Hyperparameter tuning

JENS BAETENS

Hyperparameters vs parameters?

Hyperparameters vs parameters?

Parameters

- Interne waarden van het model
- Worden geoptimaliseerd bij training

Hyperparameters

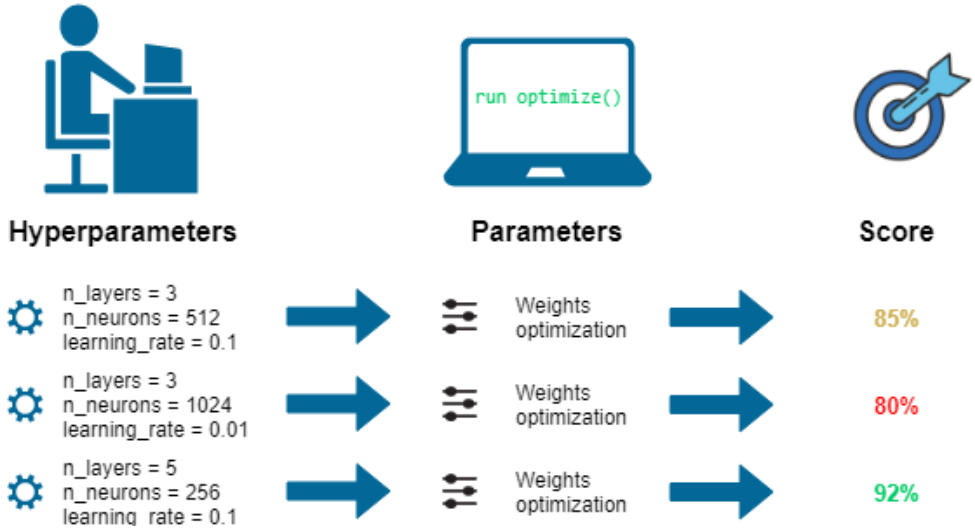
- Configuratie van het model
- Vrij te kiezen

Hyperparameter tuning

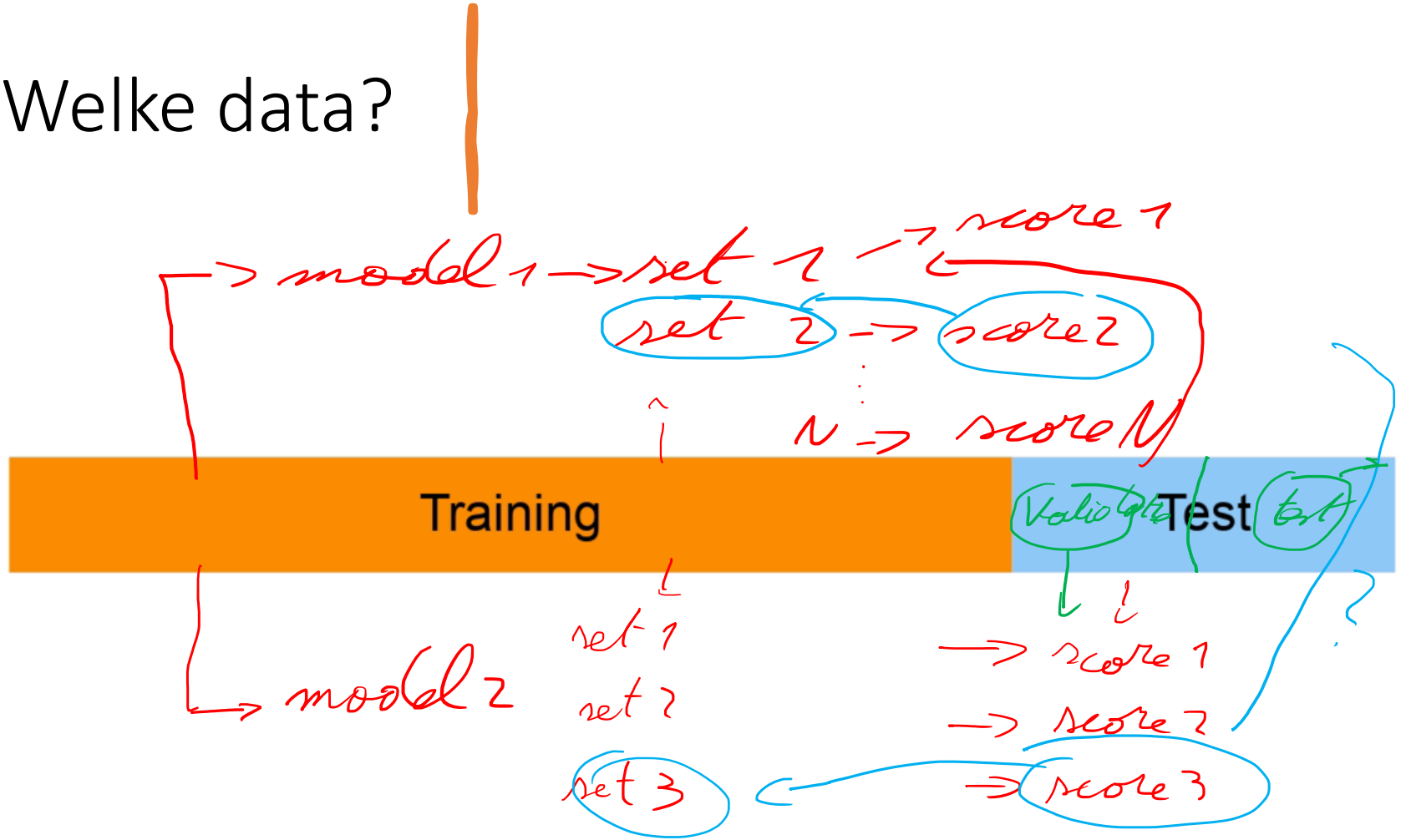
Wat is de beste configuratie?

