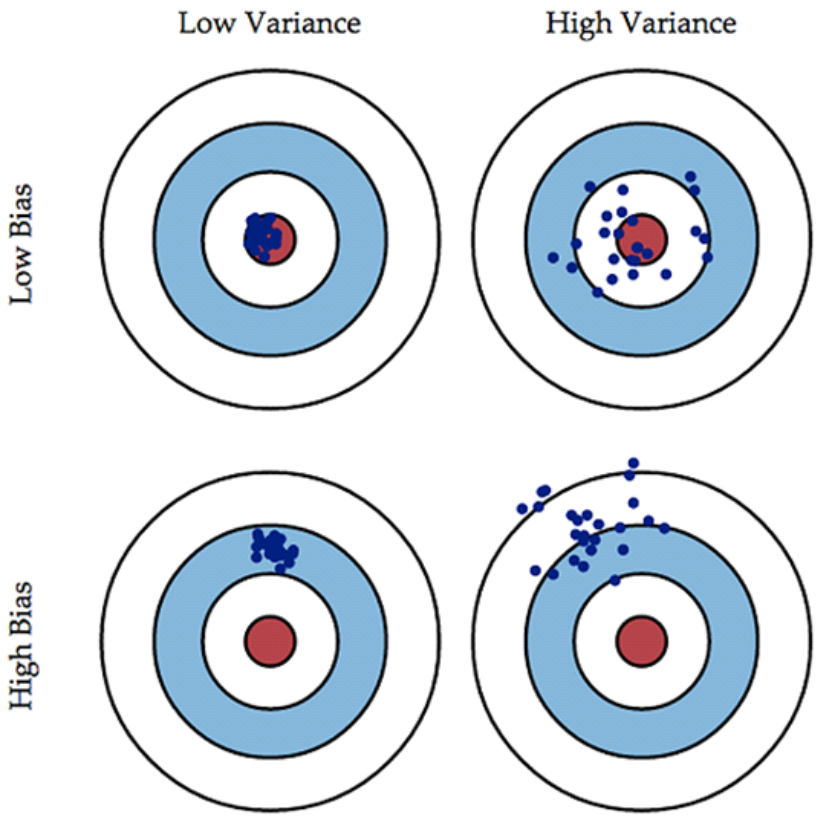
PowerPoints machine learning

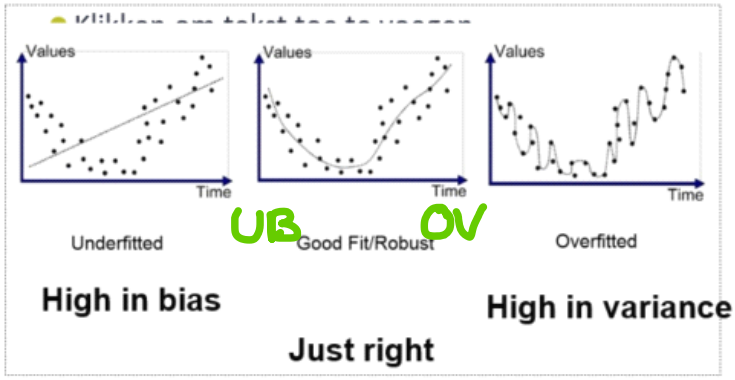
Sunday, 25 November 2018 16:28

Rules of thumb for sound machine learning

- Verify data quality
- Verify tussenresultaten
- Measure effectiveness using a ground truth on a hold out test set
- Compare against baseline
- Model selection (the simplest one is the best)

- Regression/continuous scale
  - Mean Absolute Error $\frac{1}{m}\sum_{i=1}^{m}(\hat{y}-y)$
  - Mean Squared Error $\frac{1}{m}\sum_{i=1}^{m}(\hat{y}-y)^2$





High variance – a.k.a. overfitting

- Causes for overfitting:
  - Too many features (e.g. classify 100MP images)
  - Not enough training data
  - Learning on a poor sample (not representative)
  - Doing model selection on the training set

