Week 1
Analytics continuum
Analytics:
Descriptive – What happened? (wat is er gebeurt?)
Diagnostic – Why did it happen?(waarom gebeeurt)
Predictive – What will happen?(wat gaat er gebeuren)
Prescriptive – What should I do?(wat moet ik doen)

4 V's
Volume – Data at rest(data op zn plek)(hoeveelheid)
Velocity – Data in motion(data beweeegt)(snelheid)
Variety – Data in many forms(data is verschillende vormen)
Veracity – Data in doubt(waarheid van data)

Uncertainty important
- Data factual, answering straightforward(Feitelijke gegevens, eenvoudig antwoord)
- Data contains uncertainty almost always(Gegevens bevatten bijna altijd onzekerheid)
- Need statistics to estimate possible solutions (Statistieken nodig om mogelijke oplossingen te schatten)

Manual engineering becomes infeasible when(handmatig engineering wordt onhaalbaar als)
- Data volume grows(data hoeveelheid groter wordt)
- Data complexity grows (many features)(veel features)
- There is less time to develop(er is weinig tijd)
    - Machine Learning!(Als al deze dingen gebeuren dan moet je machine learning doen)

Data science requires light programming skills
- Python:
    - Easy to learn
    - Efficient to code in
    - Supports functional programming (required)
    - Hooks to fast libraries for Data Science
    - Interactive Notebooks

Week 1 is goed

Week 2 introduction to machine learning

Wat is machine learning ?

Machine learning is elk proces waarbij een systeem de prestaties uit ervaring verbetert, een grote verschil met normal programming is dat machine learning zich zzelf verbeterd door middel van observatie. ML kan zelf onderscheiden het verschil tussen bijvoorbeeld verschillende getallen. Als je '4 en 4' hebt en '5 en 5' dan zegt de systeem zelf welke 4 is en welke 5 door middel van ML.

'Learning Task(T) is elk proces waar een system performance verbeterd gemeten(P) door ervaringen (E) .

Stel dat T(taak) is het voorspellen van huis verkoop prijzen

Experience is dat er een zelfde soort huis is voor 250k en een ander zelfde soort voor 200k dan voorspelt ML dat de huis verkoop prijs in die prijs klasse hoort) die voorspelling gebeurt dmv performance en experience uit het verleden

Ander voorbeeld

T= je wilt je spam filteren

P=percentage van spam/ham zijn er uitgefilterd

E= mails dat zijn gelabeld als spam/ham

voordelen ML

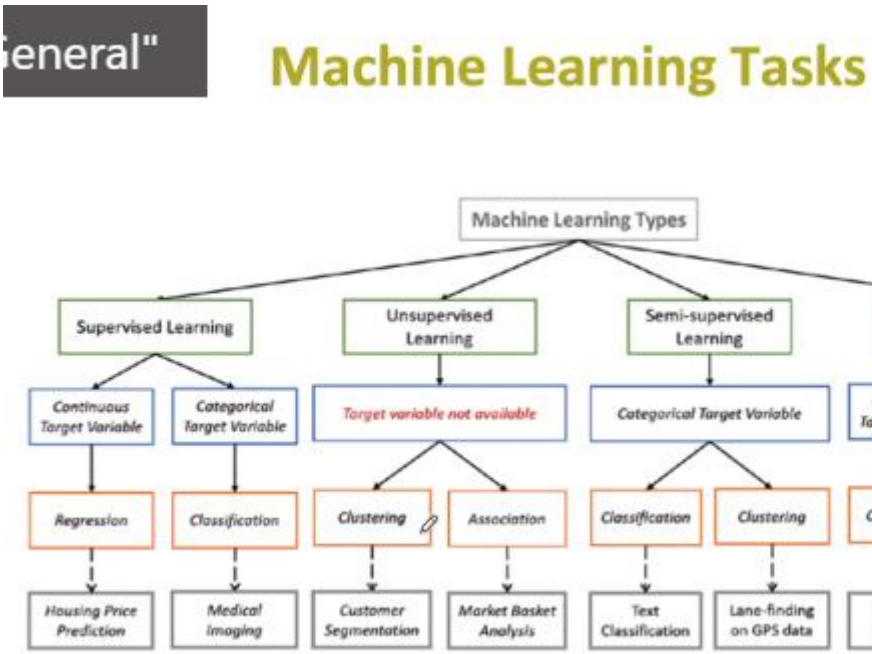
-past zich automatisch aan naar individuelen en veranderingen

-ontdekt nieuwe kennis

-imiteert menselijk intelligence(stel je gaat opzoek naar film dan weet de machine op wat je let en geeft je daarop aanbevelingen)

-menselijk engineering is soms heel moeilijk

Machine learning types



Supervised learning gebruikt een dataset wat een ground truth bevat.

ground truth is where you have actual labels that are been recorded about what the correct outcomes for certain experiences are….. dus in de dataset staan de verschillende huizen prijzen waarop de ML gebruik maakt op de nieuwe huisprijs te voorspellen

onder supervised valt regression dit houdt in dat je een tabel hebt met allemaal huis prijzen en er een lijn door het midden gaat, waarbij je kan concluderen dat hoe groter de oso hoe duurder de prijs dat doet dan de machine ook. Ook voorspel je hier een waarde

Onder supervides valt classification ook dit houdt in dat er voorspeld wordt welke classifiactie

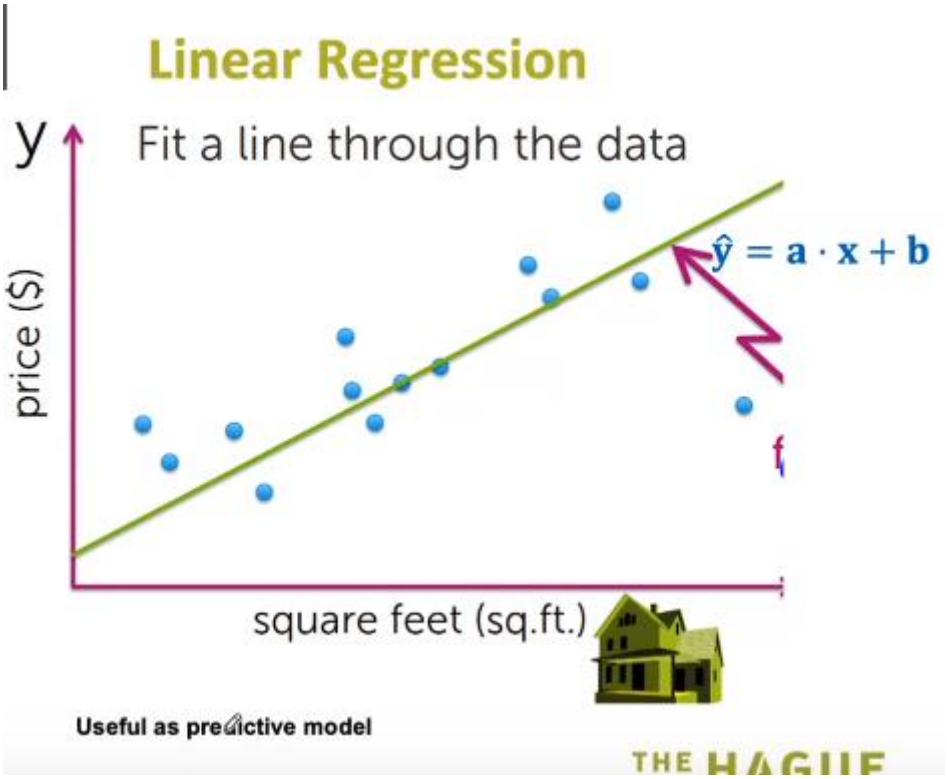
unsupervised heeft wel een dataset maar geen ground truth waardoor er geen correcte waardes beschikbaar zijn, dan komt clusteren van toepassing waarbij de specificaties van een oso worden vergeleken met de andere osos en in die ‘cluster’(groep) worden geplaatst. Een ander is is association waarbij er word gekeken welke producten er samen worden gekocht

semi supervised leraning je hebt een dataa set maar een geringe grounf truth

reinforcement leraning is wat een automatisch stofzuiger heeft, het heeft geen map van mijn oso maar leert van zelf hoe het eruitziet en past zich daaraan

liniear regression

check onder die tabel en dan zet je lijn in het midden en de machine voorspelt dan een prijs voor een nieuwe oso door middel van de ‘ecperience’ van andere huizen. Dus als er een oso is met een bepaalde afmeting dan kijk je naar de lijn die verteld dan als het ware wat de prijs voor de huis is



Echtt prijs is Y
Voorspellende prijs is Y met dakje

Y as is target variable

X is het getal voor sq2

b is waar de lijn begint

a is hoe stijl de lijn moet gaan

machine learning & linear regression

hypothese: we kunnen de waarde nauwkeurig voorspellen van een target(in dit geval de huisprijs) met een feature(in dit geval groote van het huis) door middel van een linieare regressie

(Response variable wordt ook wel target genoemd

Explanatory varablie wordt ook wel feature genoemd)

leer een linear regressie functie voor het voorspellen

meest gebruikte functie voor voorspellingen

Jupyter notebook is handig voor data scientist, why?

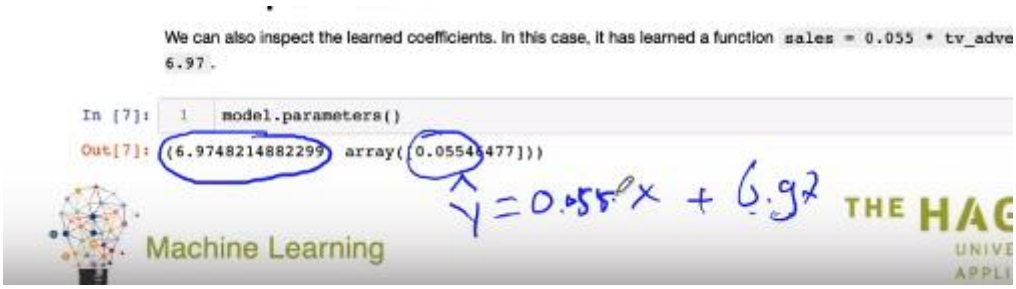
Je werkt interactief(korte script, direct inzicht)

Visueel

High level view on ML

-preparing Data : what data do we have?
-Model selection: What is our task? What is a suitable model?
-Train
-Evaluate

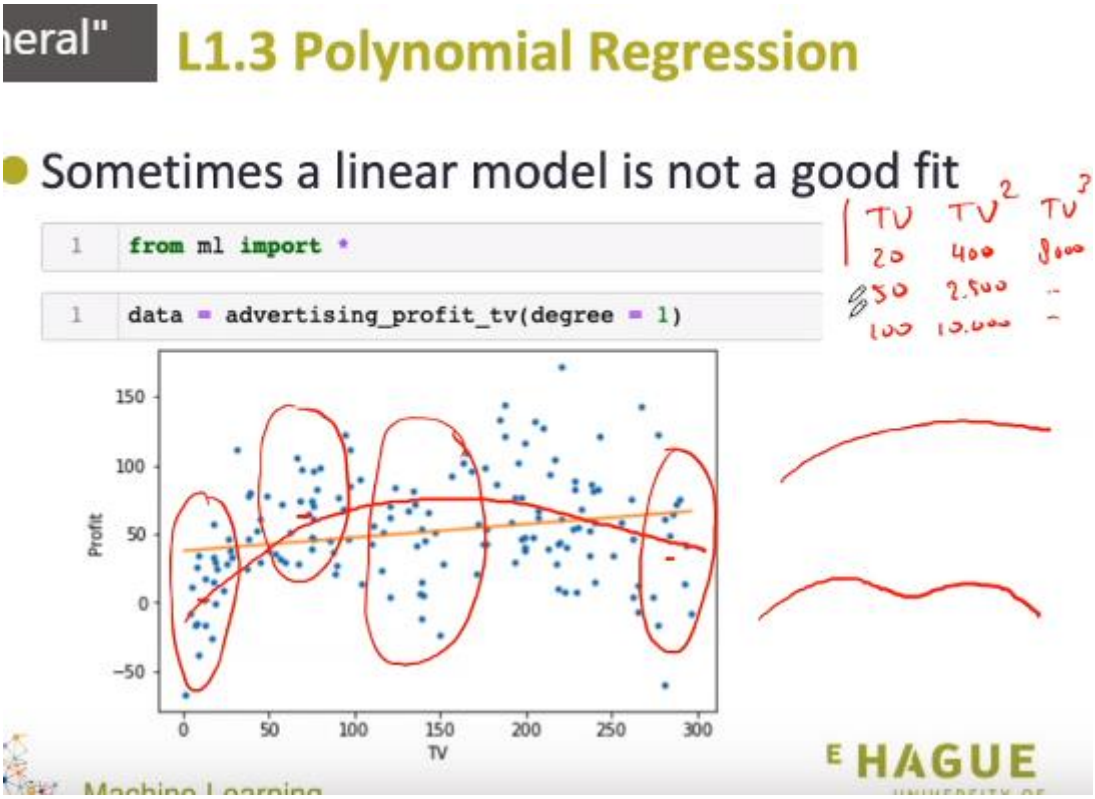
Hoe goed een lijn is gefit noemen de loss/cost function

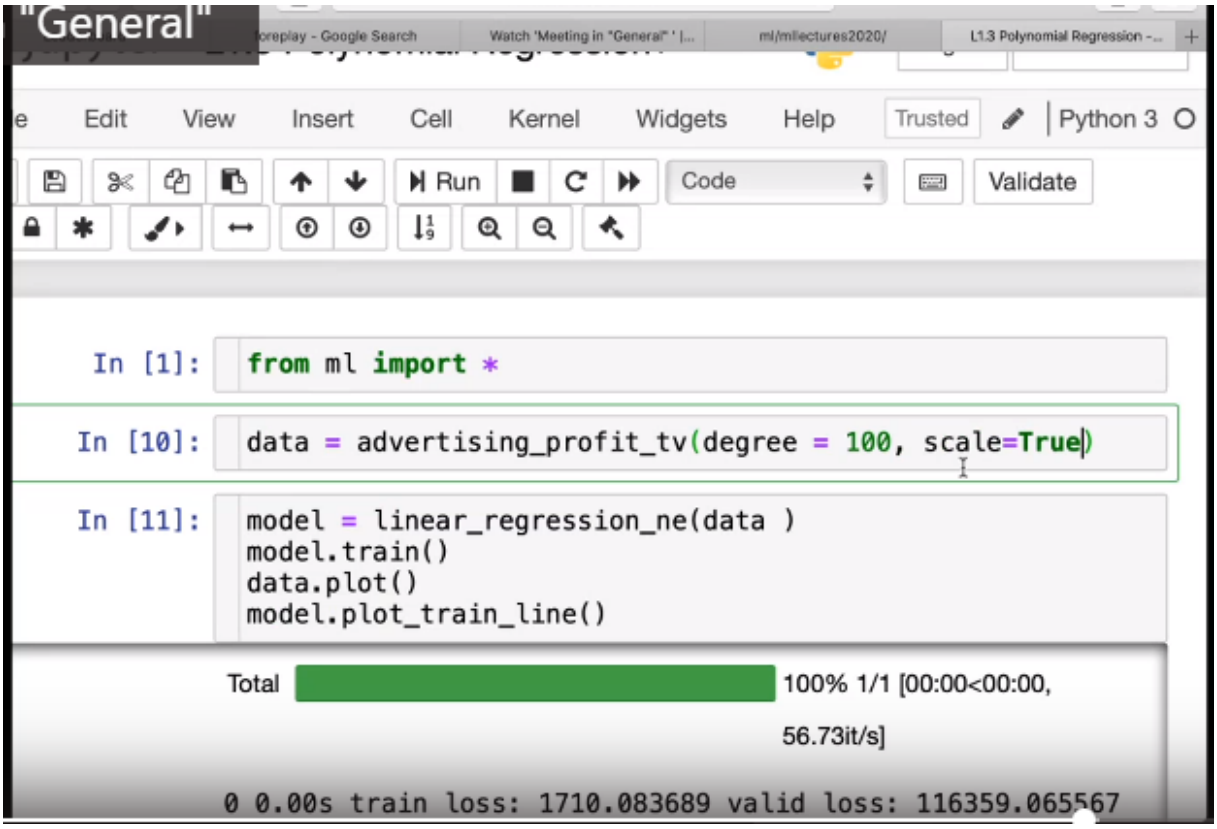


Bij multivariate regresion veranderd alleen stap 1 Data dit is anders bij liniear regression

L1.3 polynoom regressie

Om een polynoom parabool te krijgen in je grafiek moet je x/\2 of x/\3 gebruiken en fit het als een multivariate regression



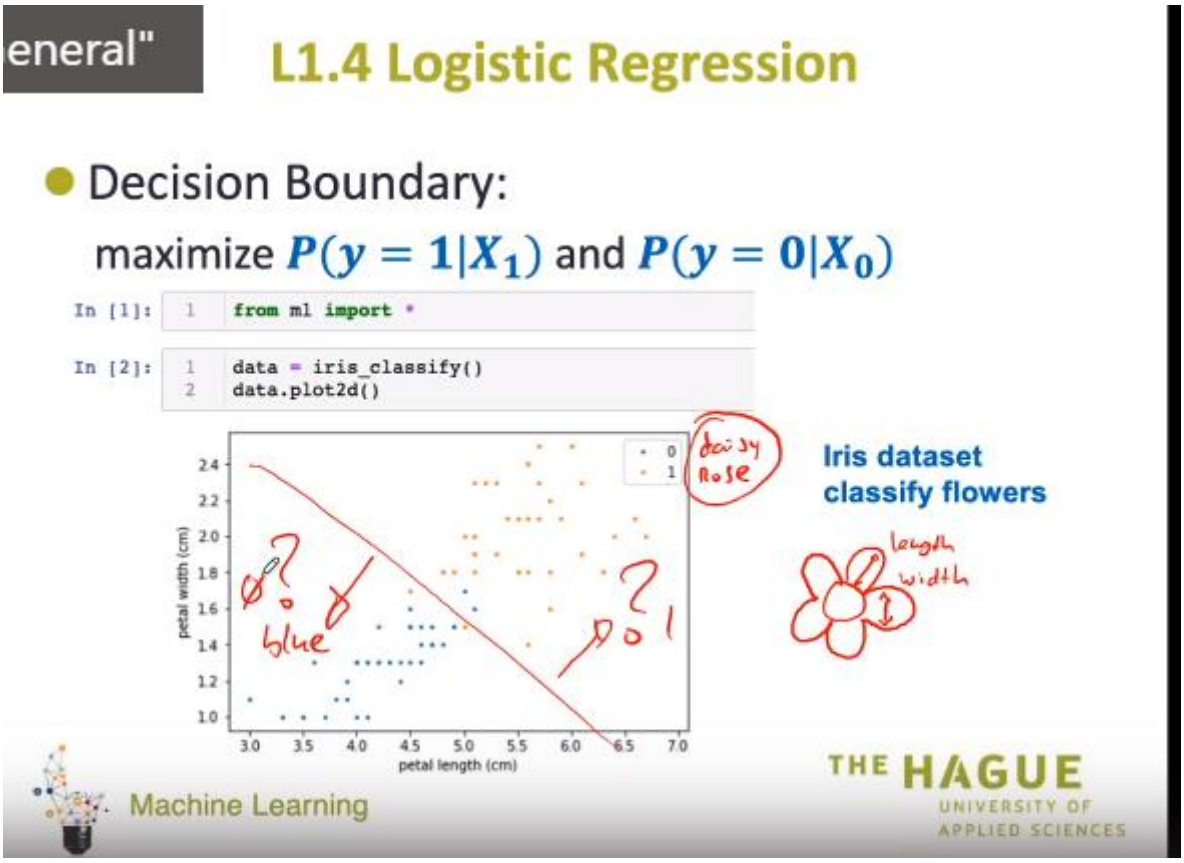


(Met feature scaling bypass je het probleem met grote getallen)

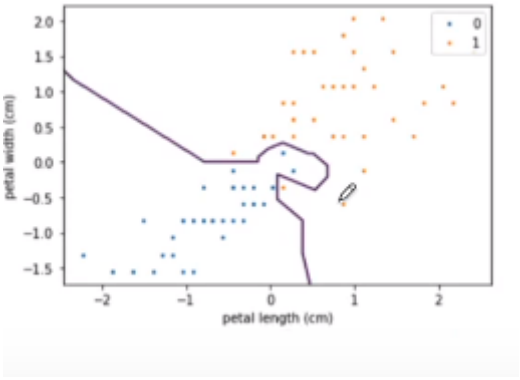
Gebruik scale= True om rare lijnnen te voorkomen(het is wel gedetaillerd (precies lijn))

Wat is een verschil tussen een regressie en classificatie?