Overloop alle mogelijke combinaties

Kies het model met de hoogste score



Welke data?



Welke data?

Steeds evalueren op ongeziene data

- Model dat beste werkt op validatie daarom niet het beste op de testdata

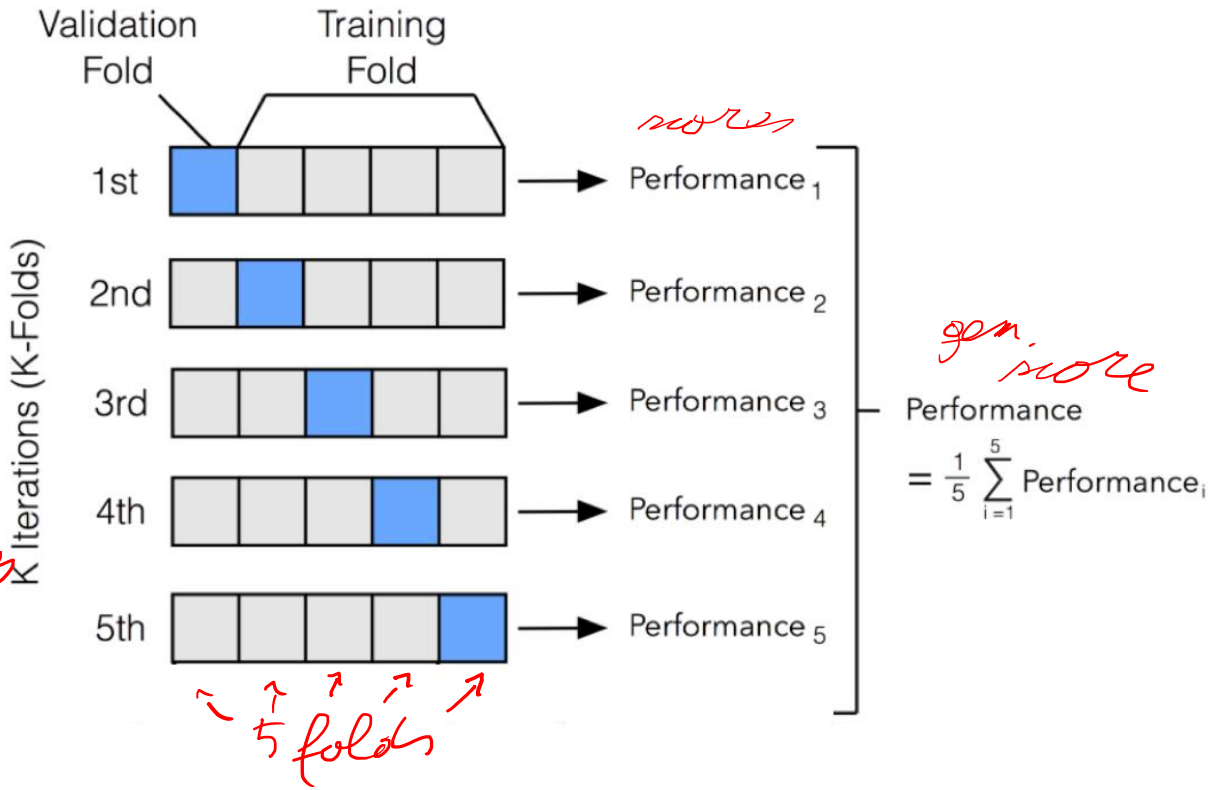
Daarom zowel validatie als test set nodig (holdout methode)