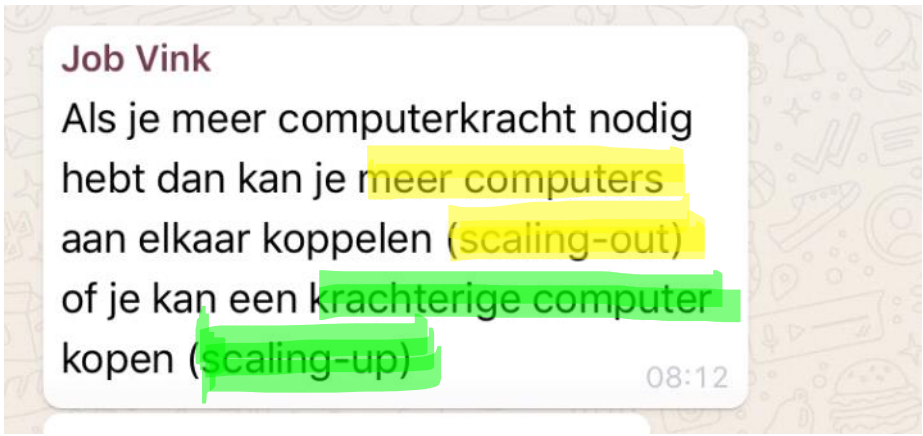
Regularisation

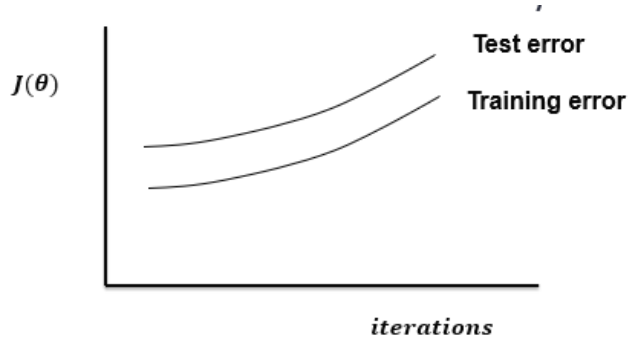
Eel penalty toevoegen

- Reduce risk of overfitting

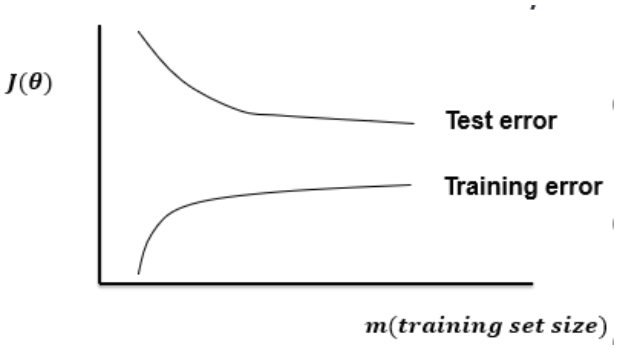
Feature Engineering

1. Check Data Edges
   a. Check number of rows and columns
   b. Check first few rows and last few rows
   c. Formatting ok? Are the values within realm of reality?
2. Variable Identification (Codebook)
   a. Per set: Where did data come from? How was data collected? Technical information about files. (How many, size, format)
   b. Per variable: Position, name, label, values, data type, numerical/categorical, predictor/target variable, summary statistics.
3. Univariate Analysis
   a. Check if it is a normal distribution
4. Bi-variate
   a. Check the correlations
   b. Plot the data in different graphs to understand the data
5. Missing Values
   a. Find NaN values and delete or replace them
6. Outliers
   a. Find outliers and delete or place them
7. Variable transformation
   a. Mean normalisation
   b. Find a group of outliers and change the scale or multiply everything with a log
8. Variable creation
   a. Change categoricals in numbers
9. Evaluation
   a. Check if the cleaned data has better results than the same model on the raw data. Use Mean Root Squared Error.