bij classificatie gaat het om 0 en 1 en dan heeft 0,3 geen nut



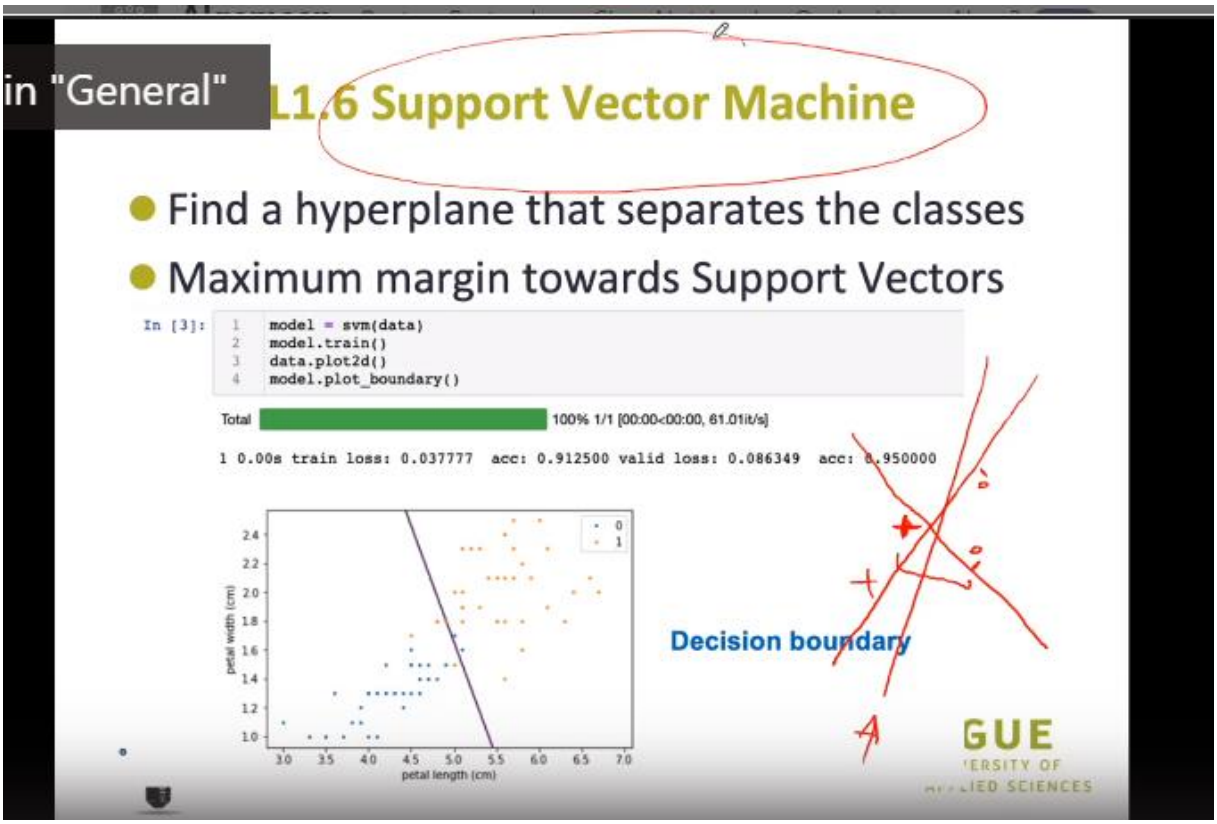
Hier gaat het om 0 en 1 dus je trekt een zogenoemde decission boundry lijn en als je ‘ding’aan de kant van blue is is het 0 en als je ‘ding aan de kant van red is is het 1



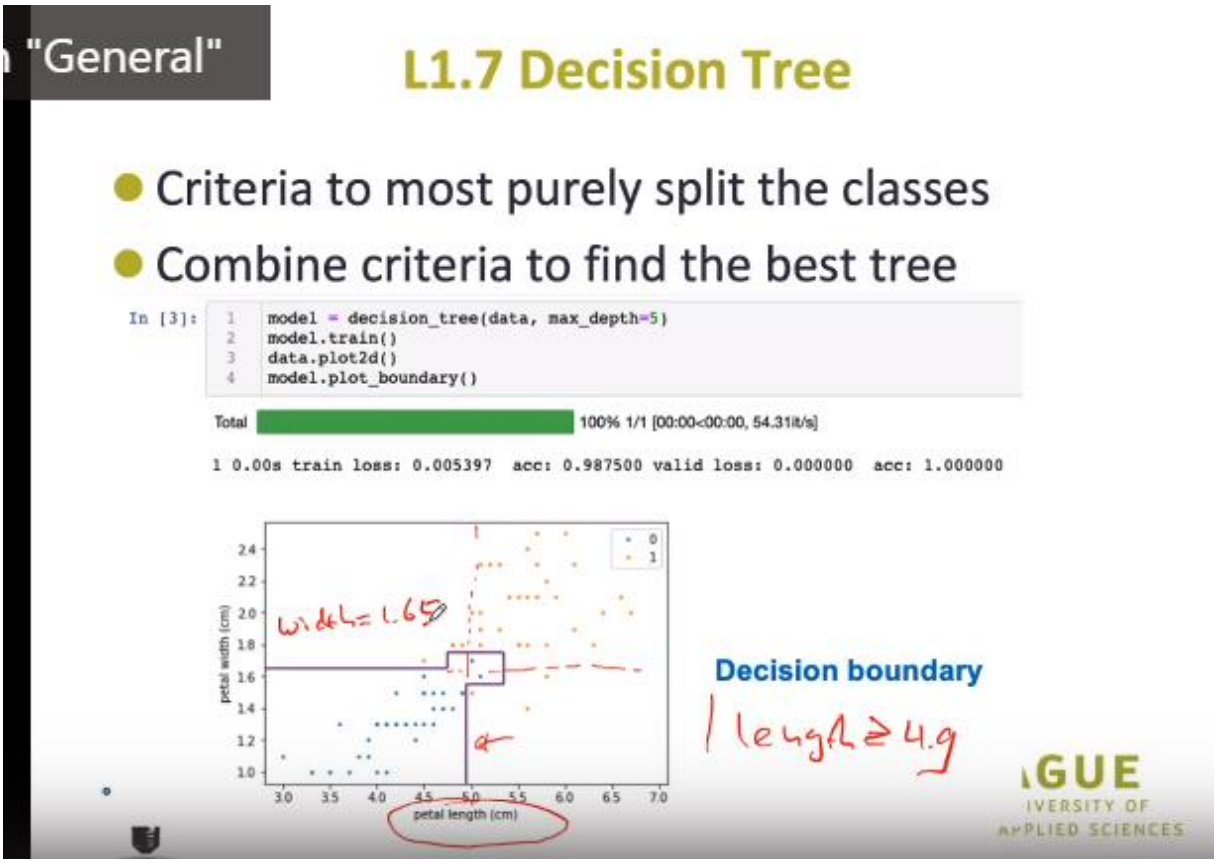
K nearest neighbor is de lijn deelt de verschillende kleuren van elkaar en als de punt dichtbij is bij een van de blauwe of gele hoort het bij die puntje

Die lijn is precies het midden tussen 1 blauwe en 1 gele puntje en die lijnen connecten elkaar

Er is pricies in he tmidden van 2 bollen(blauw en oranje) een lijn getrokken en die lijn is verbonden met elkaar om zo de grenzen vast te stellen als je twijfelt waar het puntje in geplaatst moet worden\



Support vector machine is een lijn tussen de 2 die het gelijk verdeeld



Lijn die de verdeling uitmaakt dus voor ‘length = 4.9’is alles blauw en alles ernaa is geel

Dat geld voor elke hoek lijn in de grafiek

In deze foto zijn er 6 decisions(elke lijn is 1 decision)

Week 3 model estimation

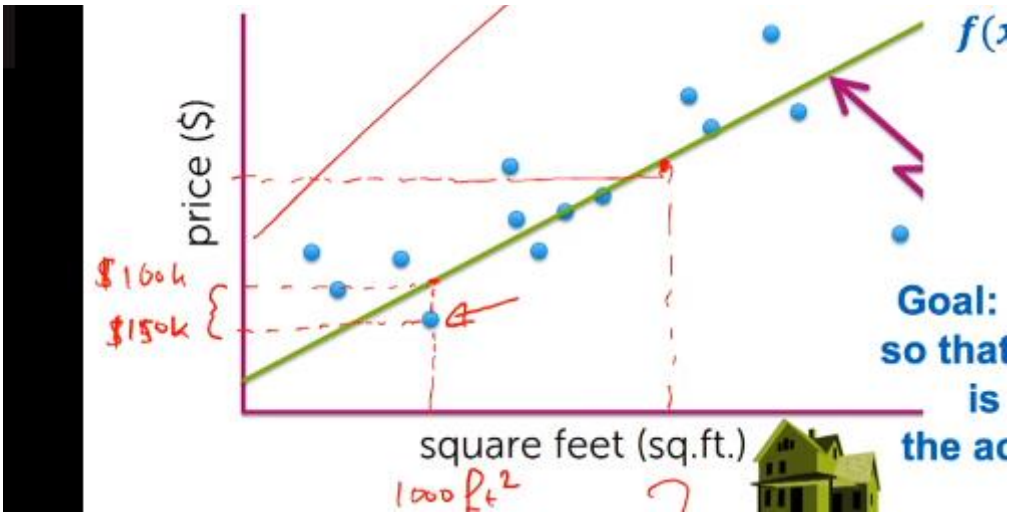
Hoofdonderdelen van supervised Learn algoritmes

Regression : schat de relatie tussen een doelvariabele(vb, price van huis) en kenmerken(vb,grootte en buren van variabele)

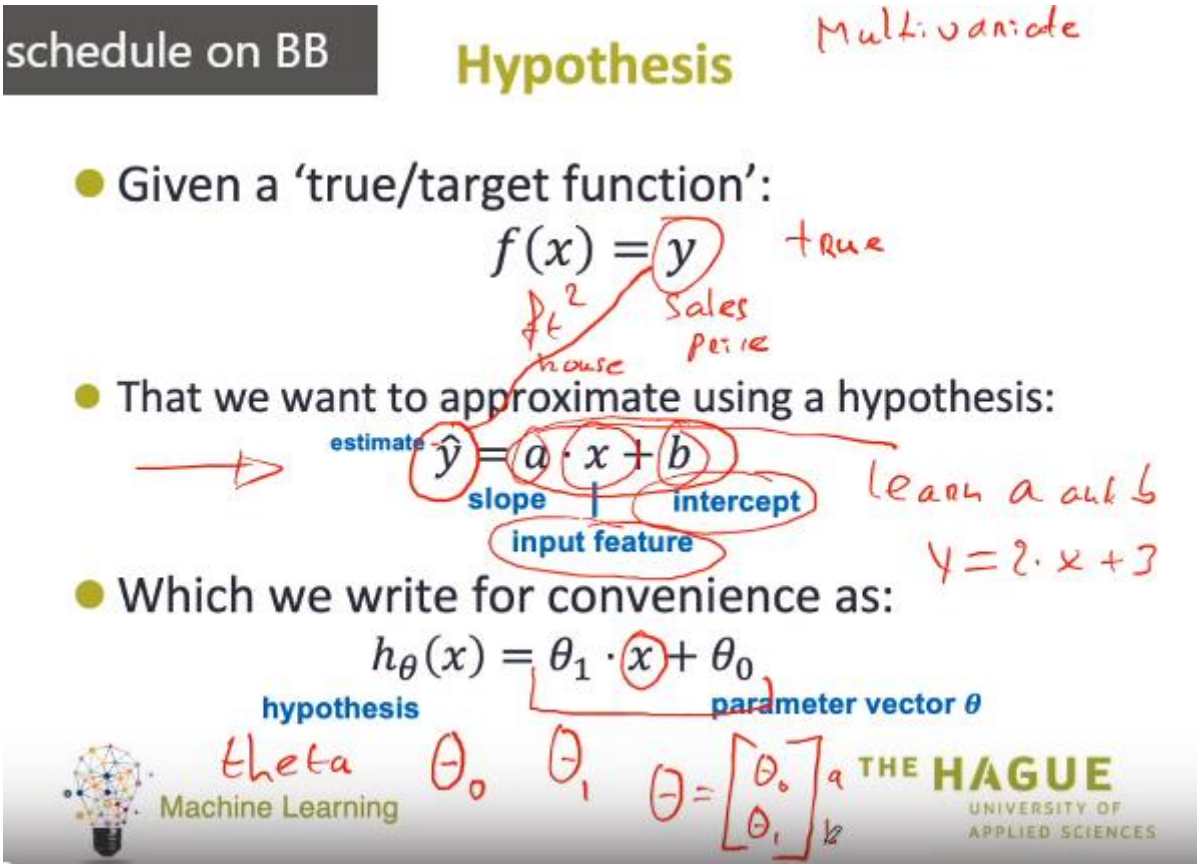
Classification : schat tot welke afzonderlijke klasse een waarneming behoort(soort bij groep)

Hoe kunnen we meten of een liniear regressie optimaal is?

Het verschil tussen de actuele stip en de lijn moet minimaal zijn dan is het niet perfect maar wel in de goede richting, maar als je kijkt naar een linieare lijn wat veel hoger is gezet dan is de lijn zeker niet optimaal omdat het verschil tussen de blauwe stip en de lijn te groot is.



Hypothesis

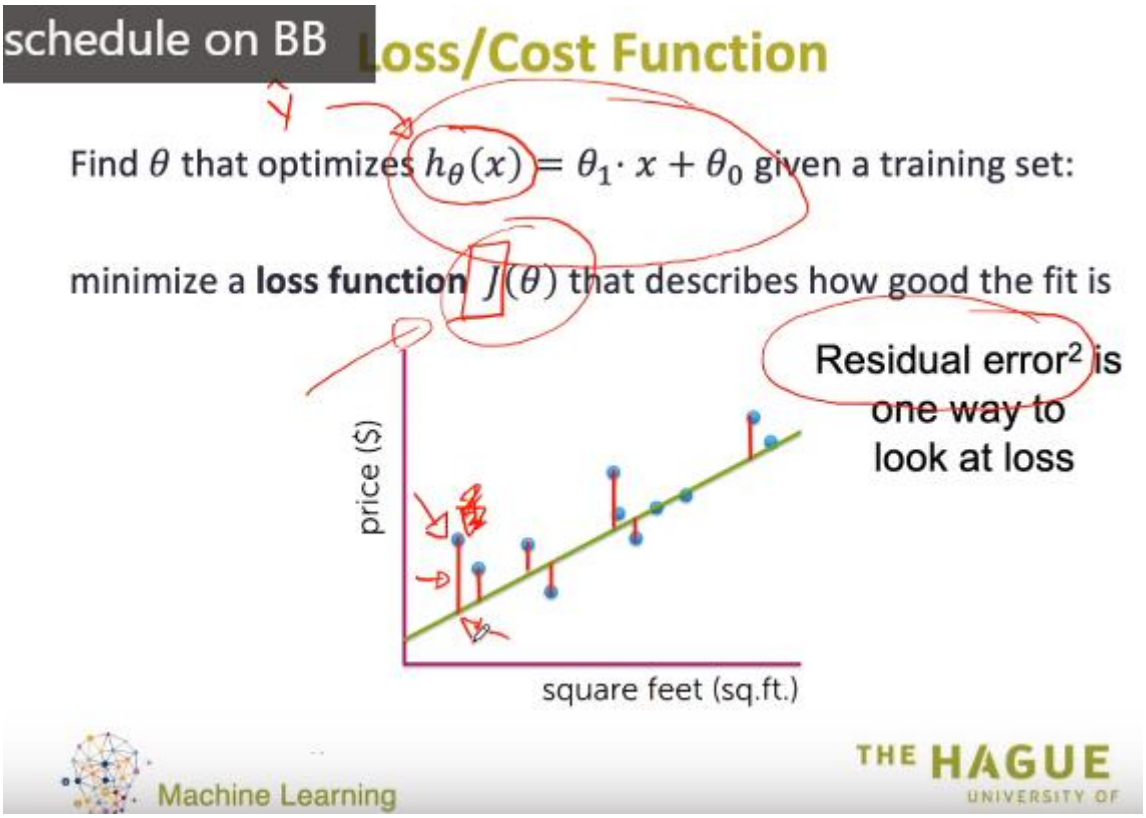


Het is belangrijk een hypothese op te zetten hiermee kan je veronderstellen dat je een goed voorspellende model hebt.

Punt 1 : dit is de gegeven functie(true value) dus voor zoveel vierkante meter heb je een huis voor deze prijs.

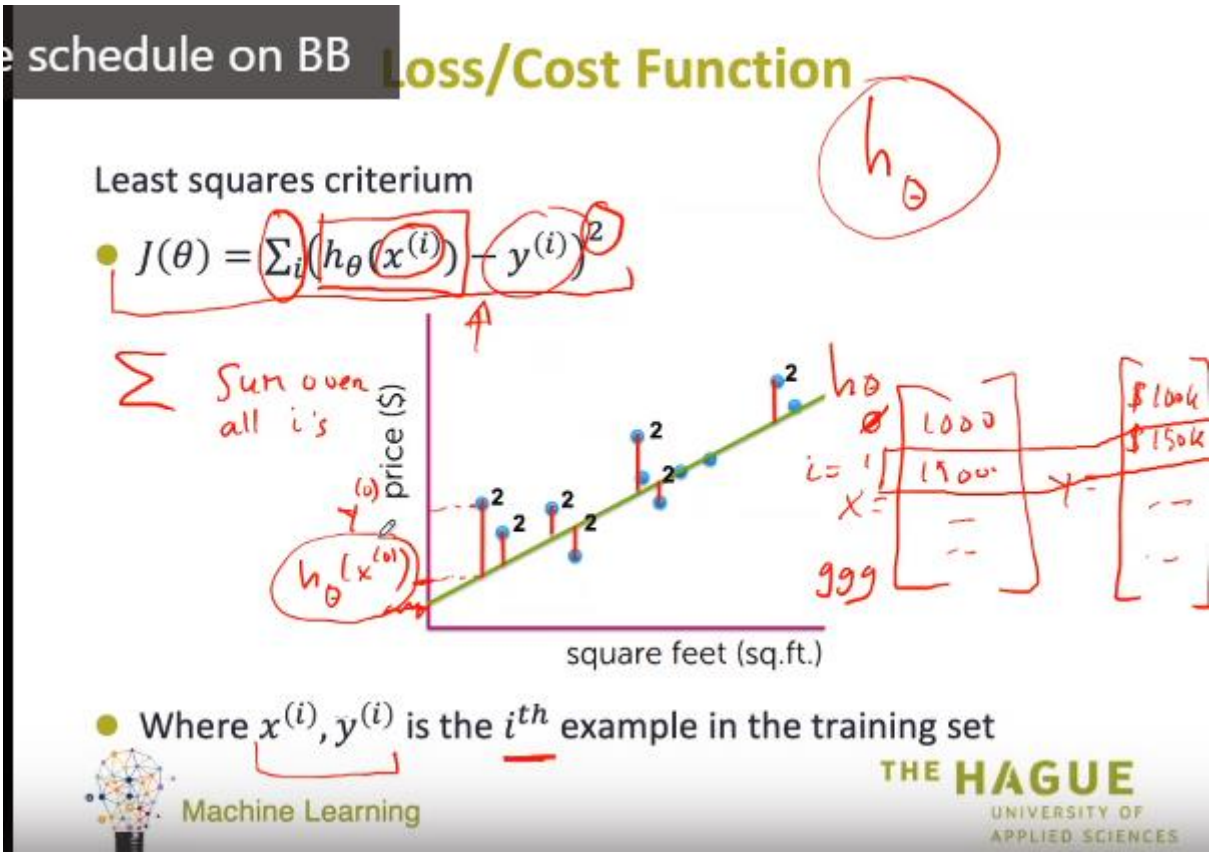
Punt 2 : deze functie is de voorspellende y waarde dus hoe de lijn dient te lopen(hierbij is het belangrijk A en B goed te laten leren om een optimaal model optzetten)

Punt 3 : Maar binnen ML schrijven we de functie bij punt 2 zo op omdat er binnen ML vaak een multivariabele voorkomt en als die groot zijn heb je niet genoeg letters over van het alfabet daarom wordt er gebruik gemaakt van zo een anders geschreven functie(die 0 met streep wordt een theta genoemd)



J staat voor Loss function en een Loss functipon geeft aan hoe goed het model is gefit, meestal is dan het minimum waarde de beste fit

Residual error geeft het verschil aan tussen predicted en echte waarde, zo kan je ook laten zien hoe goed de fit is binnen het model



X staat voor oso groote

Y staat voor oso prijs(dit is ook gelijk de blauwe stip)

I is als je de x en y nummerd dus 1000m2 = 0 , 100k euro = 0

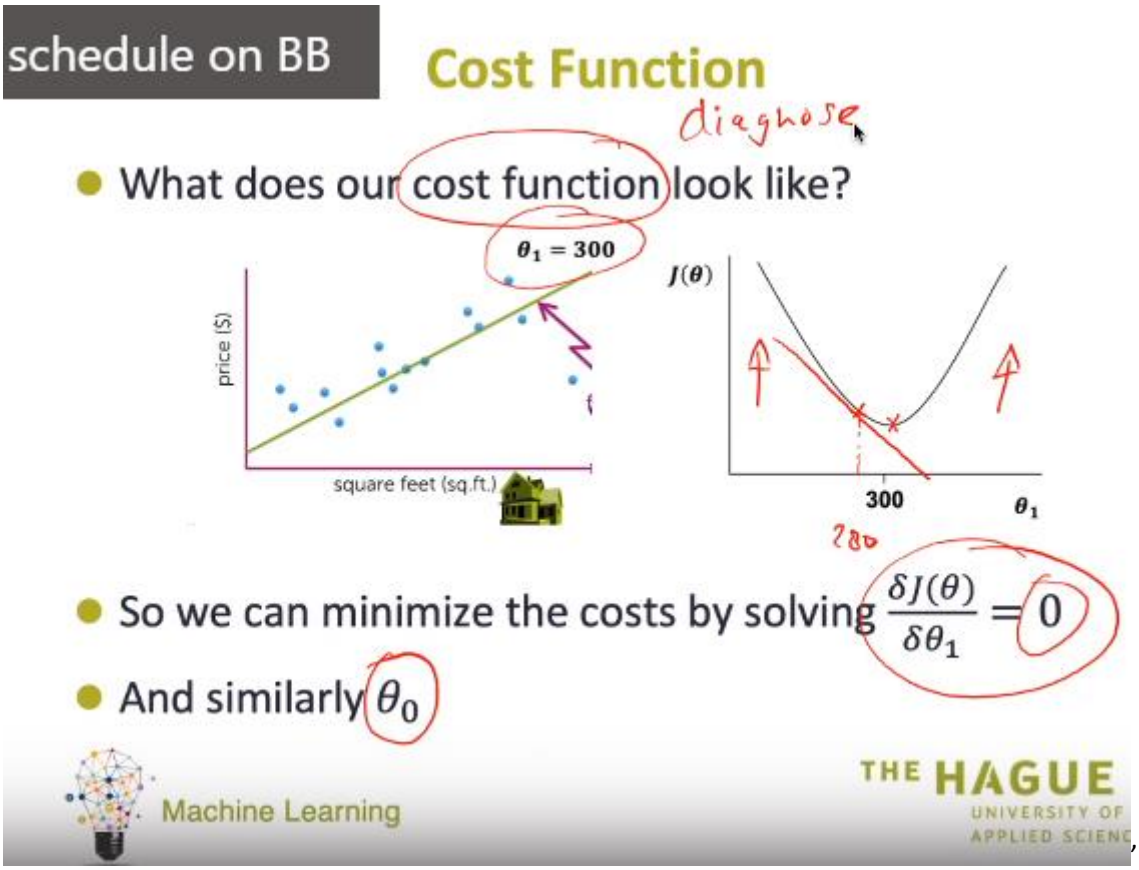
Dus de i is deze X hoort bij deze Y dus deze grootte bij deze prijs.

