



Odisee
DE CO-HOGESCHOOL

Data Science – Week 8



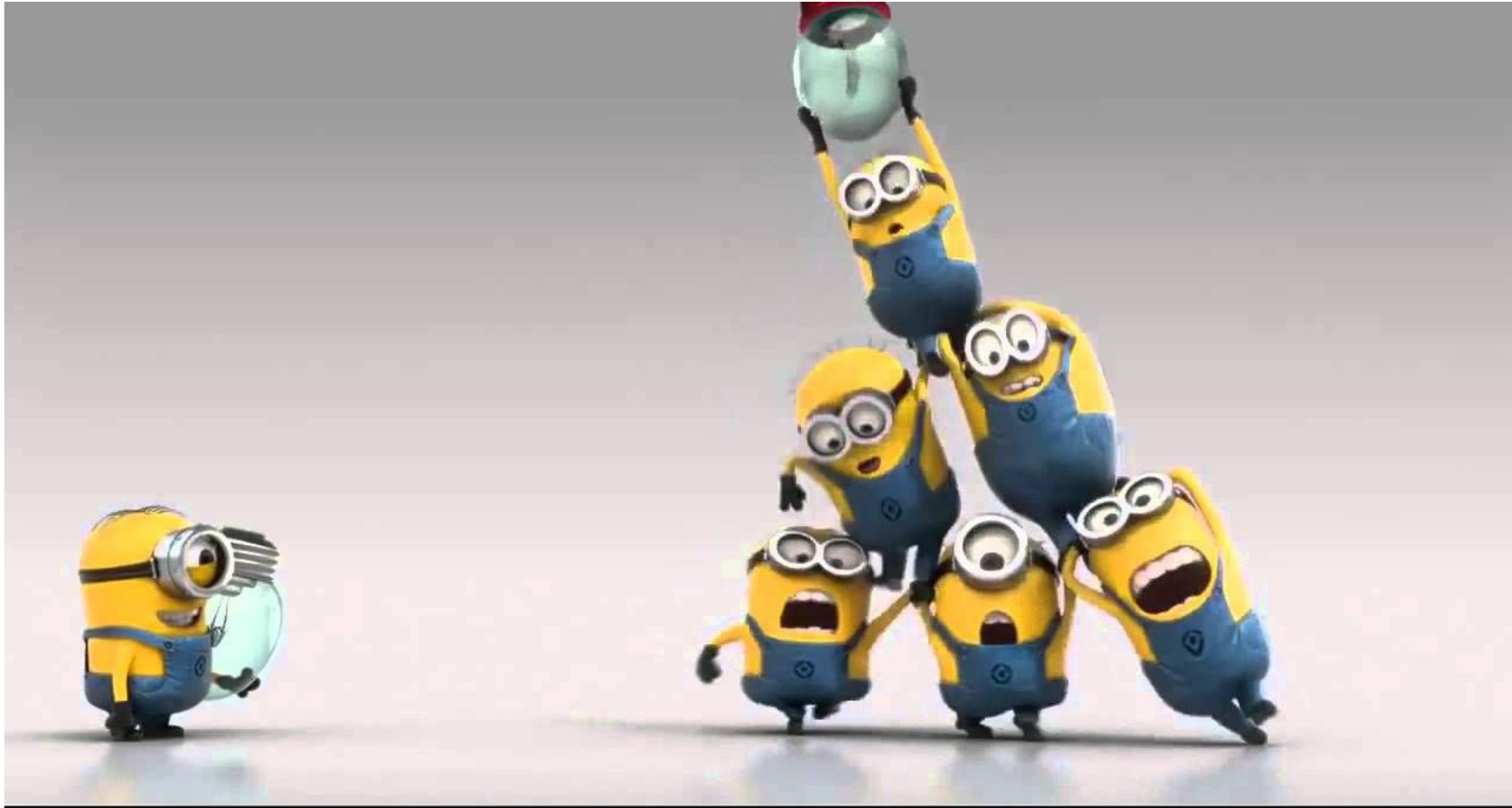


Recap vraagjes



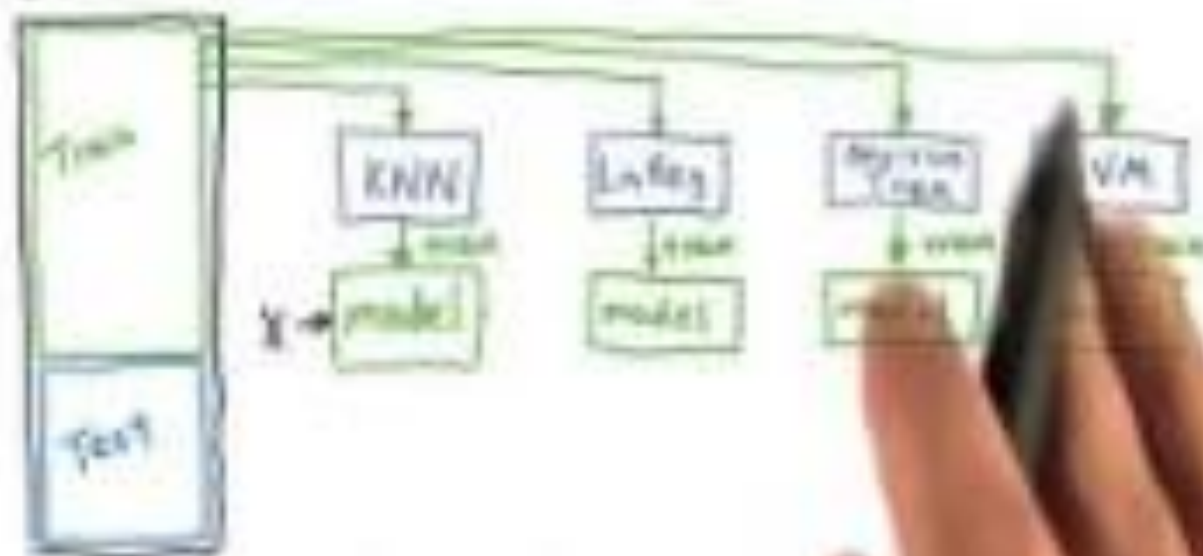
Ensembles





Ensemble learners

Data

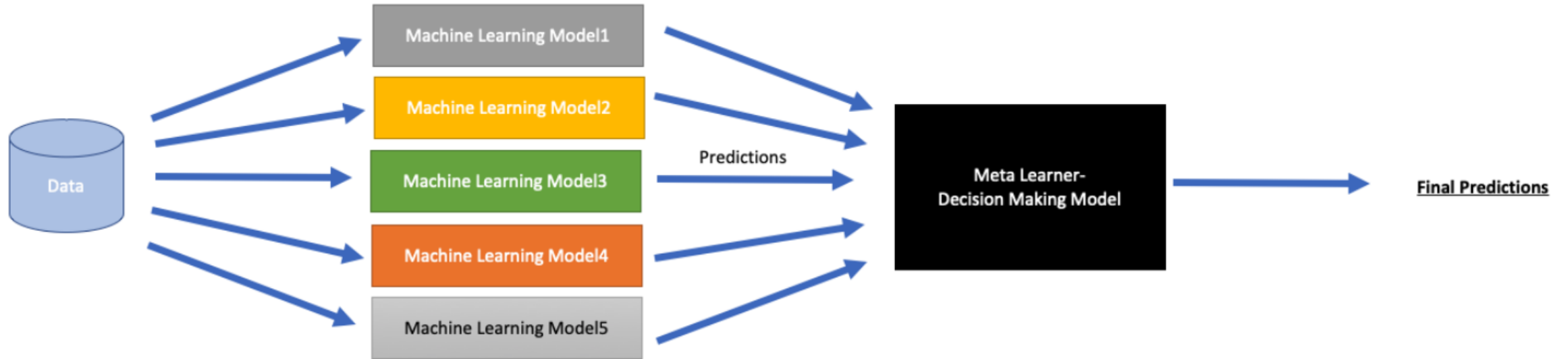




Kenmerken

- ▣ Meerdere modellen die samen de voorspelling maken
 - Kan voor zowel classificatie als regressie
- ▣ Meerdere technieken kunnen gecombineerd worden
- ▣ Doel is om de resultaten van elk individueel model te verbeteren

