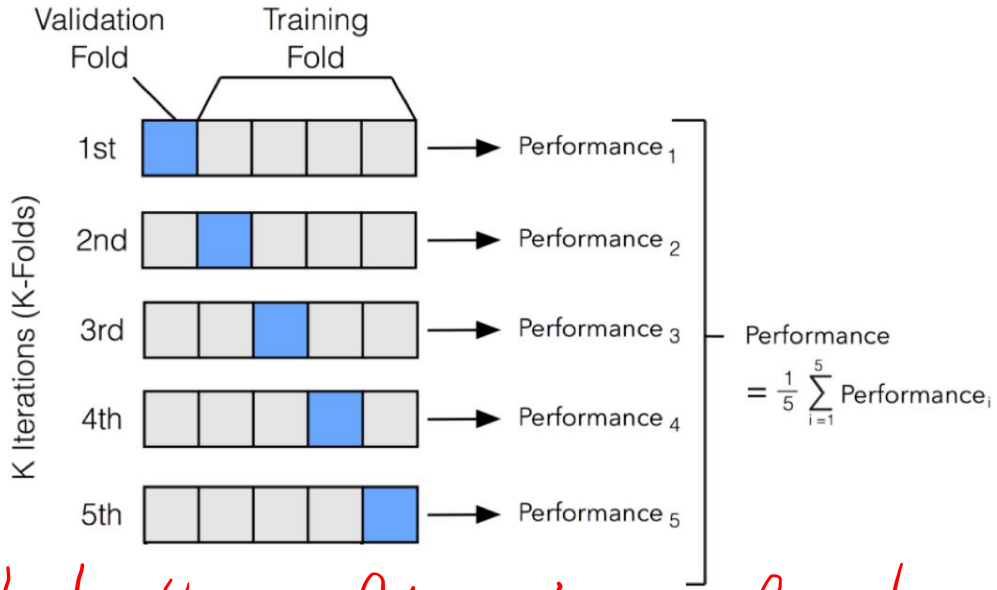
WELKE DATA?

Steeds evalueren op ongeziene data

Daarom zowel validatie als test set nodig (holdout methode)

Veel gebruikte percentages: 70/15/15 tot 98/1/1 bij Big Data

K-FOLD CROSS VALIDATION



DATA LEAKAGE

Data van buiten de trainingsdata gebruikt voor het model te trainen

Probleem: bereikte performantie minder betrouwbaar voor ongeziene data

Vooral risico bij:

- Tijdreeksen → data uit de toekomst
- Meerdere lijnen die tot dezelfde persoon/klant behoren
  ↳ duplicaten

DATA LEAKAGE

Hoe minimaliseren?

- Geen features die ingevuld worden na de target (behandeld voor …)
- Maak gebruik van pipelines zodat scalers en imputers werken binnen de folds en geen data van de gehele dataset gebruiken.
- Pas op met oversampling
- Hou de testdata volledig apart! → niet voor scaling the fitten / null-waarden

Voorbeeld: https://www.kaggle.com/c/the-icml-2013-whale-challenge-right-whale-redux/discussion/4865#25839

Grootte van de bestanden en de timestamps maakten het mogelijk om de classificatie uit te voeren.