De Ho(x(i) is de verwachte prijs dus die op de lijn hoort

Die sigma i staar voor (opsommen alle i's)

Als je die hoge 2 hebt geeft dat altijd een parabool als lijn voor je

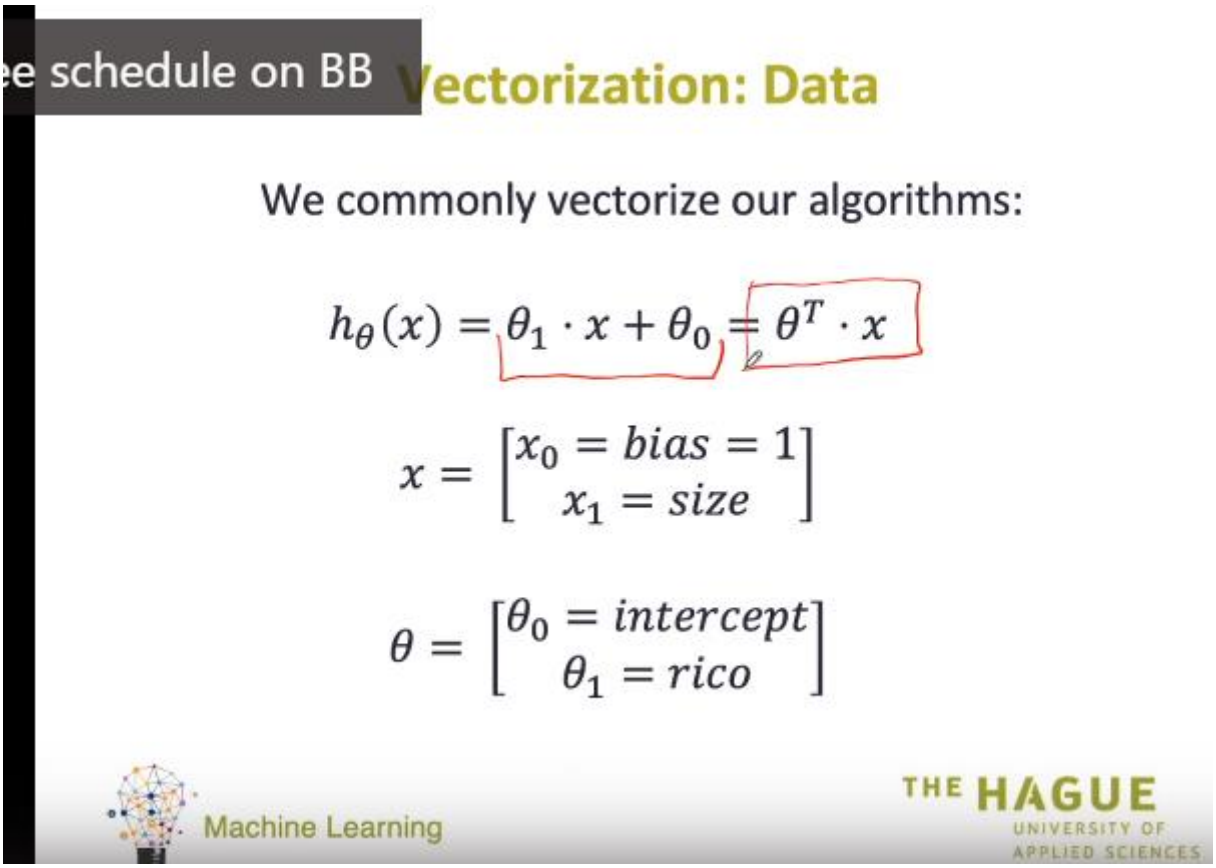
Met deze J functie bereken je hoe goed het model fit

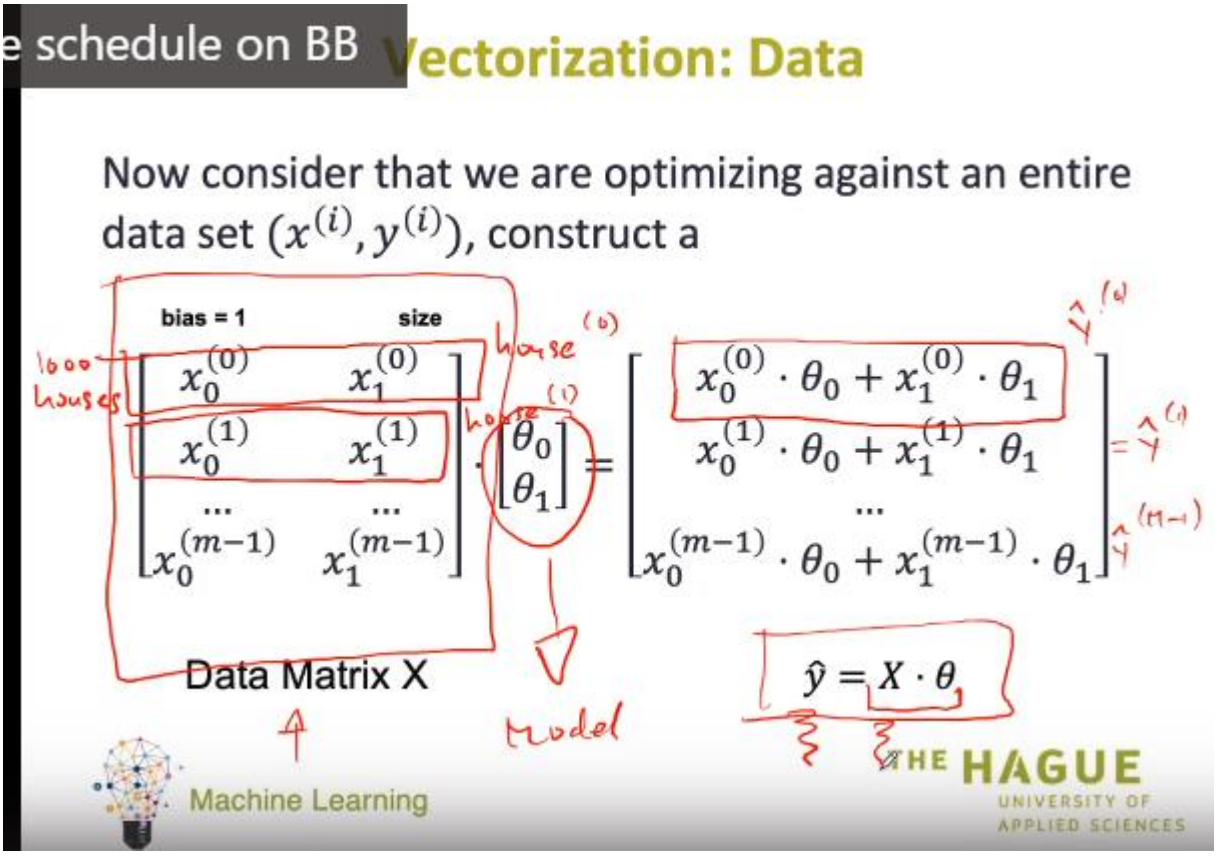


Minimize the cost is het laagste punt, daarvoor los je die functie op daarmee vind je de meest minimum punt vindt.

Je ziet dat het een cost function is door ^2

Bij cost function zoek je altijd het minimum





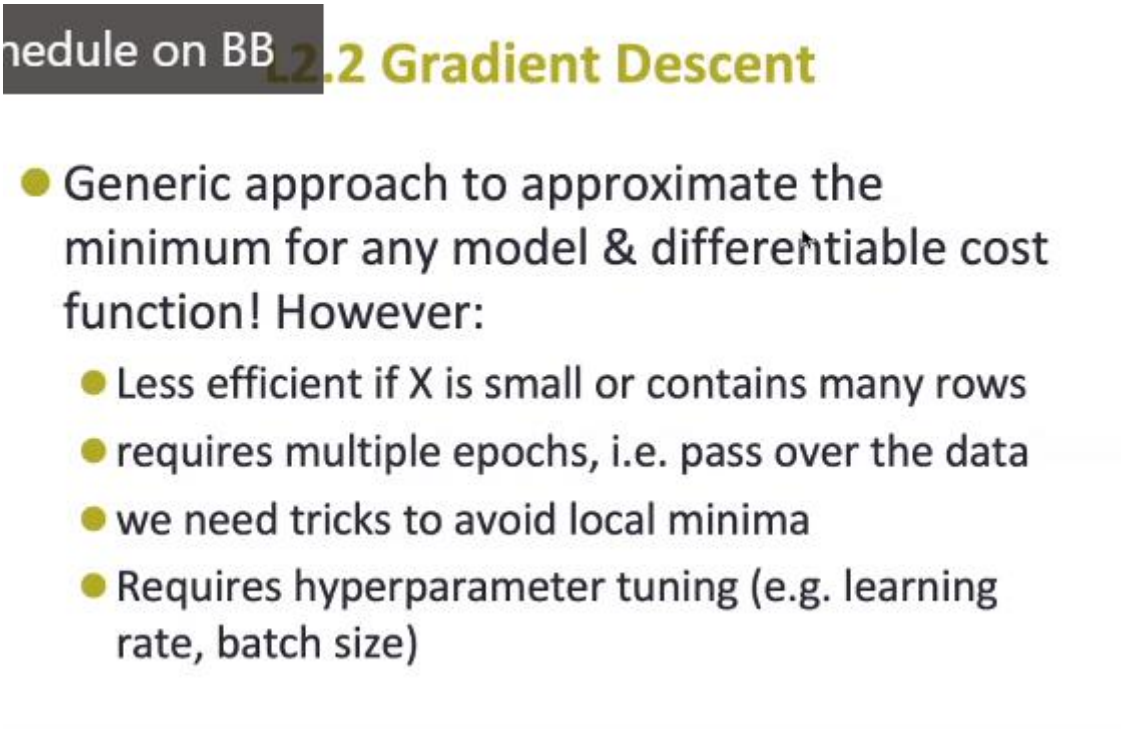
Vectorization is gewoon het vergemakkelijken van de functies, dus een lange functie kan je verkleinen naar een kortere functie(net als bij wiskunde waarbij je die functies korter kon opschrijven) . Laatste rechter kolom is de kolom met predictions

Ook is bij data matrix de rijen bijvoorbeeld een huis en kolommen zijn features. Daarno komt er een klieine [ ] wat het model representeerd en de rechter [ ] representeerd de predictions

- Solving: $\frac{\delta}{\delta\theta}J(\theta) = 0$

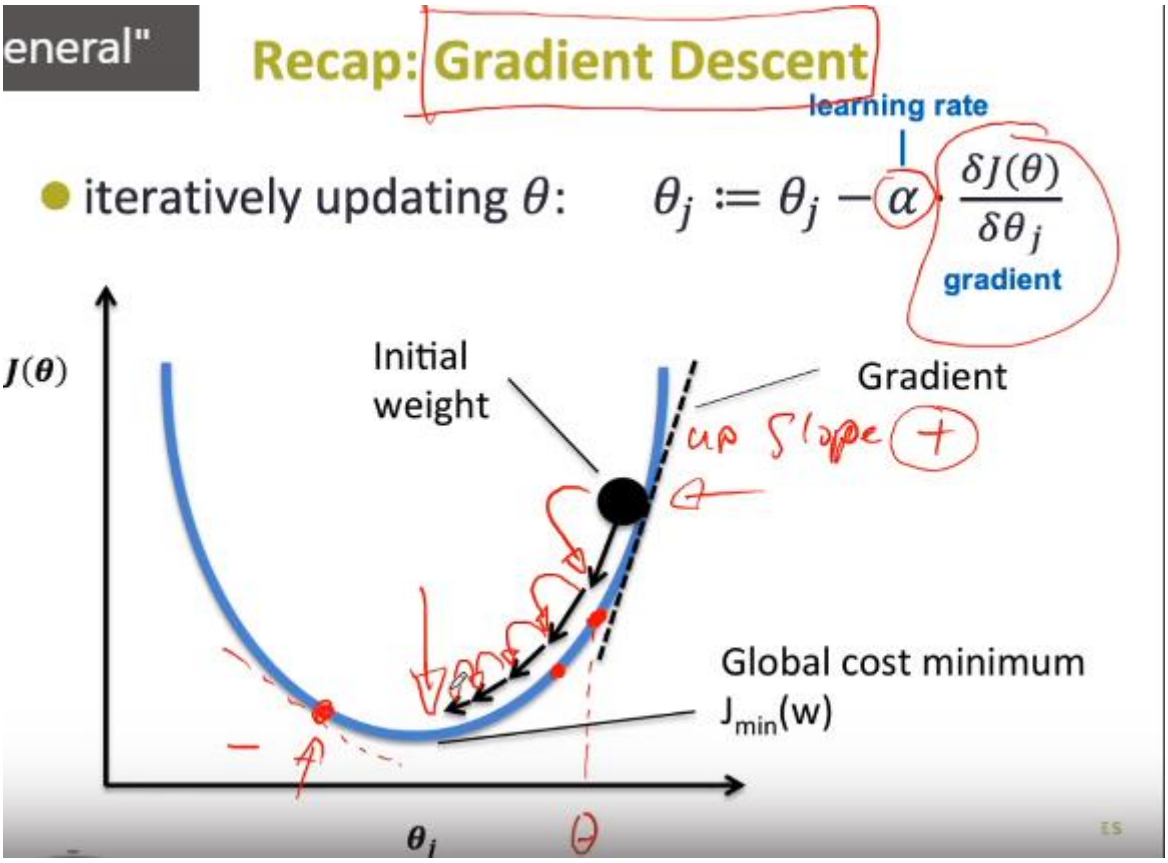
Dit is de solving functie deze werkt niet altijd als deze niet werkt dan heb je andere opties zoals;

- Normal equation(analytical)
- Gradient descent (deze is een truck om het optimale minimum punt in een model te vinden of verschil cost functie)



Deze benadering werkt voor elk model, tenzij:

- De data klein is of veel rijen aanwezig zijn
- Heeft meerdere epochs nodig
- We hebben truckjes nodig om local minima(extreme waardes) te vermijden
- Heeft hyperparameter tuning nodig



Om optimale model te vinden gebruiken we gradient descent. Het doel is het vinden van minimum punt.

Bij een random punt begin je en gebruik je de gradient solve formule om te kijken of je naar beneden moet of omhoog.

Altijd de slope(de lijn)(gradient) trekken vanaf de rechter kant van de punt. Als die naar boven gaat is het een + en als die naar benenden gaat een –

Als de cijfer voor de sloop hoog is is het een stijle lijn en minder hoog cijfer is het minder stijl

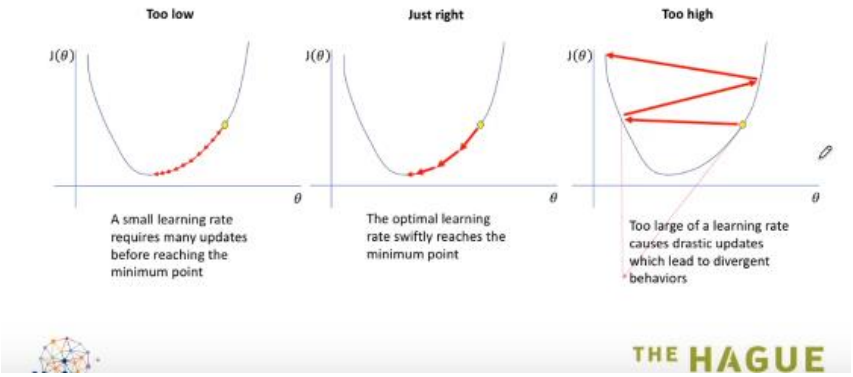
Dus deze gradient geeft aan welke richting we op moeten voor het minimum punt

Hoeveel stappen hebben we nodig om het minimale punt te behalen?
-hangt af van learning rate(a) en data

chedule on BB

Gradient Descent

- Need to find a good learning rate α

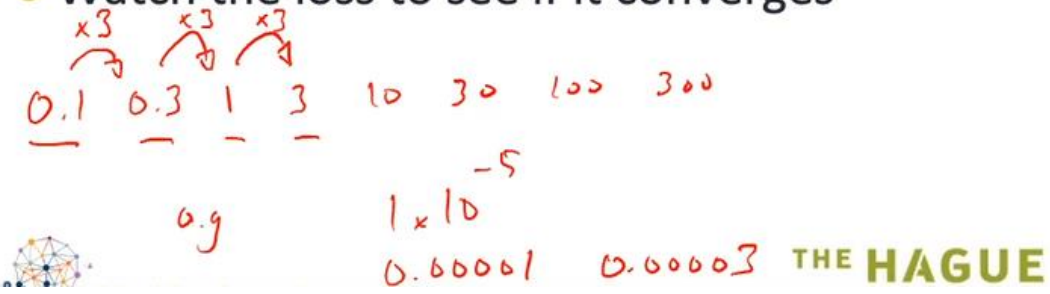


(learning rate is hoe groot je elke stap moet nemen) middelste learning rate is het beste

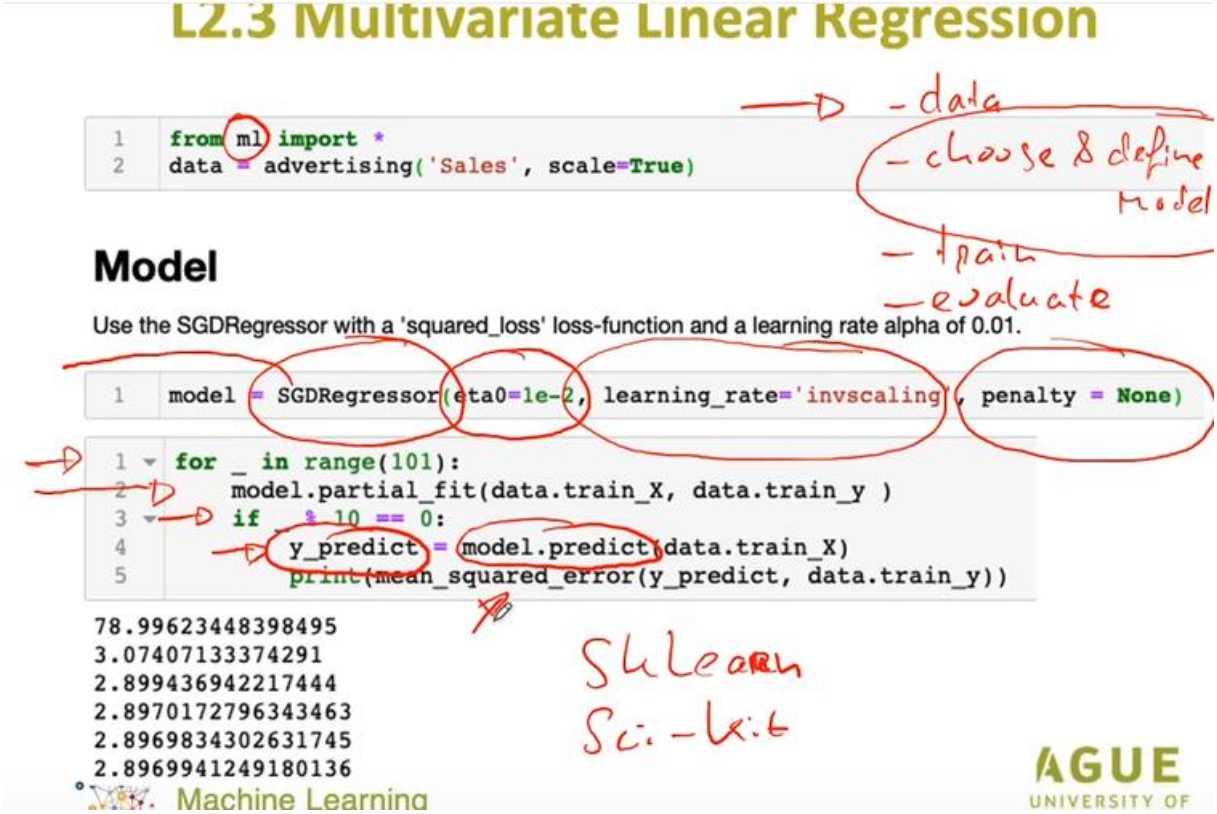
Gradient Descent

- Hyperparameter: learning rate α
- Challenge: need to find a good α
- Rule of thumb, try the sequence

  1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3, etc.
- Watch the loss to see if it converges



Bij rule of thumb doe je x3 maar als je in de buurt bent van 1(0,9) dan rond je dat getal naar een mooi rond getal dus 0,9 wordt 1 en 9 wordt 10



Waarom dalen de nummer en stijgt het bij de laatste ?
-omdat er bij de ene laatste het minimum is bereikt en bij de laatste gaat het over de minimum dus stijgt het weer.

We hebben gezien dat linieare regressie, multivariate en polynominal regressie de zelfde modeelen zijn dus is het belangrijk om de data te scalen.