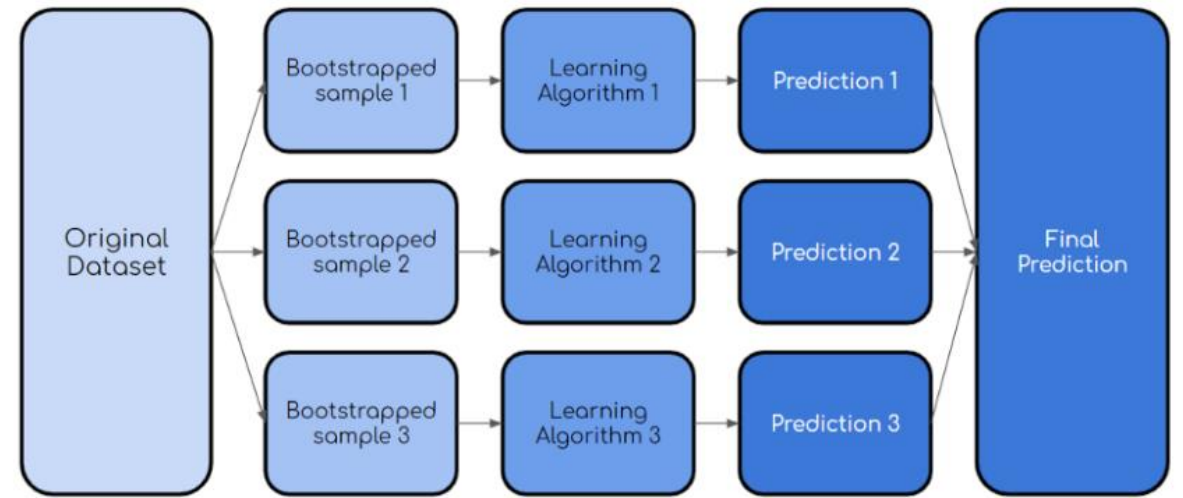
Stacking



- ▣ Decision making model: Majority voting, mean, apart model, ...
 - Kan ook werken op voorspelde klassen of op de kansen van elke klasse

Bagging of Bootstrap aggregating

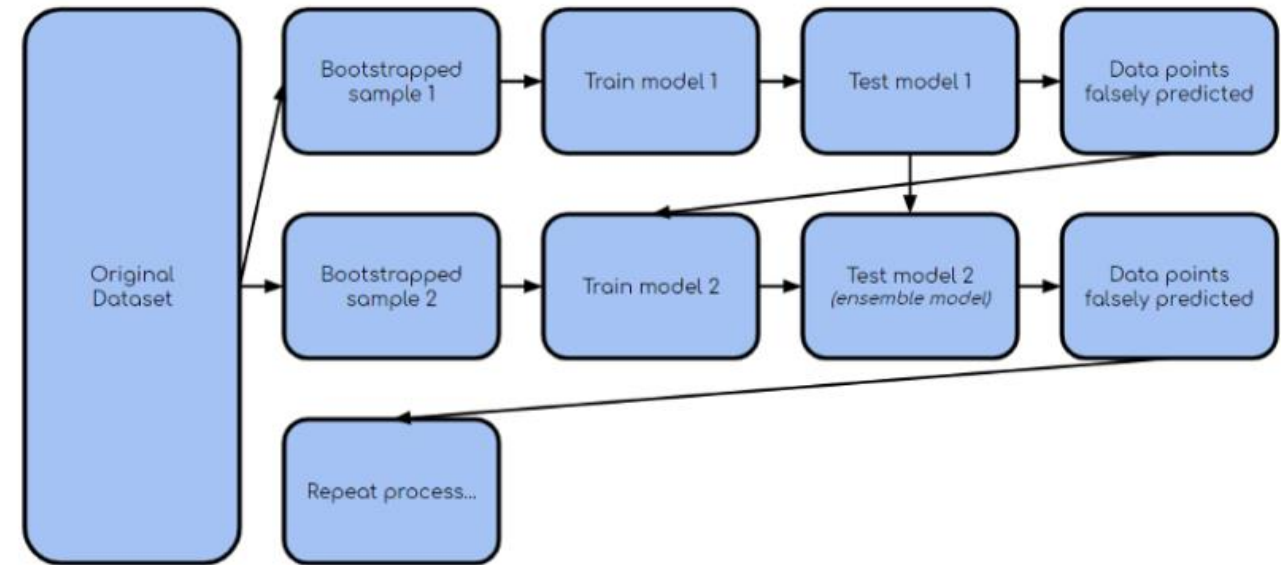
- ▣ Meerdere modellen
 - Zelfde techniek
- ▣ Train elk model met deel van de data
 - Vaak ongeveer 60%
 - Sampling data met terugleggen
- ▣ Finale predictie
 - Majority voting voor classificatie
 - Mean voor regressie
- ▣ Goed tegen overfitting