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

K-fold cross validation

+ → accurater model score

− → K trainingruns nodig



Hoe kiezen welke techniek het beste is?

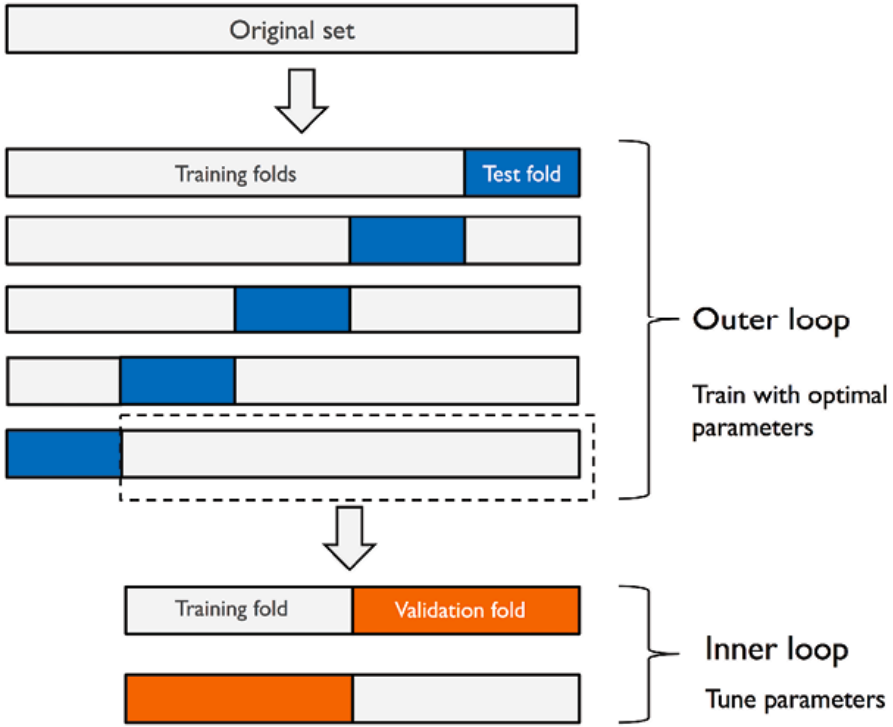
- K-fold cross validation -> ga op zoek naar het beste model voor 1 techniek

- Zelfde concept kan ook toegepast worden op meerdere technieken
  - Dus ook cross validation voor de testdata

  - Wordt nested cross validation genoemd

    - Voordeel: Betere beoordeling van de techniek
    - Nadeel: Meer fits/trainingen nodig dus dit duurt langer

Nested 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
- Meerdere lijnen die tot dezelfde persoon/klant behoren

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.

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!