Classification

Regression describe or predict real values

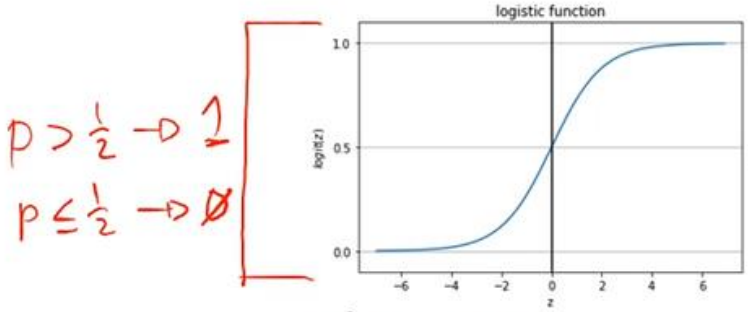
- House price

Classification: predict a class

- Medical diagnosis (e.g. diabetes I, II, none)
- Predict outcome legal case
- Customer sentiment/attitude in stores

L2.6 Logistic Regression

$y \in \{0, 1\}$



$$logit(z) = \frac{1}{1 + e^{-z}}$$

logit(x) may be interpreted as $P(y = 1|x; \theta)$

$p > \frac{1}{2} \rightarrow 1$

$p \leq \frac{1}{2} \rightarrow 0$

$$h_\theta(x) = logit(\theta^T \cdot x) = \begin{cases} 1, if\, h_\theta(x) \geq 0.5 \\ 0, otherwise \end{cases}$$

linear

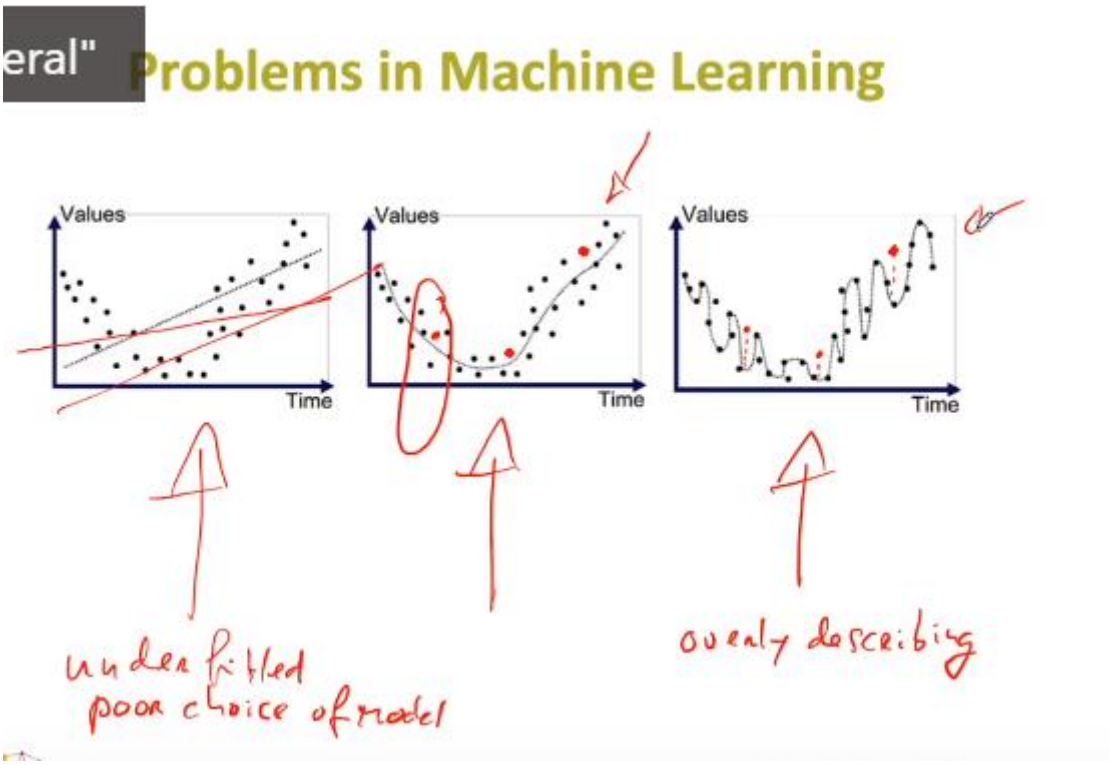
De cost function is een sqaurred function, waarom kan er een local minima dus?
omdat het een functie is van Y en dat heeft een local minima

schedule on BB

What works best?

Optimizer	Pro	Con
Normal Equation	Analytical Exact solution Fast when #features < 100	Does not always work Slow #features > 100
Gradient Descent		
- Liblinear	General purpose simple	May not easily converge
- Adam	A lot faster	Not as good on colinear features No L1 regularization??
- Lbfgs	Less problems converging	May consume much memory
- SAGA	Fast with HUGE data and L1 regularization	Must scale data

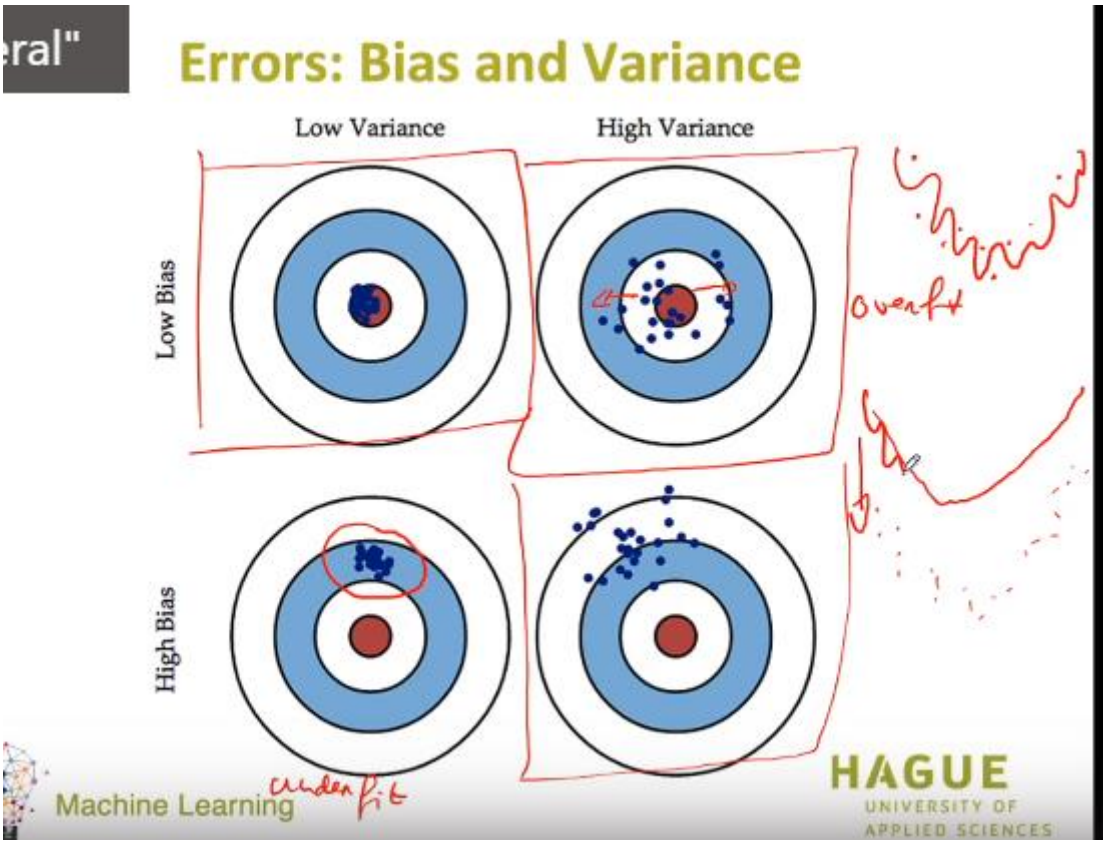
Week 4 training en validatie



Linker model is underfitted dus een poorchoice, de lijn valt niet goed samen met de punten

Rechter model is overfitting het is te precies en een afstand van een niewue prediction en de lijn is te groot.

De middelste is een good choice want de lijn is tussen alle punten en een prediction kan goed gemaakt worden



Bias is een systematische fout

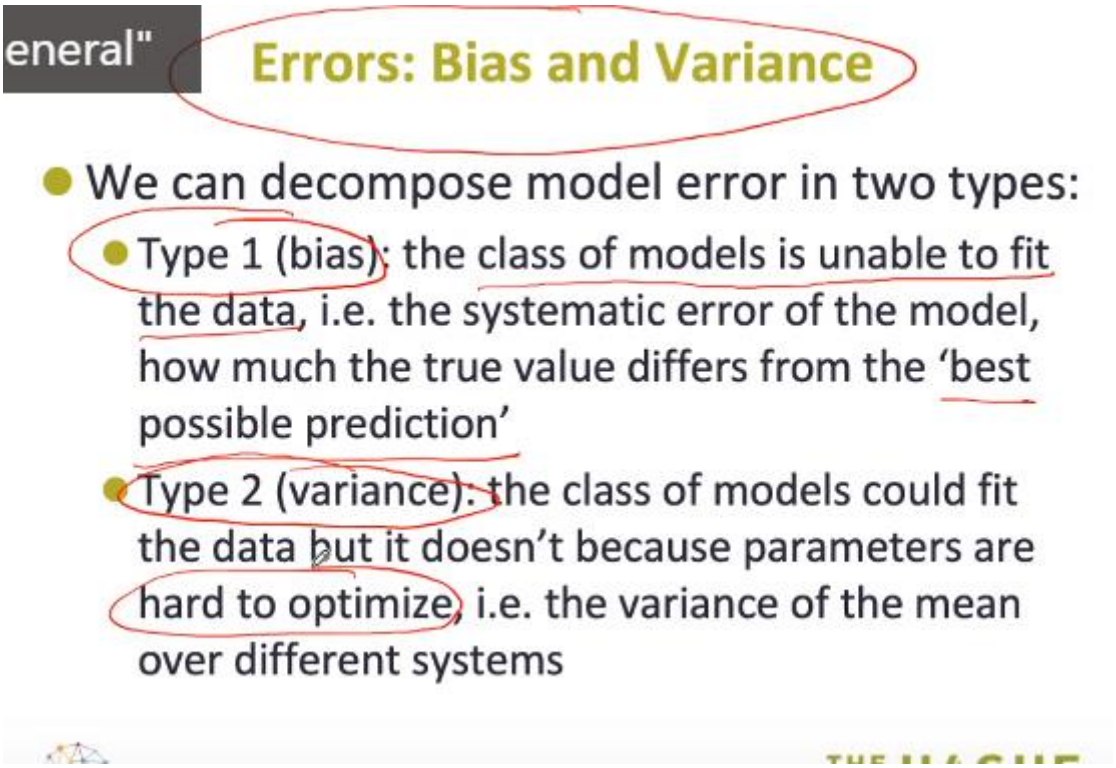
Variance is variatie

Mooiste is lage bias en lage variatie

Hoge bias en lage variatie is underfit

Lage bias en hoge variatie is overfit

Hoge bias en hoge variatie moeten we ook niet hebben

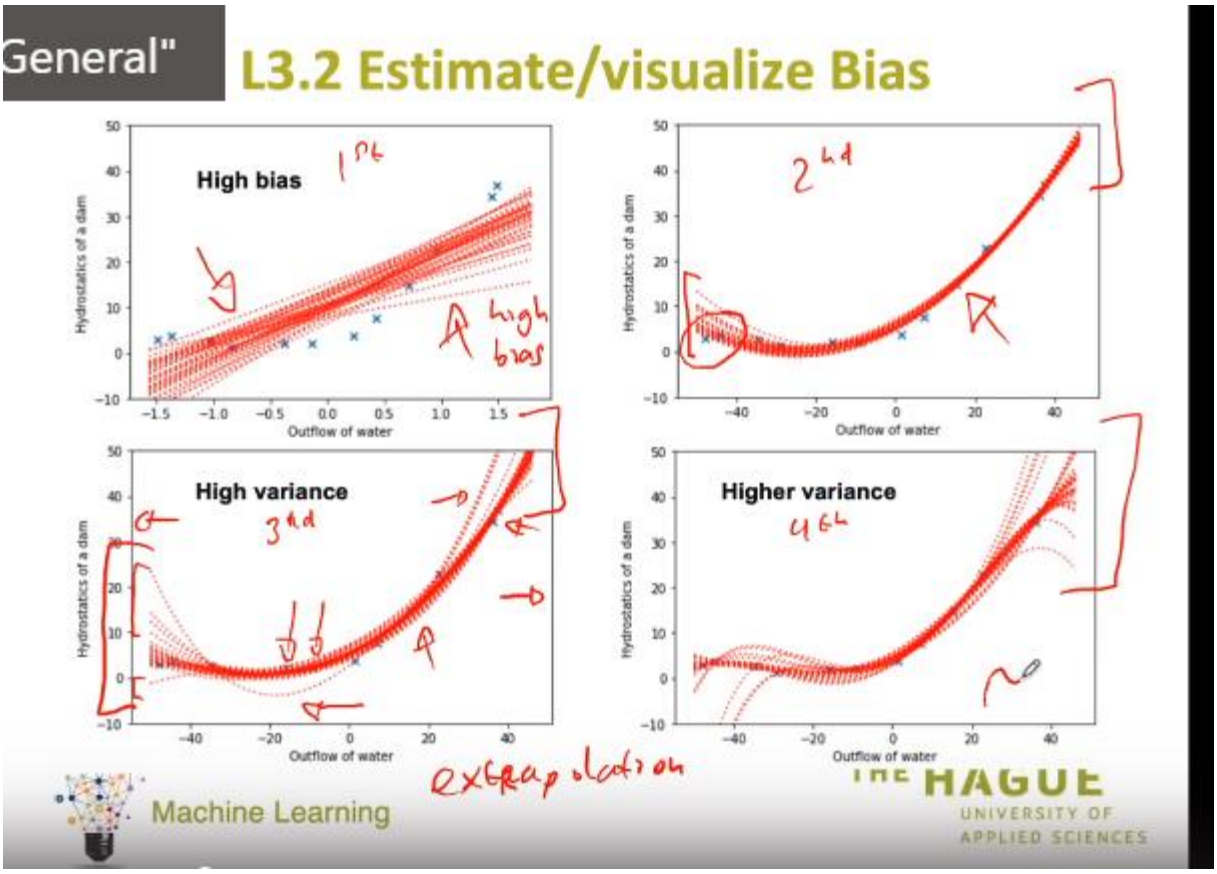


Er zijn 2 typen errors

Type 1 error ; bias : de verschillende modellen zijn niet toepasbaar op de data, bijvoorbeeld; er is een systematische error van de model

Type 2 error ; variance : de verschillende modellen kunnen passen op de data maar het doet niet omdat er parameters zijn die te moeilijk zijn om te optimialiseren

Hoe weet je hoeveel parameters voldoende zijn?
bij hoe goed het model fit op de validation data

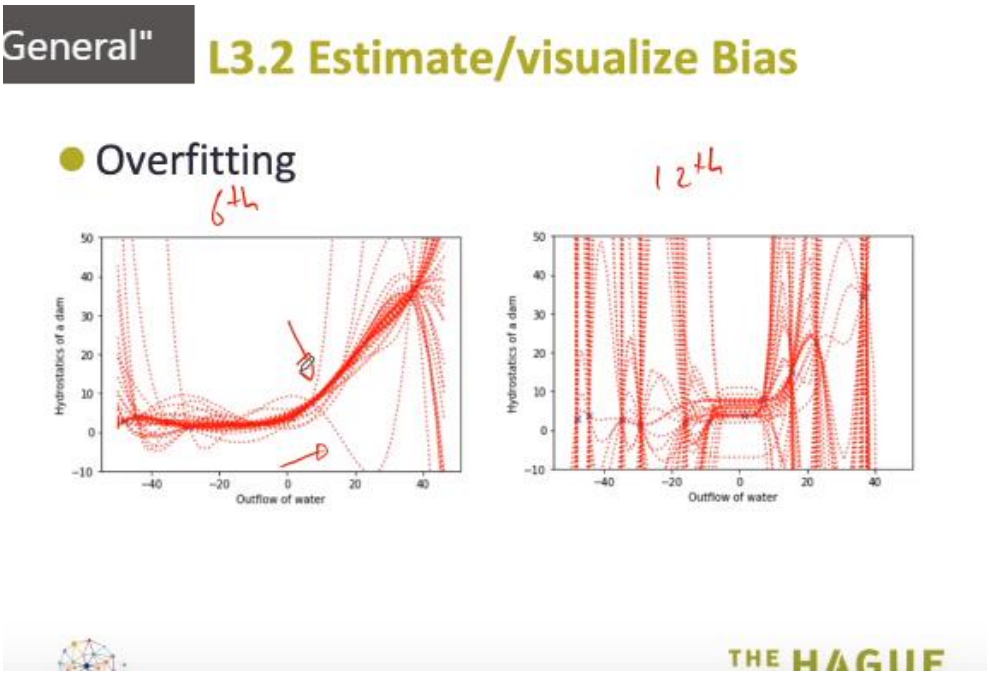


Hier zie je verschil tussen bias en variatie goed

4th is ook overfit

Let op het woordje extrapolation

Extrapolatie is wat buiten de lijnen ligt(3rd)



Hoge bias == underfitting

Oorzaken voor underfitting

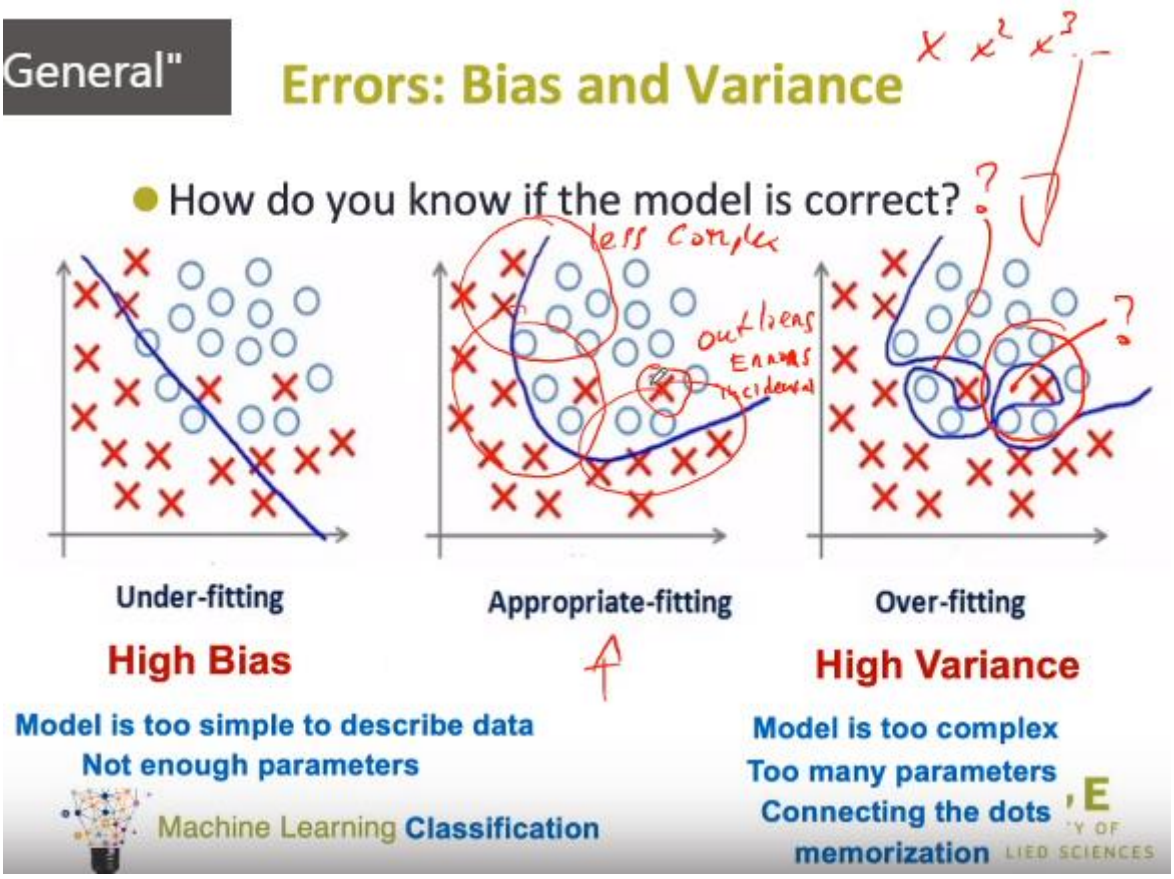
- Model is te simpel
- Niet genoeg informatie(puntjes)(features)
- We hebben het model niet genoeg getraind

Hoge variance == overfitting

Oorzaken voor overfittng

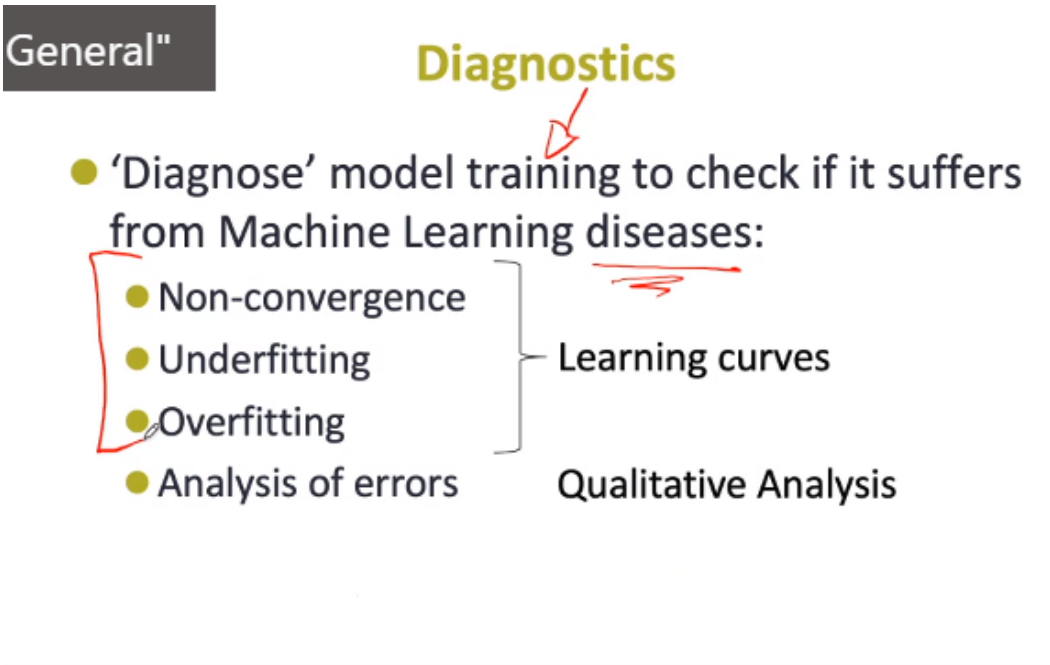
- Model is te complex
- Teveel informatie(puntjes)(features)
- Niet genoeg training data
- Slechte sample(het is niet representatief voor het geheel)
- Het model 'over trainen'

Wat is het oorzaak als een model overfit?
de model generaliseert niet



Error bias op een classifiaction

Bij middelste grafiek zijn er een aantal uitschieters(zie rode kruis bij blauwe rondjes) die moet je negeren dus is niet erg want de blauwe overheerst



Door een diagnose te doen dan check je of het lijd aan machine learning problemen

Deze problemen kunnen zijn

Leer lijnen

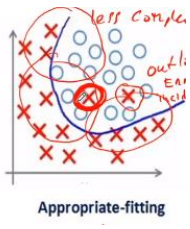
-non convergence(het vind geen minimum)