Boosting

■ Bagging maar

- ▬ Verkeerd geclassificeerde data komt terug als input van het volgende model
- ▬ Training sequentieel
- ▬ Prediction parallel
- ▬ Adaboost meest gekende





XGBoost

- ▣ Heel snel om te trainen
- ▣ Eenvoudig te implementeren en optimaliseren
- ▣ Haalt goede scores in competities
- ▣ Meer info: <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

When should I use XGBoost?

EXTREME GRADIENT BOOSTING WITH XGBOOST



Deep learning
with TensorFlow



Overview

Features of XGBoost

- Regularized boosting (prevents overfitting)
- Can handle missing values automatically
- Parallel processing
- Can cross-validate at each iteration
 - Enables early stopping, finding optimal number of iterations
- Incremental training
- Can plug in your own optimization objectives
- Tree pruning
 - Generally results in deeper, but narrower, trees





Natural Language Processing

Is dit spam of geen spam?

■ Hoe zou je hier een model voor opstellen?

- Welke features?
- Data cleaning?

Attention My Dear,

You have been a lucky winner of \$3.2MILLION from western union west Africa continent as value customer who use western union to transfer money from one country to another. Right now your first payment of \$5000 Is about to send today through western union money transfer You are advise to Contact Mr Peter Charles with your full information.such as

Your name.....

Your country.....

Your phone number..

Your adders.....

To Enable him send your First Payment of \$5000 today.For more information contact Mr Peter Charles. Telephone number: +229 63012985. Email (wu293780@gmail.com) he will keep sending your payment until your total fund is Completed \$3.2MILLION.usd

Best Regards

Mr.Don Alex



Feature Engineering

- ▣ Mail omzetten naar een set van features
 - Features = woorden in de mail
 - Gebruik deze voor classificatie
- Zijn de woorden afhankelijk van elkaar?
 - ▣ Komen woorden vaak samen voor in tekst?




Feature Engineering

- ▣ Mail omzetten naar een set van features

- ▬ Features = woorden in de mail
- ▬ Gebruik deze voor classificatie

- ▬ Zijn de woorden afhankelijk van elkaar?

- Komen woorden vaak samen voor in tekst?
- Ja, maar we veronderstellen dat ze onafhankelijk zijn (Naïeve Bayes veronderstelling)


$$P(\text{Spam}|w_1, w_2, \dots w_n) = \frac{P(w_1, w_2, \dots w_n | \text{Spam}) P(\text{Spam})}{P(w_1, w_2, \dots w_n)}$$

Regel van Bayes



$$P(\text{Spam}|w_1, w_2, \dots w_n) = \frac{P(w_1 | w_2, \dots w_n, \text{Spam}) P(w_2 | w_3, \dots w_n, \text{Spam}) \dots P(\text{Spam})}{P(w_1, w_2, \dots w_n)}$$

Veronderstel onafhankelijkheid



$$P(\text{Spam}|w_1, w_2, \dots w_n) = \frac{P(w_1 | \text{Spam}) P(w_2 | \text{Spam}) \dots P(w_n | \text{Spam}) P(\text{Spam})}{P(w_1, w_2, \dots w_n)} = \frac{P(\text{Spam}) \prod_{i=1}^n P(w_i | \text{Spam})}{P(w_1, w_2, \dots w_n)}$$

Noemer niet nodig



$$p(\text{Spam}|w_1, \dots, w_n) \propto p(\text{Spam}) \prod_{i=1}^n p(w_i | \text{Spam})$$