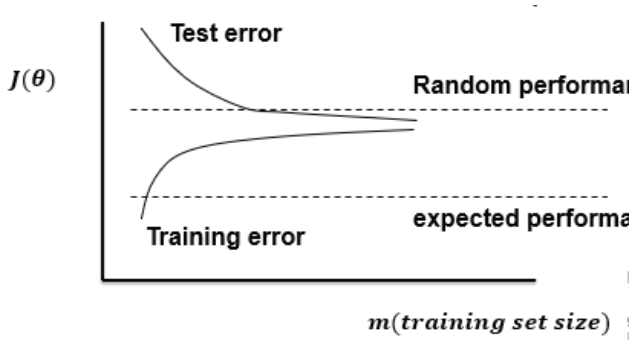


Fout, learning rate moet lager



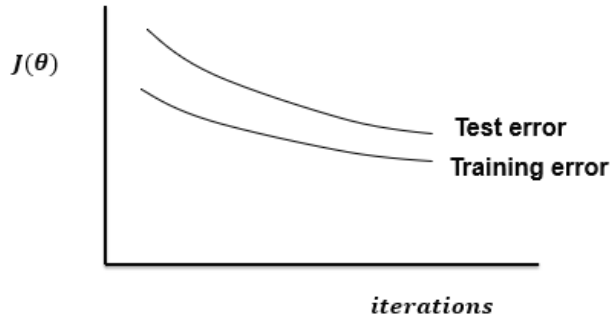
Fout, overfitting want traning error is veel lager dan test, (dus te veel getraind op training set).

- **Lambda moet dan hoger.**
- **Polynomials of features minders**
- **Verhoog traing size**

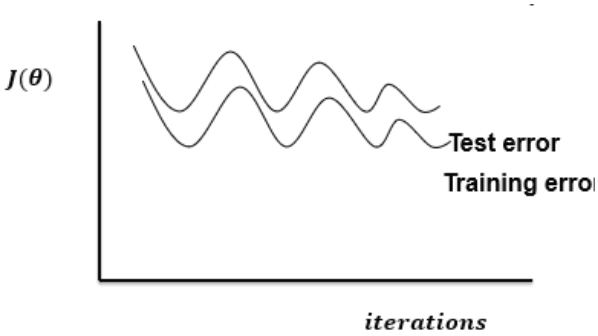


Fout, de lijnen moeten lager komen.
Underfitting, te gegeneraliseerd. (to high in bias)

- Lamba lager maken
- Meer features (wordt specifieker)

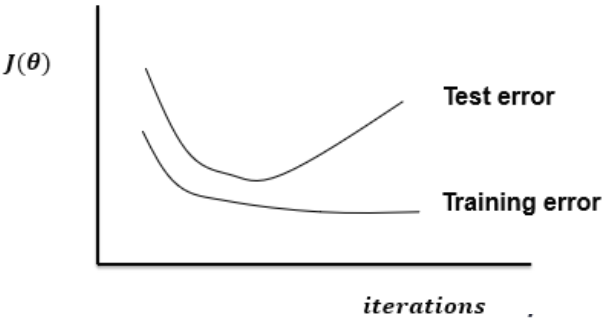


Goed



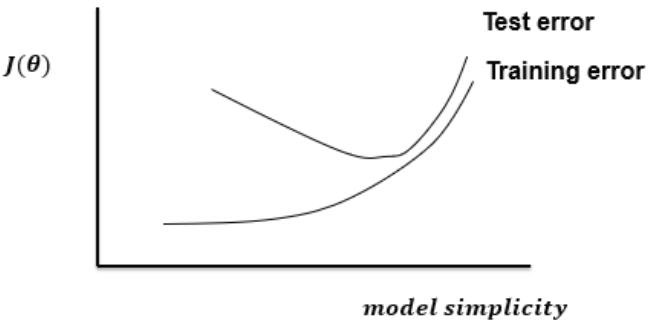
Fout, learning rate is te groot.

- Kleiner maken learning rate



Fout, overfit.

- **Get more training examples**
- **Try smaller set of features**
- **Try to increasing regularization**



Model complexity/simplicity tunen om te voorkomen dat het model high bias of high variance heeft. Dat kan bijvoorbeeld zijn door een juiste feature selection te maken, of een juiste order polynomial.