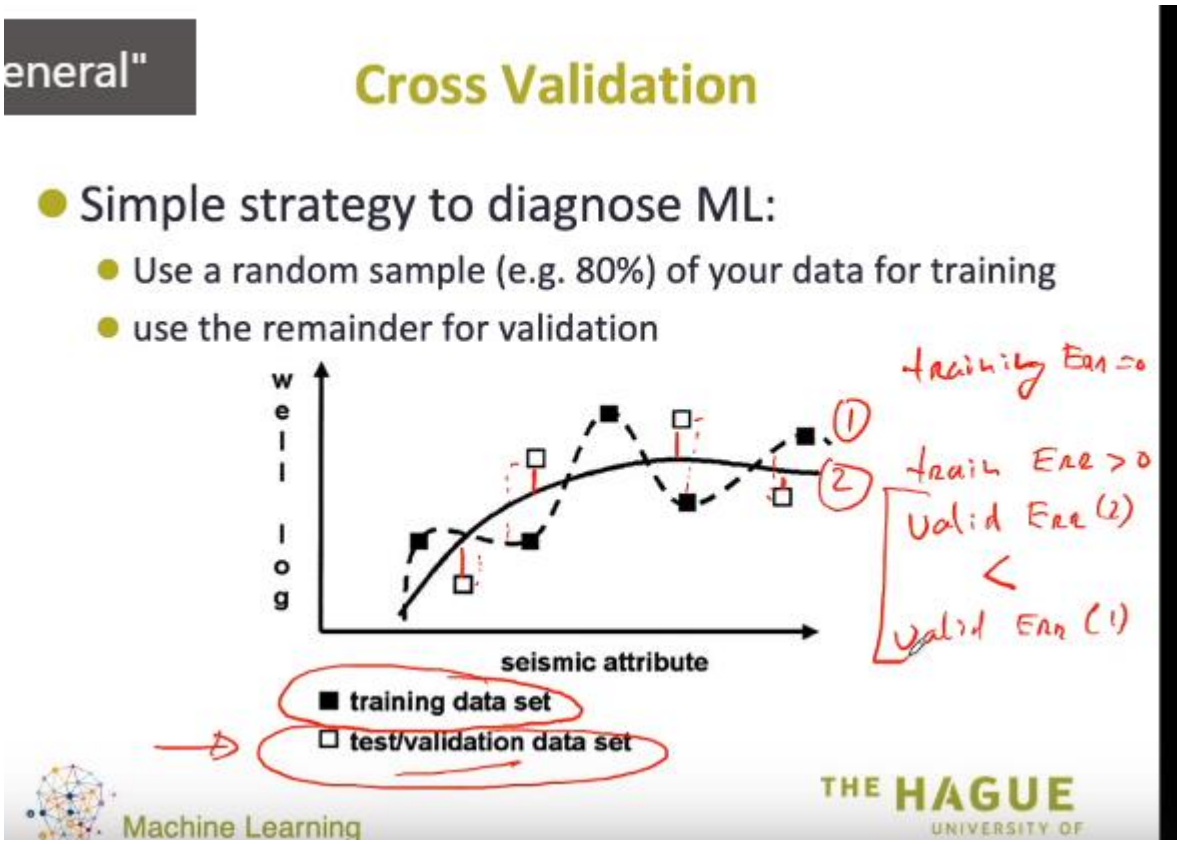
-underfitting

-Overfitting

Qualitative analyse

-error analyses(check foto eronder, met deze diagnose check je waarom die rode kruis er staat bij de bluawe rondjes groep, je zoekt een reden)(het is belangrijk voor ons project om zulke fouten te beschrijven waarom ze zijn gemaakt)





Simpele strategy voor een ML diagnose:

Gebruik een random proef van je data voor de training

Gebruik de rest om te valideren

Je hebt zwarte en witte blokken

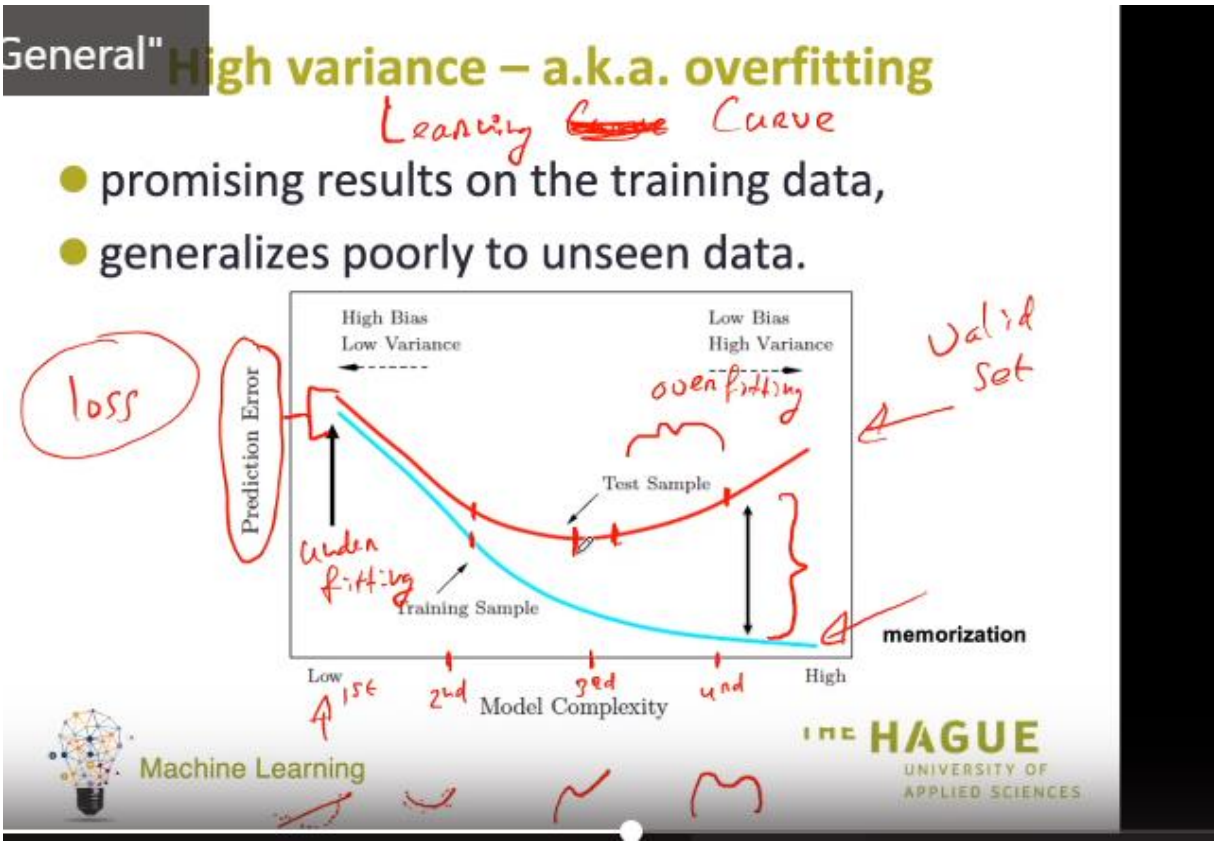
Zwarte blokken zijn lijn 1

Witte blokken heeft lijn 2

Training error voor 1 is =0

Training error voor 2 is >0

Validation error voor 2 is < dan validation error voor 1



Veelbelovende resultaten op een training data

Genereerd slechte tot ongezeine data

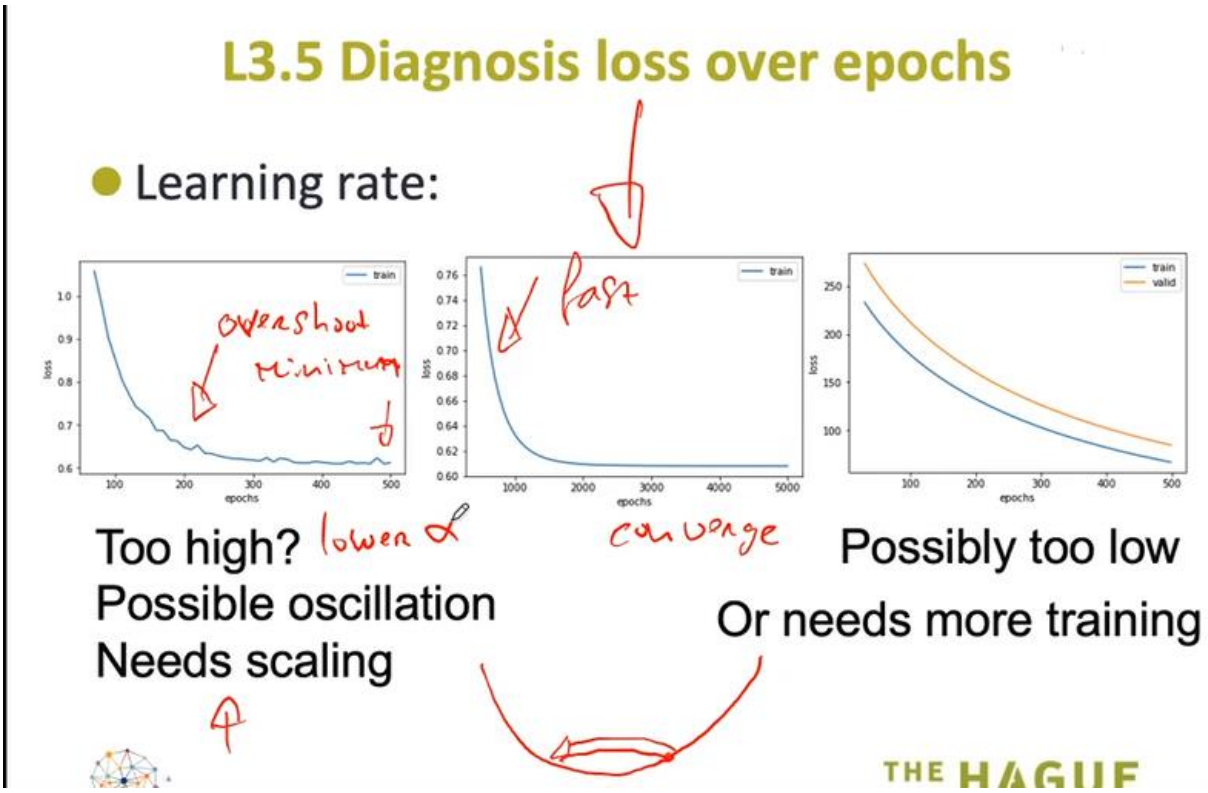
Optimale punt is bij de streep links van 3rd

Recht van 3rd is over fitting

Links van 2nd is underfitting

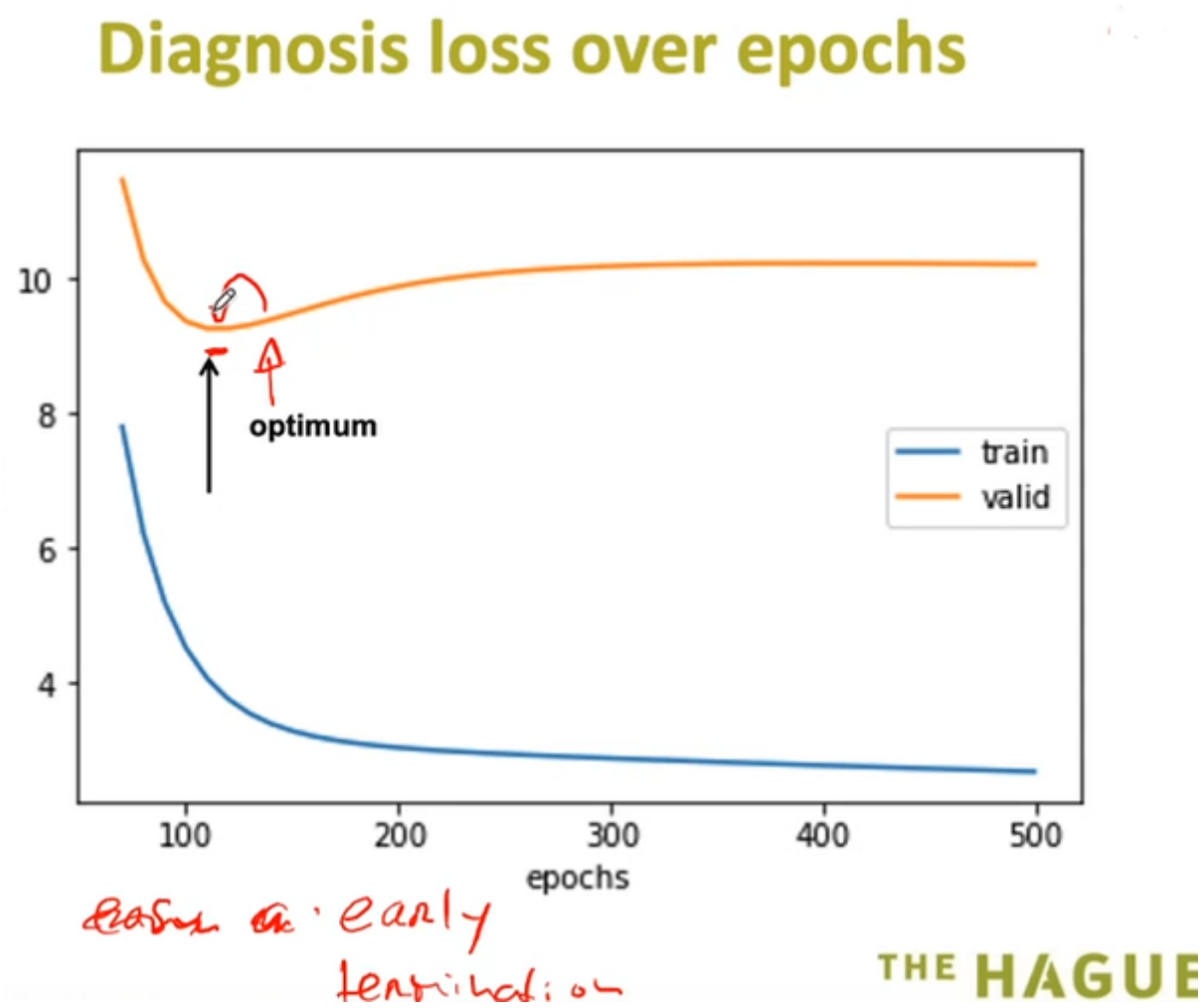
1,2,3,4 laat de order polynominal zien daar onder

Epoch is een pass over het gehele data set

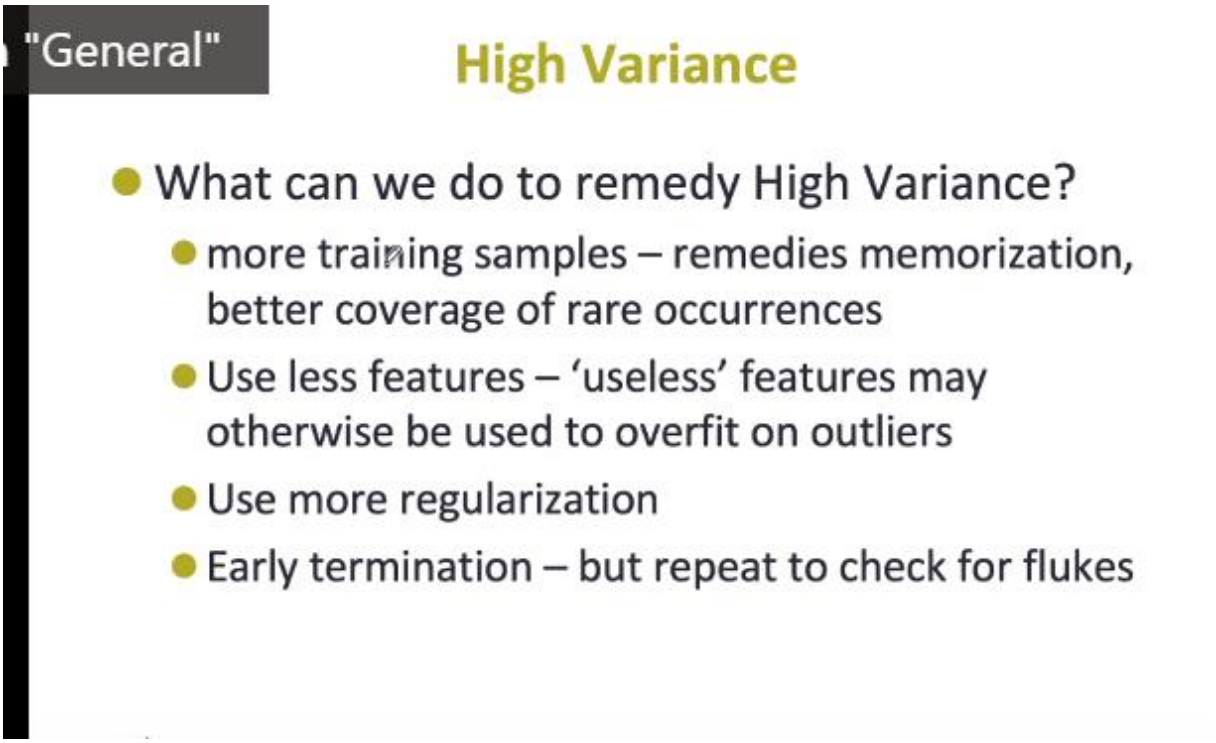


Linker foto hoe kan je dat oplossen?
verlagen van learning rate

Het is waarschijnlijk aan het Oscillation
en daarvoor gebruik je Scaling



Je doet je ding je ziet het stijgt van het optimale punt dan doe je early termination en ga je terug naar het optimale



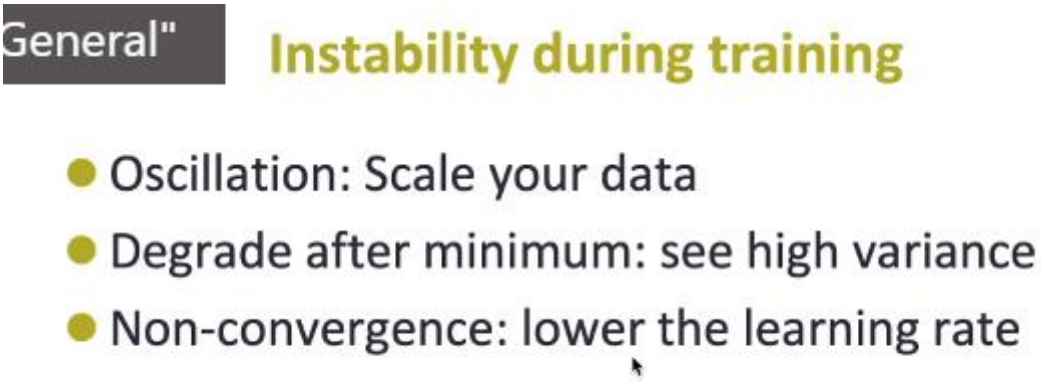
Wat doe je tegen High variance

- Meer training
- Gebruik minder features
- Meer gebruiken van regularization(in orde brengen)(model vergemakkelijken)
- Early termination(dit is je hebt een laagste punt dus optimum punt als je iets verder gaat gaat de lijn weer omhoog dan kies je om weer terug te gaan en dan heb je optimale punt dus dat is early termination)

- What can we do to remedy High Bias?
  - Use more features, polynomials
  - Train for more iterations (if training has not reached its optimum)
  - Use less regularization

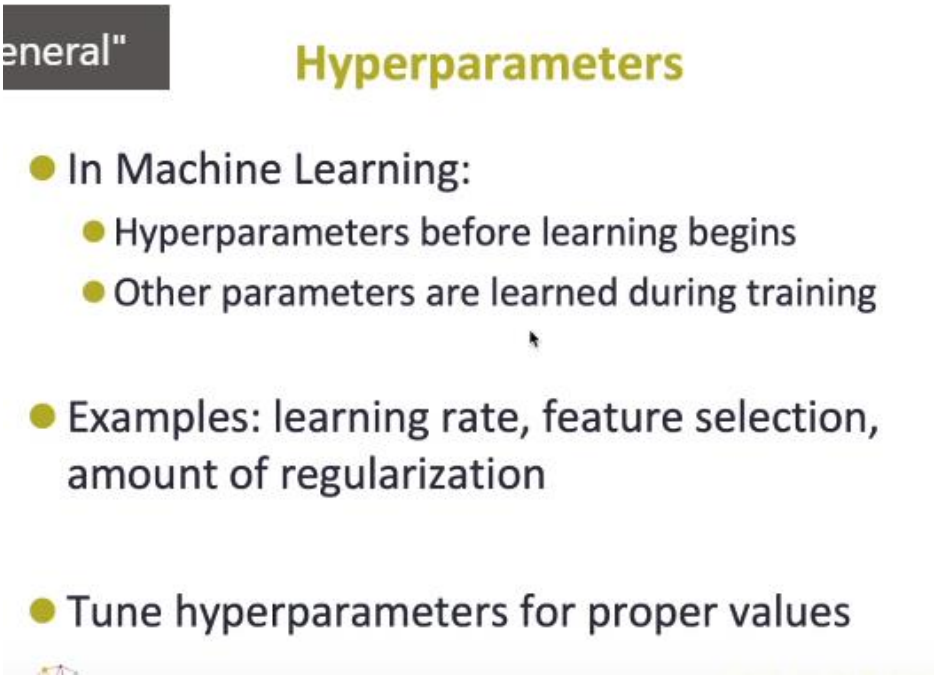
Wat doe je tegen High Bias?