Classificatie

- Dataset met 300 spam mails en 850 ham mails:

Woord	Spam Frequentie	Spam Kans	Ham Frequentie	Ham Kans
customer	100	0.33	200	0.24
advise	50	0.17	70	0.08
Africa	120	0.4	30	0.03
money	60	0.2	450	0.53
number	180	0.6	550	0.65

- Welk percentage mails is spam?
- Doe een manuele classificatie van de zin “Africa advise money”

$$\text{SPAM} = \frac{300}{300 + 850} \quad \times 0 \quad \times 0 \quad \times 0 \quad = A$$

$$\text{HAM} = \frac{850}{850 + 300} \quad \times 0 \quad \times 0 \quad \times 0 \quad = B$$

Classificatie

```
pSpam = 300 / (850+300)
pHam = 1 - pSpam
print(pSpam, pHam)

pTextIsSpam = pSpam * 0.4 * 0.17 * 0.2
pTextIsHam = pHam * 0.03 * 0.08 * 0.53

if(pTextIsSpam > pTextIsHam):
    print("Africa advise money is spam")
else:
    print("Africa advise money is ham")
```



Probleem – nieuwe woorden

- ▣ Classificatie van de zin “Europe advise money”: $p(w) = 0$
 - Hoe dit classificeren?

Probleem – nieuwe woorden

- ▣ Weglaten van de ongeziene woorden
 - = Weglaten van informatie
- ▣ Laplacian Smoothing (voeg “fictieve” trainingsdata toe)

$$P(w) = \frac{C(w) + \alpha}{N + \alpha V}$$

- ▣ α is hyperparameter:
 - kleine waarde = neiging tot overfitting
 - grote waarde = neiging tot underfitting



Probleem – nieuwe woorden

- ▣ Classificatie van de zin “Europe advise money”
- ▣ Bereken nieuwe matrix met de kansen
- ▣ Bereken de kansen van spam of niet spam voor bovenstaande zin

Probleem – Floating point underflow

- ▣ Classificatie door kansen vermenigvuldigen

- => Resultaat steeds kleiner

- => Gevaar op underflow

- ▣ Neem het logaritme om dit te voorkomen

$$\log(P(\text{Spam}|w_1, w_2, \dots w_n)) \propto \log(P(\text{Spam})) + \sum_{i=1}^n \log(P(w_i|\text{Spam}))$$



Probleem – Gelijkaardige/nutteloze woorden

▣ Niet alles dat in een mail zit is nuttige informatie:

- Html tags
- Cijfers, leestekens, speciale symbolen, ...
- Hoofdletters
- Stopwoorden
- Vervoegingen, werkwoorden, ...
- Te korte woorden

Tekst omzetten naar feature vector

▣ Via bag of words

- ▬ Multi-hot
 - Bijhouden of het aanwezig is in de tekst of niet (0 of 1)
 - Meerdere woorden op 1
- ▬ Aantal keer dat het woord in de mail aanwezig is
 - Count
- ▬ Term frequency – inverse document frequency
 - Verlaag impact van woorden die in veel mails voorkomen

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_i}\right)$$





Huiswerk / zelfstudie





Studeren voor Test

- ▣ Theoretische test gepland op 19 december 2023 om 13:30
 - ▣ Alle theorie van alle lessen
 - ▣ Op papier, geen code te schrijven
 - ▣ Gesloten boek, enkel stylo en studentennummer nodig voor de test
-
- ▣ Voor Microdegreestudenten: Als het niet lukt om je vrij te maken, kan er iets ingepland worden op een avond in die week.



Zelfstudie

- ▣ De volgende cursussen zijn NIET te kennen leerstof voor de test:
 - <https://www.kaggle.com/learn/intro-to-ai-ethics>
 - <https://www.kaggle.com/learn/machine-learning-explainability>

Oefening

▣ Maak de oefening via volgende link:

▬ <https://classroom.github.com/a/olg830fU>

▬ Bevat twee delen

■ NLP

-> Dit deel kan je al maken

■ Unsupervised learning

-> Dit deel zie je in de volgende weken



Project

- ▣ Vergeet het project niet!
 - Wie is er al aan begonnen?
 - Iedereen reeds een groep?