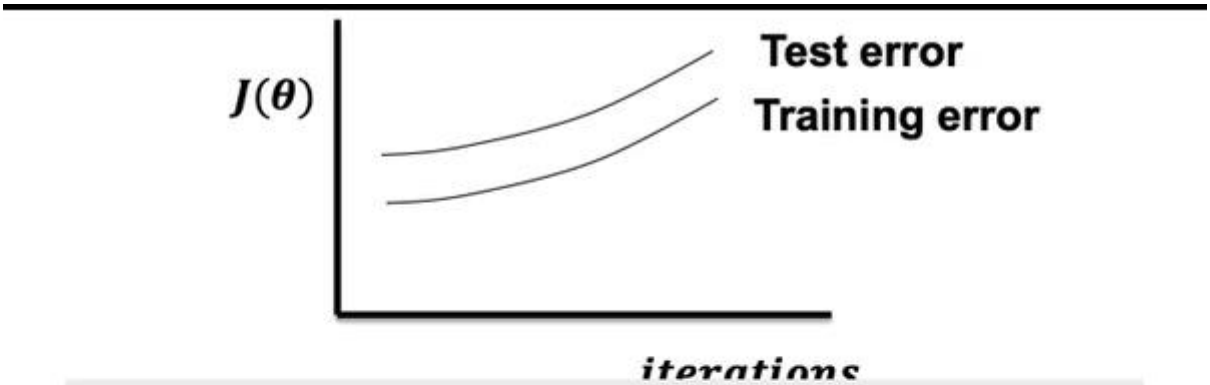
- Gebruik meer features en polynominals
- Train voor meer iterations
- Gebruik minder regularization(model vergemakelijken)



- Wanneer je merkt dat het is geen oscillaten(goed fout heen en weer) danmoet je je data scalen dus beetje mee spelen
- Degrade after minimim : zie high variance
- Non converance (geen minimum hebben): verlaag de leer tempo

Je dient altijd een diagnose te doen na het veranderen van een setting



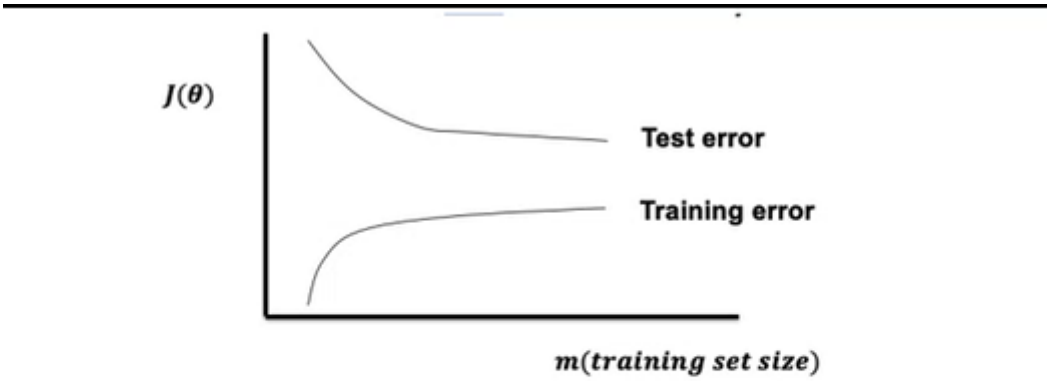


Respond at **PollEv.com/vuur**

What to do?

Top

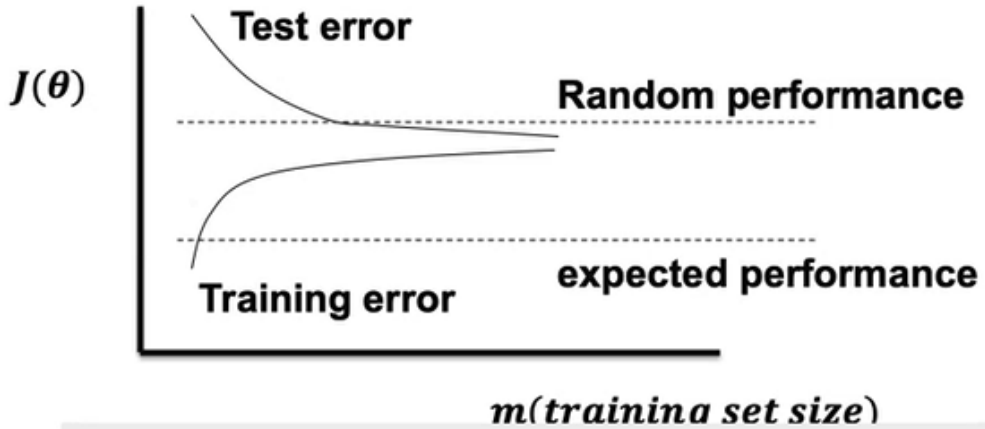
5 Lower the learning rate



Respond at **PollEv.com/vuur**

What to do next?

Increase the trainig set

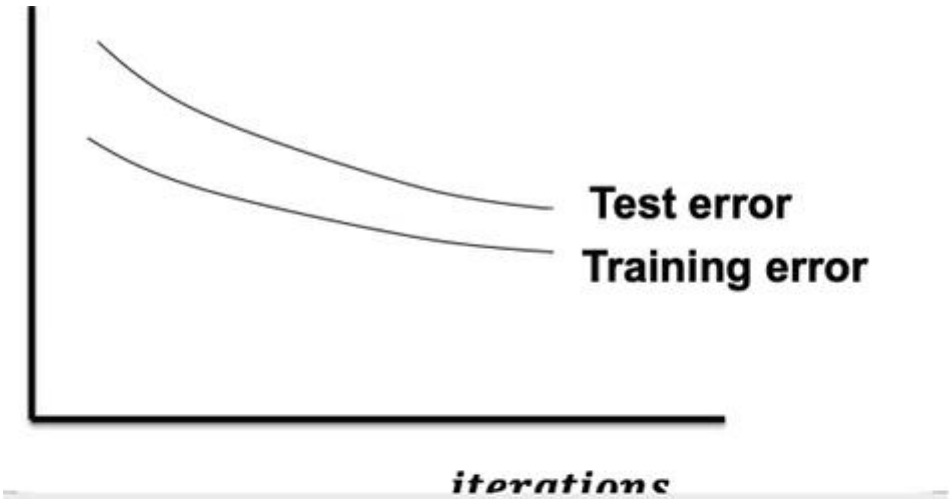


Respond at **PollEv.com/vuur**

What to do (2)?

Top

3	Add a feature



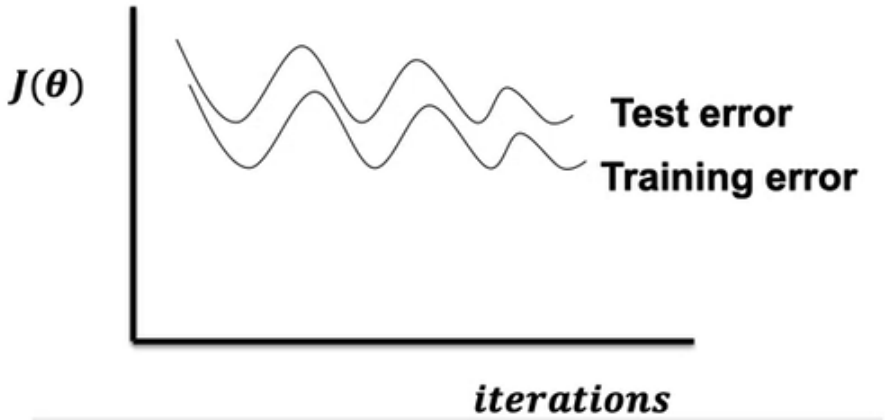
Respond at **PollEv.com/vuur**

What to do (3)?

Top

6	add iterations, because optimum is not yet reached
3	Increase learning rate

Increase learnin rate

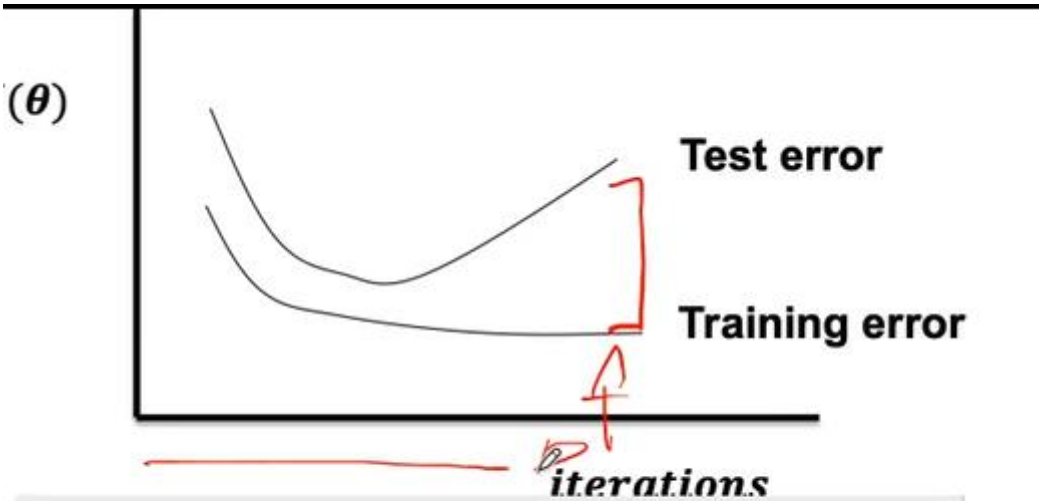


Respond at **PollEv.com/vuur**

What to do (4)?

Top

2 Scale the data

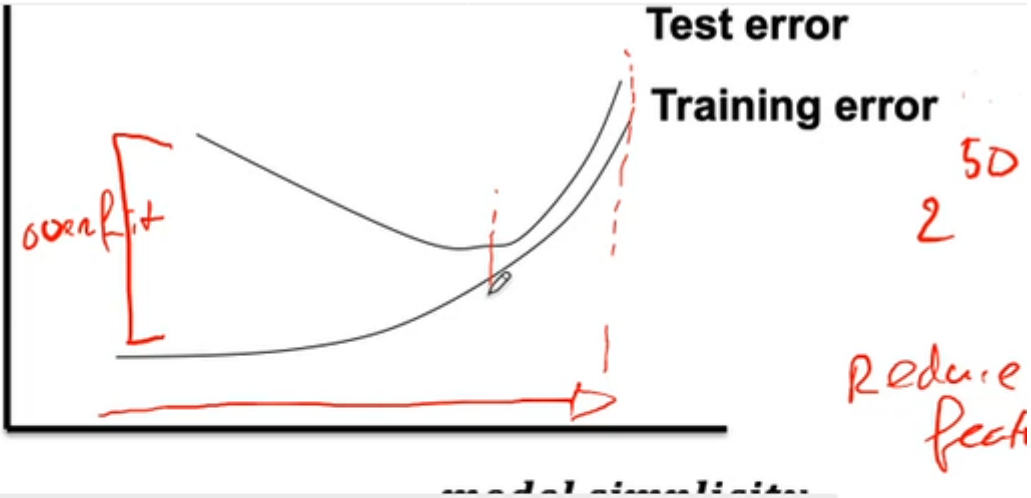


Respond at **PollEv.com/vuur**

What to do (5)?

Dit is een vorm van overfitting dus we kunnen early termination gebruiken of minder features

(θ)



Respond at **PollEv.com/vuur**

What to do (6)?

Top

0	Choose # of features where test error is minimal

Week 5 evaluation

Machine learnin is een Black box, allemaal zijn anders en je weet niet hoe het verloopt.

We maken een aantal veronderstellingen zoals;

- Over het typen model dat gebruikt wordt(regression,desiciosn vector tree,nearest neighbor,support vector machine)
- Over de data, dus wat voor data, hoe gebruik je het, selectie etc

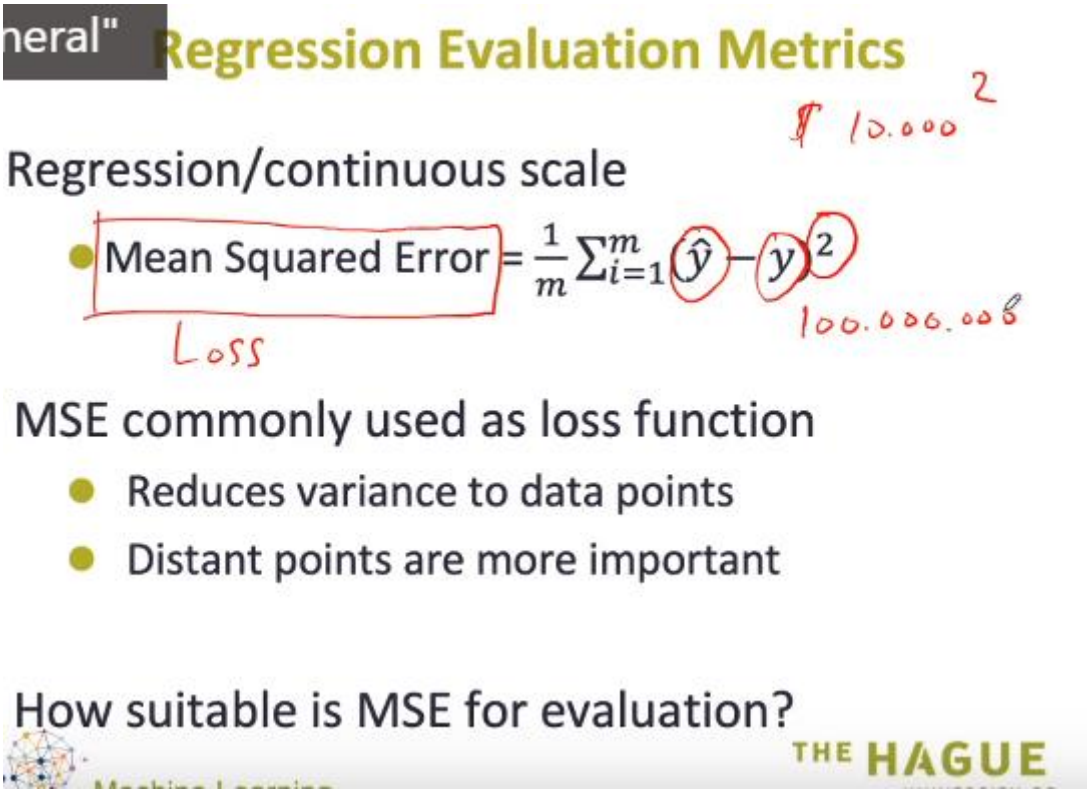
Omdat machine leaning een blakc box is moetenw e kunnen aantonnen hoe goed de ML is, dat doen we door middel van

- Validation(hebben we het model correct geleerd)
- - Evaluation(hebben we het juiste model geleerd)(hoe eefectief is het model)

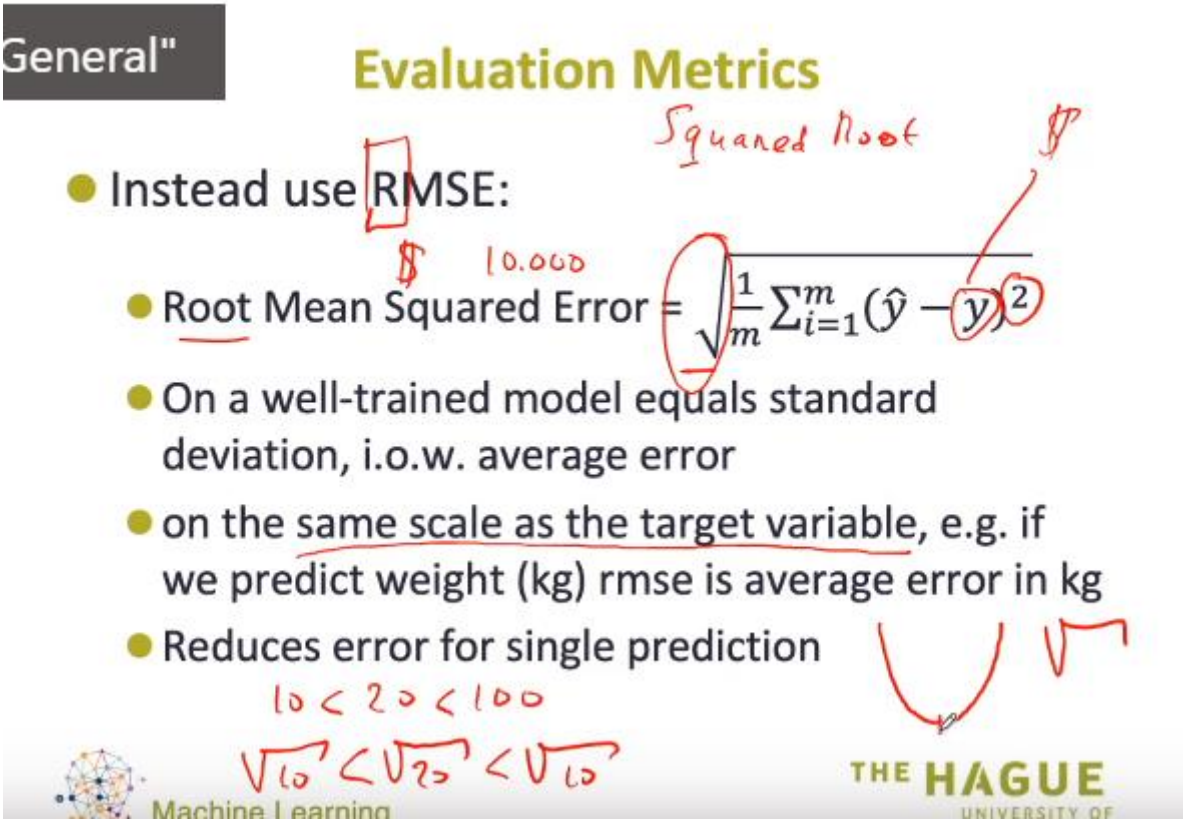
Hoe meten we de efeectiviteit voor;'

- Regression
- Classification

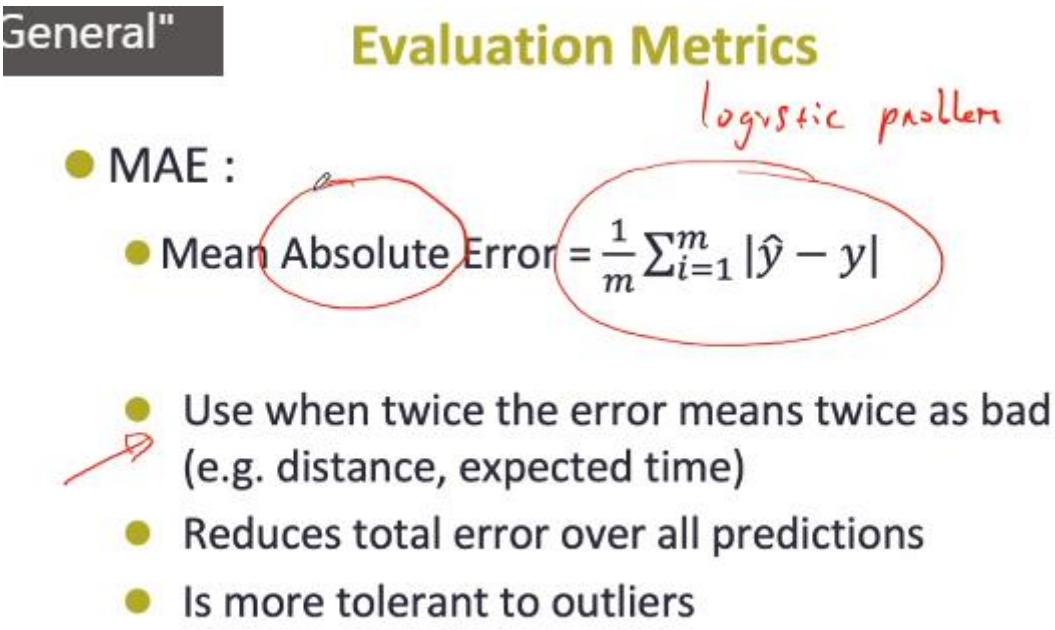
REGRESSION



De MSE is niet geschikt voor evaluation want die ^2 in de functie verhoogt het aantal dus het is niet duidelijk leesbaar(100 wordt 100000)(en dan is de waarde heel anders.

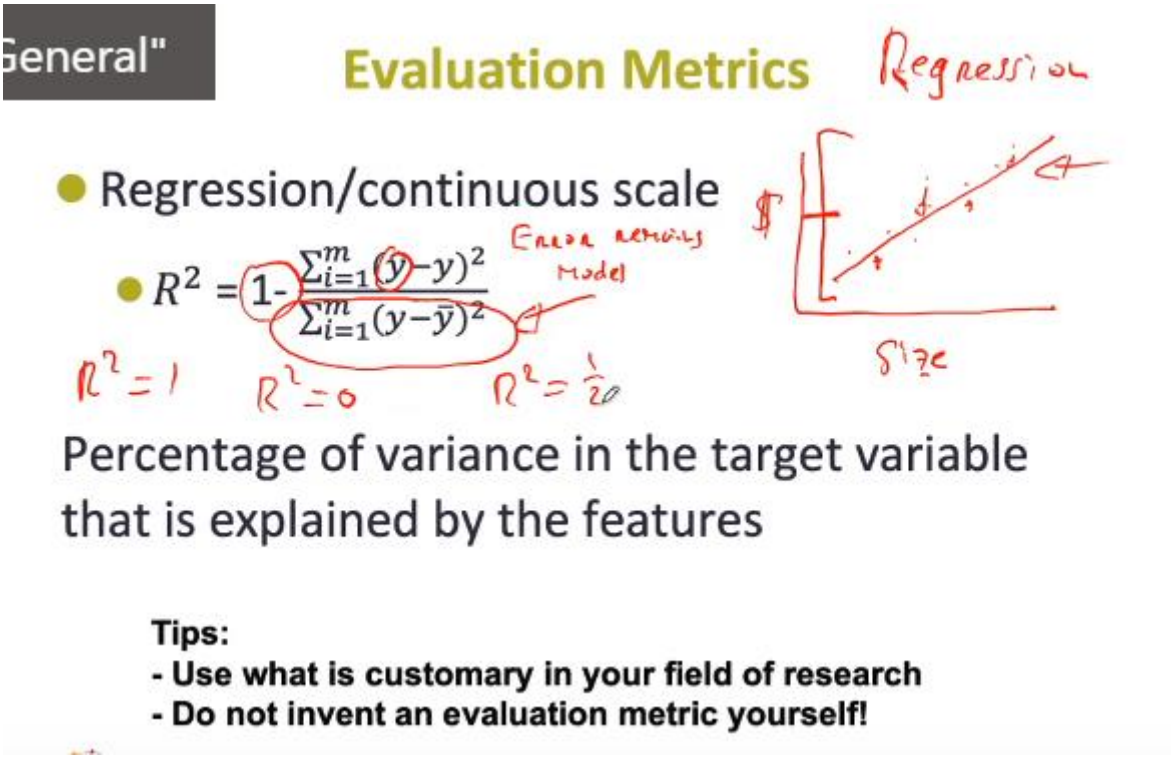


Iomdat tegen te gaan gebruiken we RMSE die ^2 en wortel spreken elkaar tegen dus (100 werd 100000 en emt die wortel wordt het weeer 100) (dus is leesbaar en de waarde blijft t zelfde)



MAE wordt meestal gebruikt binnen de logistic hierin wil je bijvoorbeeld weten hoeveel KM er totaal wordt afgelegd ipv de route die wordt vastgelegd. Dus hiermee als je 2 moddellen hebt kan je zien

met MAE welke kortste route aangeeft(voorbeeld)(de absolute in de functie betekent dat alle – een + wordt)



Dit is de laatste oploss middel voor evaluatie

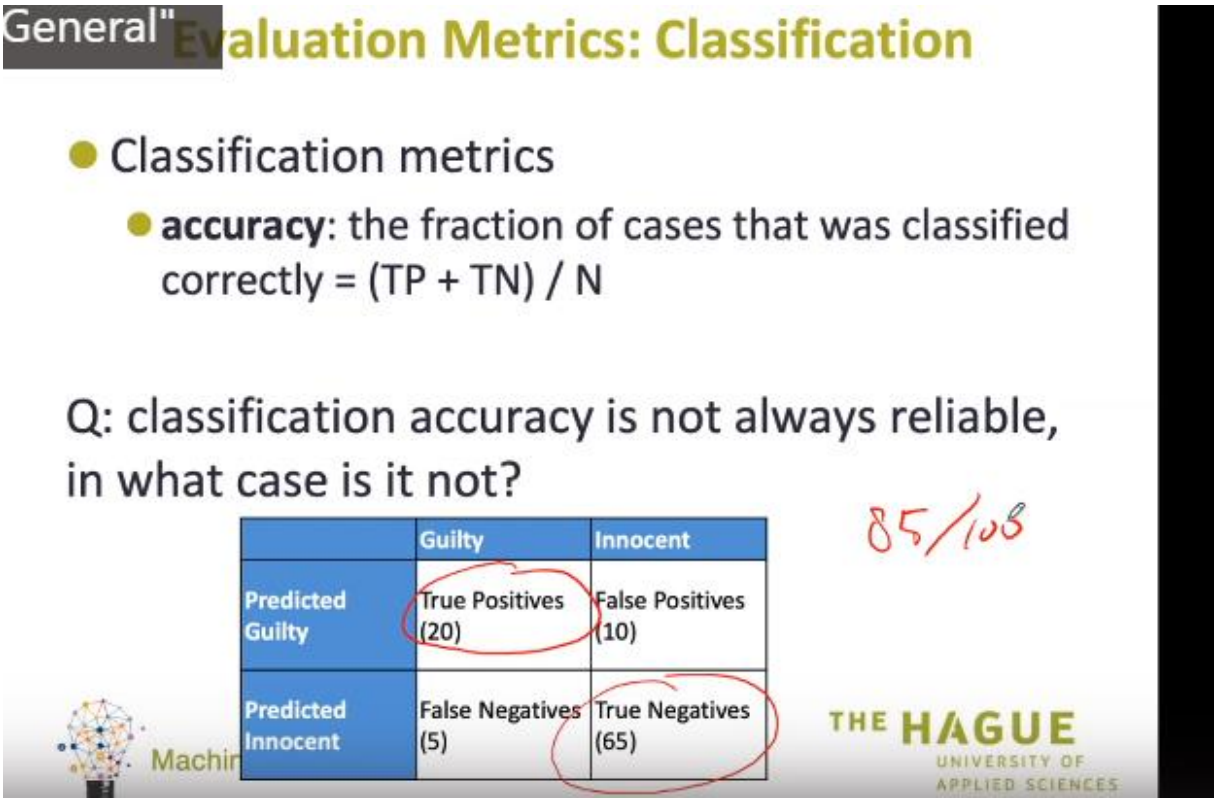
Deze R2 geeft ons de waarde voor de variatie binnen een model dus de waarde tussen andere punten binnen het model

Als R2 = 1 dan is het een goed model

Als R2 = 0 dan slecht

Als R2=0.5 dan zit je er tussen in

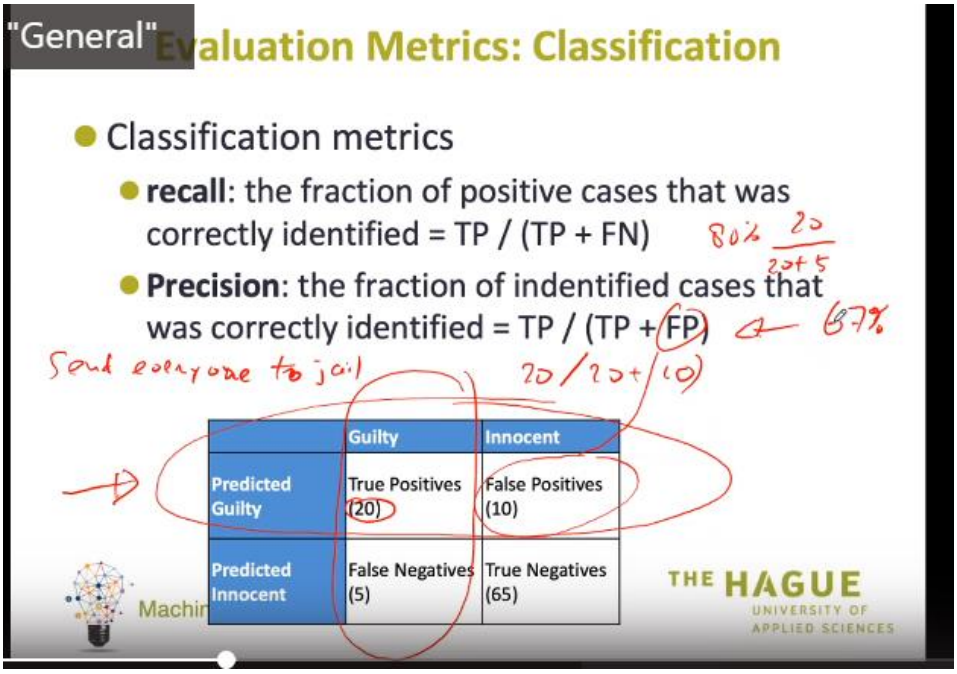
CLASSIFICATION



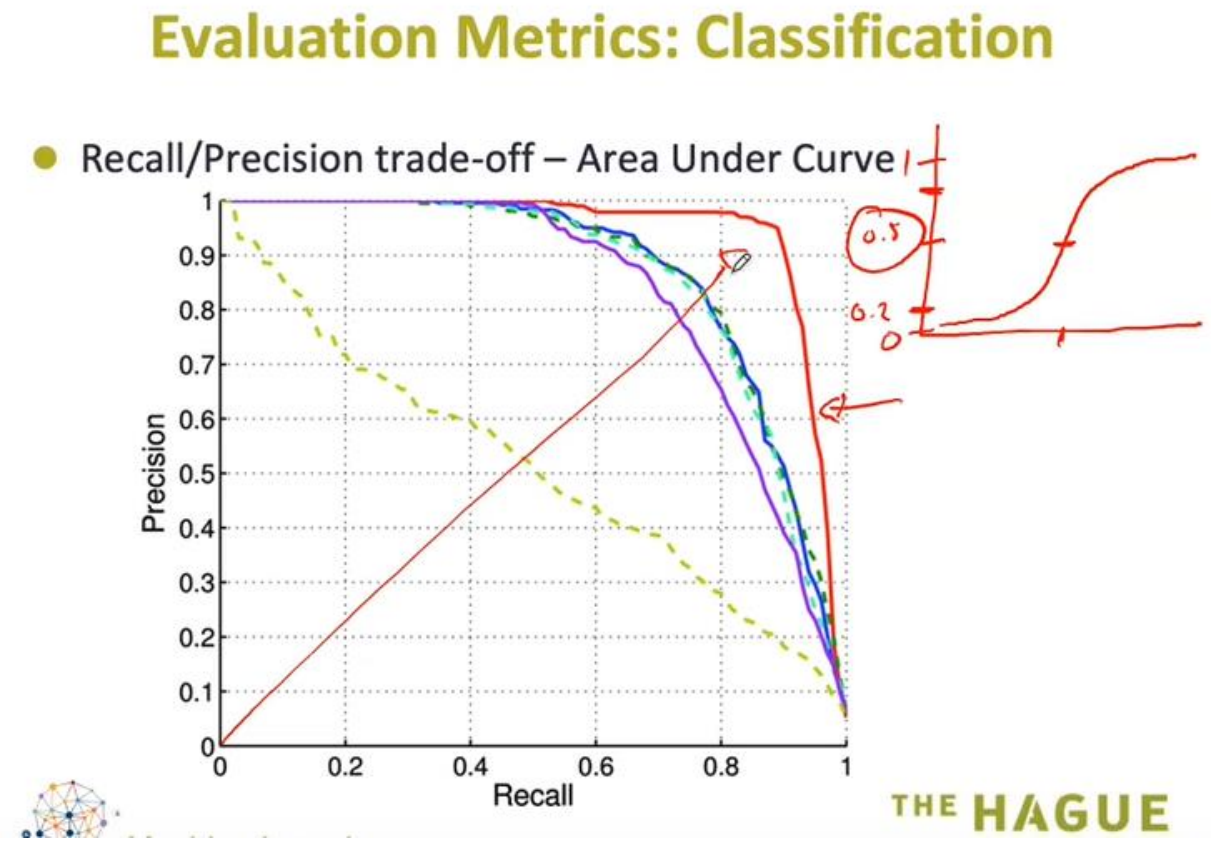
Die tabel heet een conmfusion matrix

Alleen de omcirkelde is correct positief of correct negatief de andere zijn nog twijfel geval.

Je berekent de naukweurigheid van evaluation classification dmv de 2 omcirkelde bij elkaar op te tellen en het te delen door het totaal dus in dit geval 20+65(85) / 100

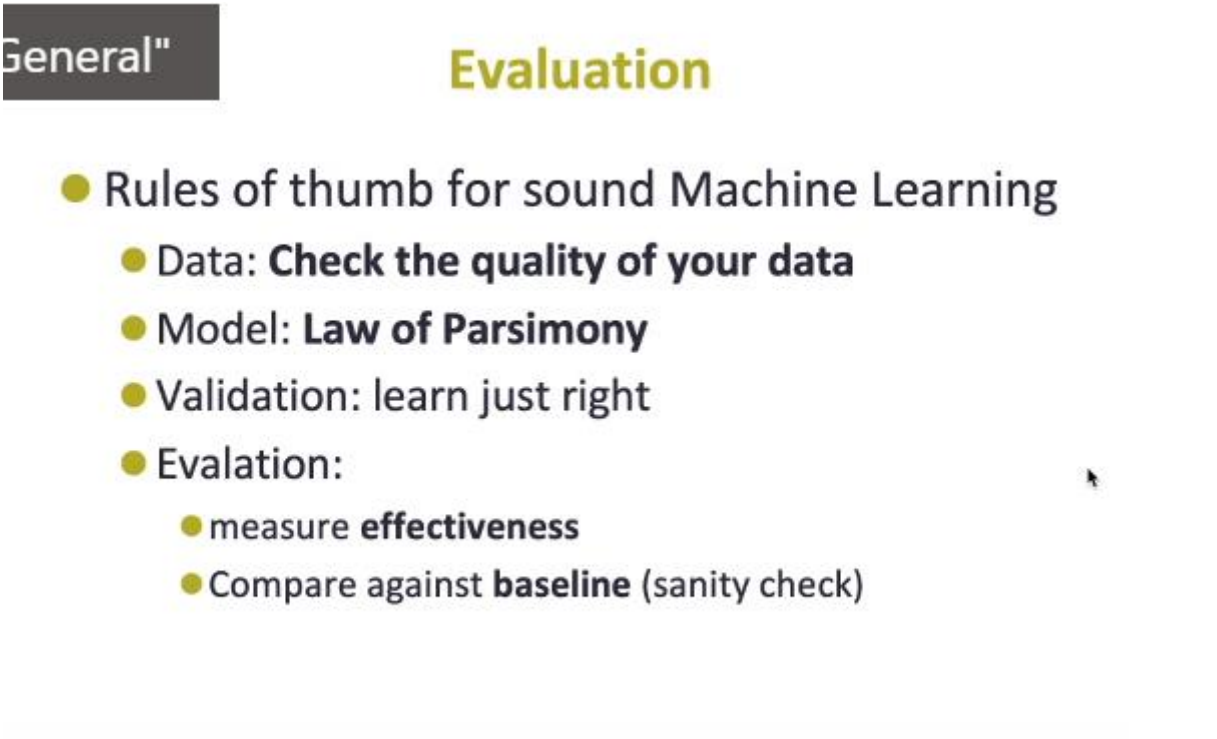


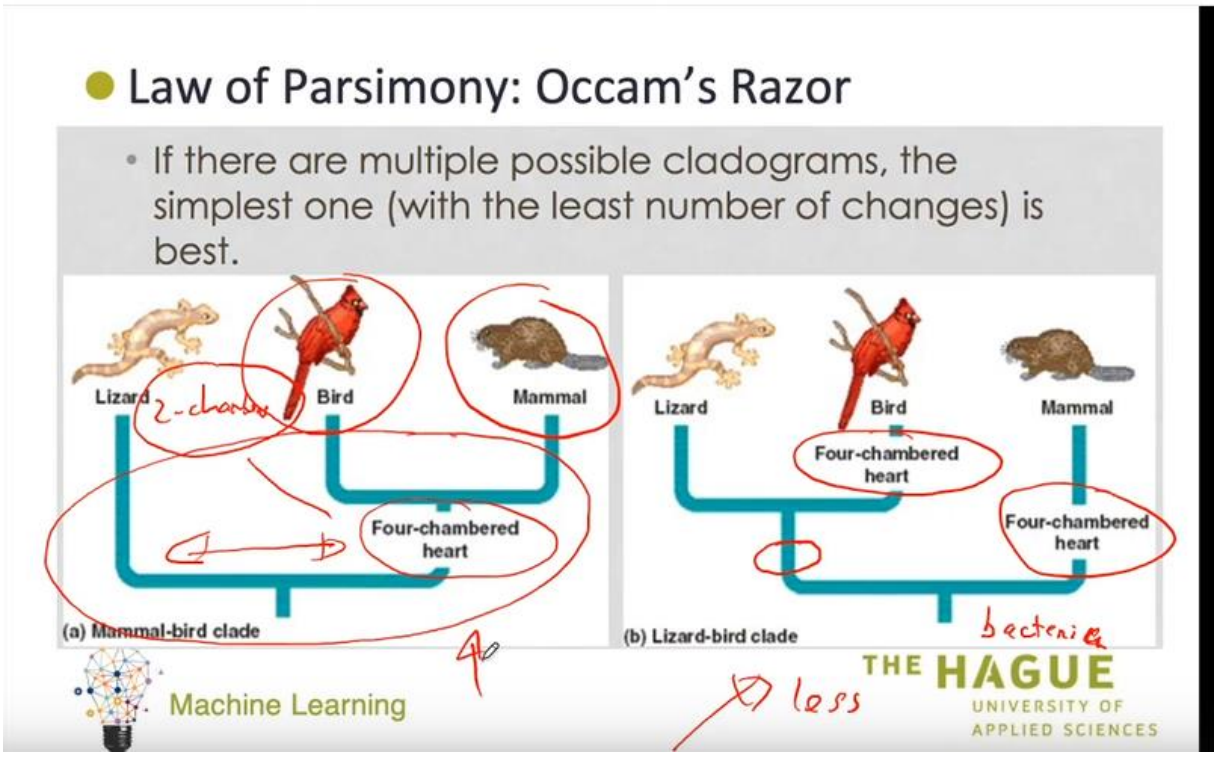
Threshold = threshold bij toetsen is 5.5 maar dat kan ook naar 6 gehaal worden



De boog die het verst is is de betere systeem(alle andere kleuren zijn andere systemen) dit is een evaluatie methode voor classification

Vuistregeles voor machine learning

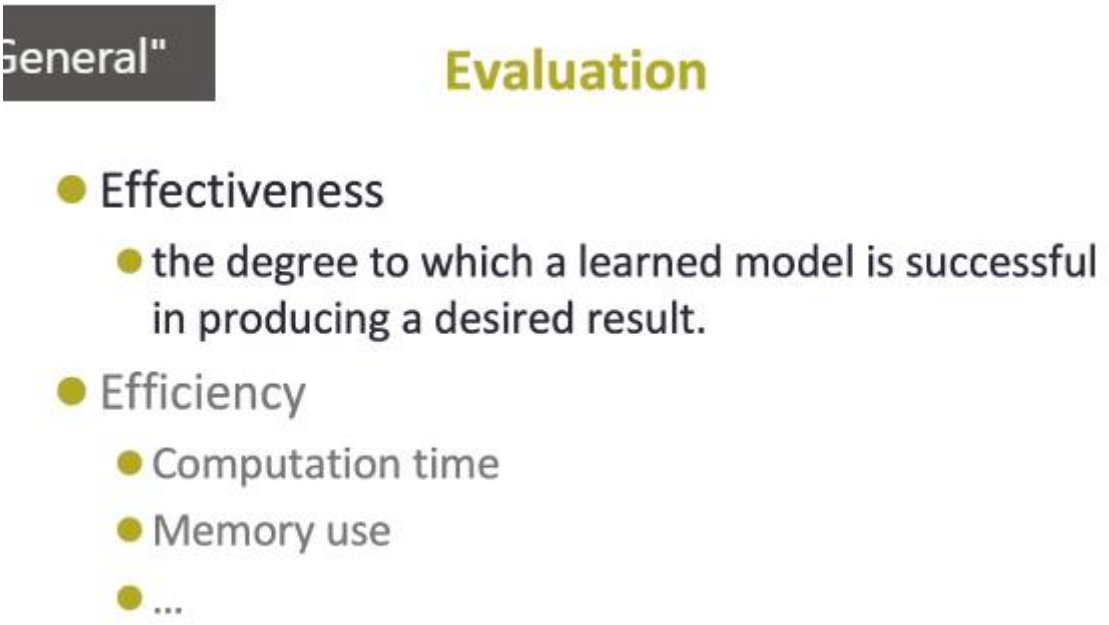




Law of parsimony jij moet degene kiezen zonder cift dikis, dus kijk rechts die heeft 2x de aftakking naar four heart terwijl linker 1 heeft.

Ockhams scheermes is de stelling dat de hypothese gekozen moet worden die de minste aannames bevat en de minste entiteiten veronderstelt, dan is het de juiste.

Wat is een voorbeeld van de law in ML?
over complex models doen het slechter dan simpele modellen

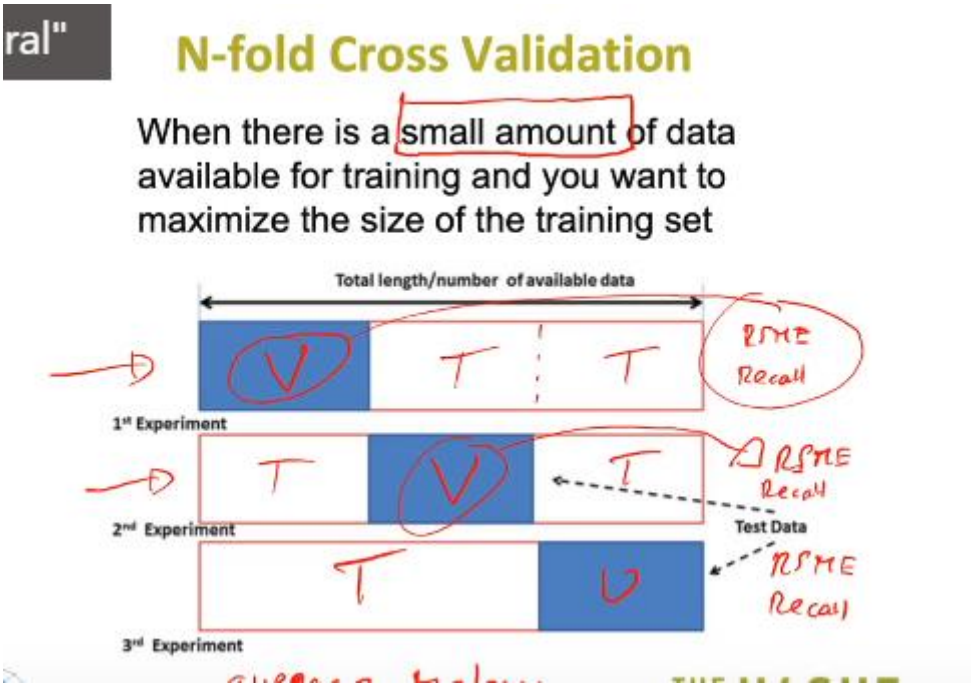


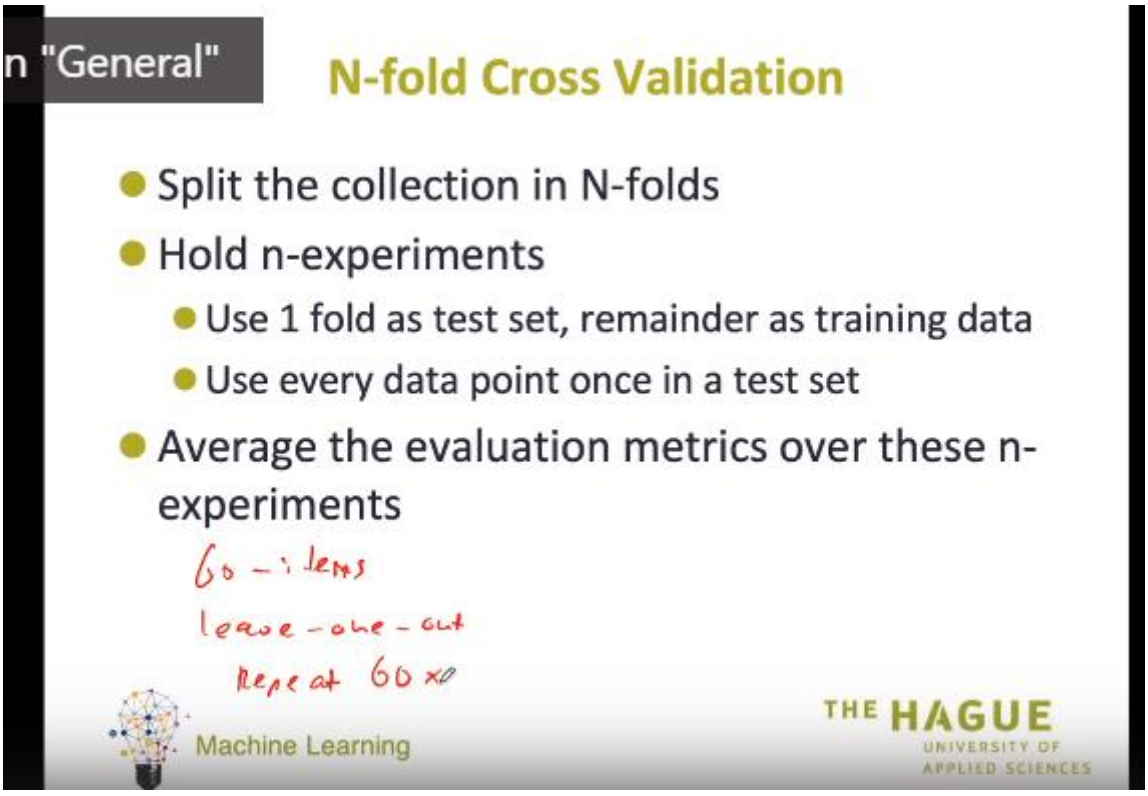
Hoe kunnen we onderwerpen goed evalueren?

Een goede methode daarvoor is

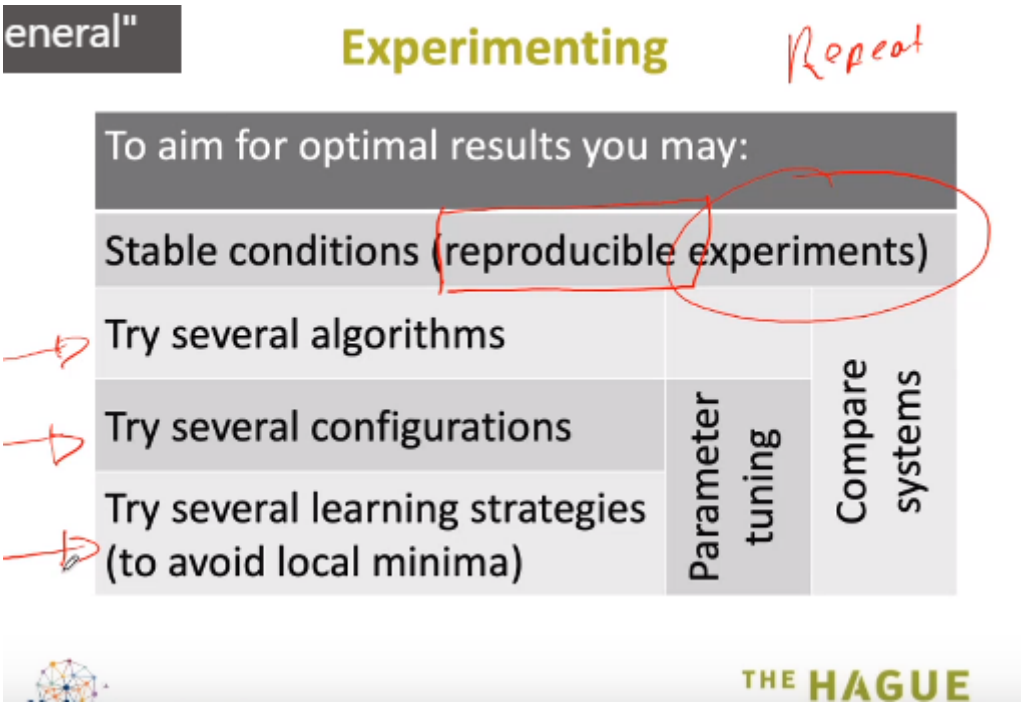
- Cross validation
    - Cross valisation gaat in op de vraag 'how well does a model generalize to unseen data?
        - Je split de data randomly door een traing en validation set
            - De traing geeft een traing aan het model
            - De validation zijn punten die het model nooit heeft gezien dus als we predictions doen op de validation set krijgen we unbiased estimates

Wanneer je een kleine data hebt en wilt de grotte van de trainigset maximaliseren dan maak je gebruik van N fold cross validation





Die N kan je elke getal invullen dus 10 is dat je m verdeelt in 10 stukjes

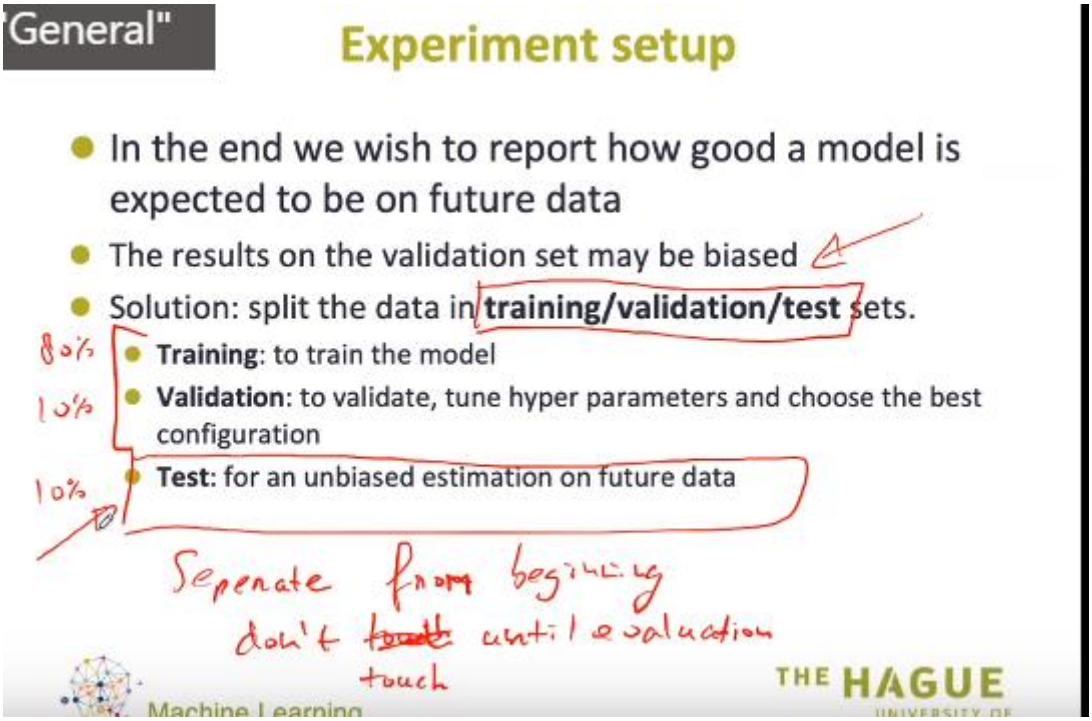


80% van je data inzetten voor training

10% validation

10% test

Maar je voert training en validation eerst goed uit en helemaal als laatst voer je test uit om een optimaal model te verkrijgen.



General"

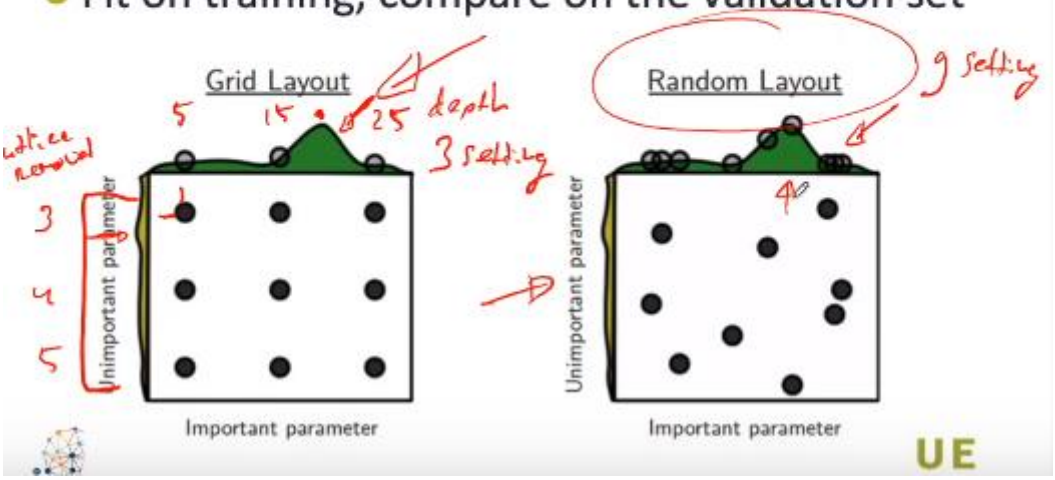
Experiment setup

- Always describe your experiment concisely
  - Dataset, version, url, labels/ground truth, features, record count
  - Preprocessing & cleaning
  - How the data is split train/valid/test
  - What hyperparameters and how they are tuned
- Results should ideally be reproducible!

General"

Hyperparameter Tuning

- Several strategies to tune parameters
- Fit on training, compare on the validation set



eneral" Comparing algorithms/models

- Compare effectiveness between several algorithms
- Especially: compare against clear and well-accepted solutions (baselines):
  - Simple lowerbound ← average
  - State-of-the-art model ←
  - Oracle system ←

Oracle system gaat altijd voor het meest optimale decission

General" Comparing algorithms/models

- Test differences for statistical significance:
  - Reduce the likelihood of false conclusions
  - Hypothesis testing is a formal procedure followed by statisticians to either accept or reject a hypothesis

"General"

Hypothesis testing

- There are two types op hypothesis:
  - H_0: NULL hypothesis, usually that the observed phenomenon is a result of chance
  - H_1: Alternative hypothesis, usually that the observed phenomenon is the result of a non-random cause
- E.g. in a coin toss experiment, H_1 could be the assumption that a coin is not fair i.e. $P \neq 0.5$, H_0 that the coin is fair i.e. $P = 0.5$

"General"

Hypothesis testing

- Similarly for Data Science experiments:
  - H_0: model A is as effective as model B
  - H_1: model A is not as effective as model B
- use statistical significance tests to estimate the likelihood of wrongly accepting H_1
- Take care in which significance test to use!

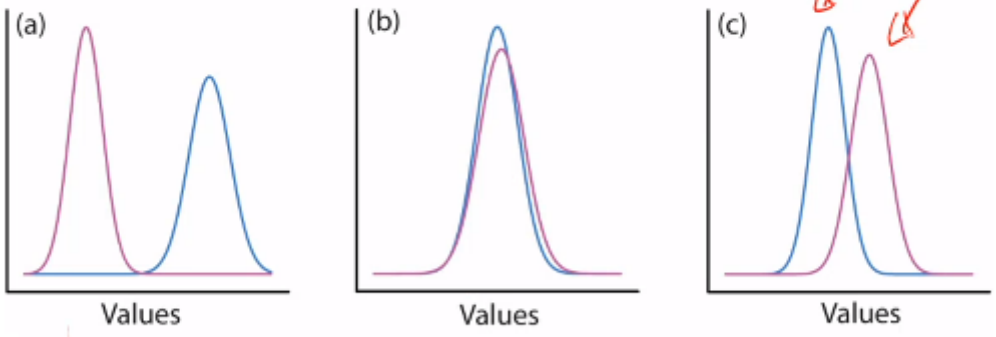
"General"

Statistical Significance

- When are two outcomes significantly different?

tails

heads



Several statistical significance tests -> choose wisely

"General"

Statistical Significance

- Several statistical significance tests -> choose wisely:
  - Student's T-test
  - Wilcoxon signed rank
  - Chi squared
  - Paired T-test
- Threshold for hypothesis rejection:
  - Common p-value < 0.05

5%

chance of drawing a wrong conclusion

General" Statistical Significance

- If you run several trials, outcomes are bound to be different. Is higher always better?
- When things are not interesting:
  - A very small improvement may not be worth the effort
  - If the outcome is more likely due to change than due to improvement
  - When we have overfitted unknowingly

eneral" Statistical Significance

- Evaluation: How do you measure effectiveness?
  - Statistic hypothesis test (e.g. $\chi 2$): rule out the probability of the observed results by chance

System A	Guilty	Innocent
Predicted Guilty	20	10
Predicted Innocent	5	65

Much better than random
P(guilty\|predicted) = 0.67
$\chi 2$ p-value < 0.00001

System B	Guilty	Innocent
Predicted Guilty	10	30
Predicted Innocent	15	45

Not better than random
P(guilty\|predicted) = 0.25
$\chi 2$ p-value = 1

Machine Learning

THE HAGUE UNIVERSITY OF APPLIED SCIENCES

System a is beter als B want als je bij A naar de eerste rij kijkt

Zijn er totaal 30 mensen van de 30 zijn 20 echt schuldig maar zijn 10 ook in jail maar zijn niet schuldig

Terwijl bij B 40 mensn er zijn en daar 10 echt schulig zijn en toch 30 personen in jail zitten

Het antwoord van de formule bij B (x2 p-value) is 1 dat betekent dat de systeem 100% fout geeft

Antwoordt bij A is kleiner dan 0.00001 dus de kant dat er een fout antwoordt wordt gegeven is zeer klein

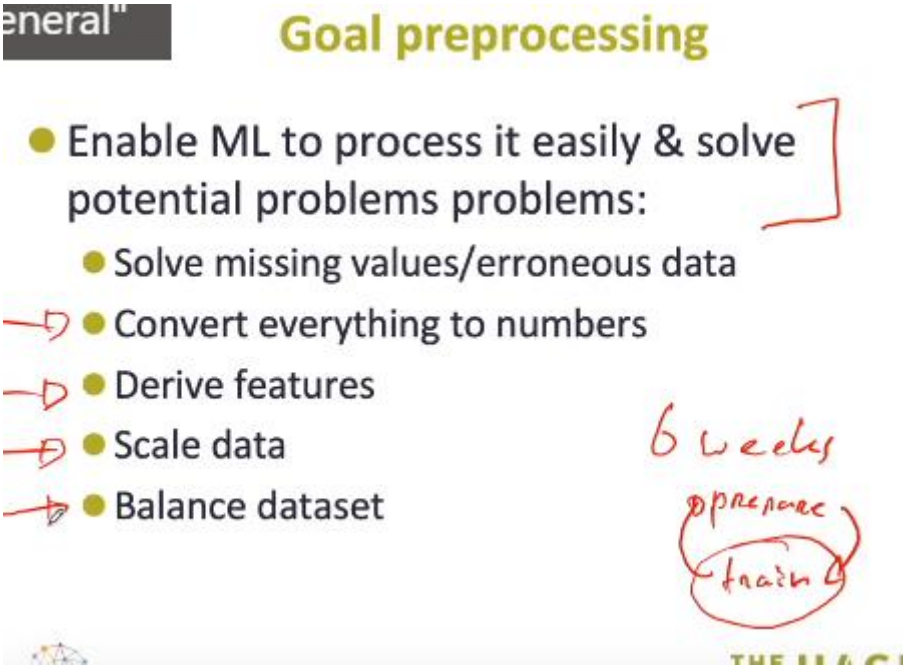
General"

Conclusions

- Be precise in formulating your conclusions:
  - Summarize results in a few sentences
  - State the extent to which the results support your hypothesis
  - If appropriate, state the relationship between the independent and dependent variable.
  - Summarize and evaluate your experimental procedure, making comments about its success and effectiveness.
  - Suggest changes in the experimental procedure (or design) and/or possibilities for further study.

Week 6 data

Datascientist besteden 80% van de tijd aan het verzorgen van de Data



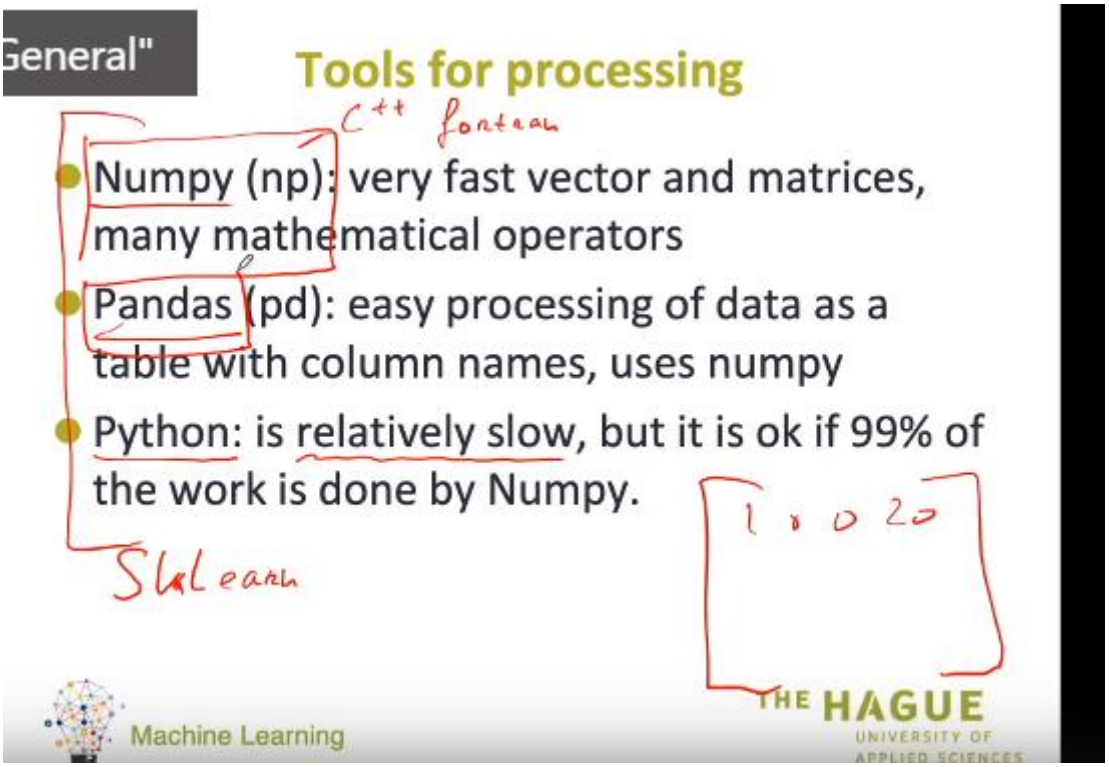
Het doel van voorverwerking is :

Om het ML proces te vergemakklijken en problmen optekunnen lossen.

De problemen die voorkomen zijn ;

- Oplossen van missende waardes of error data
- Alles omzetten naar numbers
- Features engineeren(dus goed na kunnen denken waarom dit waarom dat)
- Data scalen
- Dataset balanceren

De gebruikte tools voor preprocessing



Python is een trage tool maar waarom gebruiken we die? De reden hiervoor is omdat het meeste werk gedaan wordt door Numpy en die tool eigenlijk best wel snel is, en dus niks gemerkt wordt van de trage python.

Hoe ga je om met missing values?

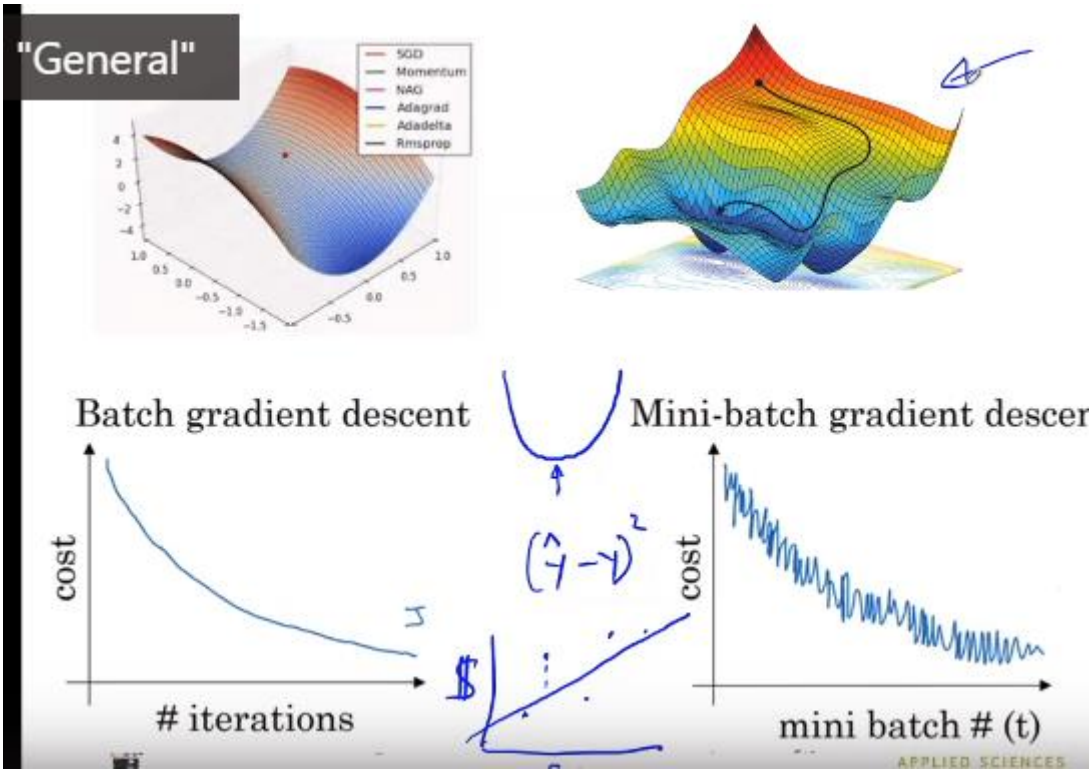
-Remove(weghalen(als je 20 rijen hebt en rij 16 mist haal je die weg)(dit is voor klein beetje goed)

- Impute(invullen(de missing value invullen dmv het gemiddelde van de wel ingevulde values)(je kan ook 0 invullen maar dat hangt van de context af, als het bijvoorbeeld gaat over het gewicht van mensen is 0 geen gode invulling maar het gemiddelde juist weer wel.) (je kan als 3e ook gebruik maken van infer dat leer je de machine om zelf in teschatten wat bij de missing values hoort)

Waarom is bij de 2de tabel geen 3de kolom met ‘expensive _big’? want big wordt gebaseerd op de anderen 2 dus als extreme of small 1 is is de anderen 2 een 0 of als small en extreme 0 zijn dan is big sws 1 snap je.

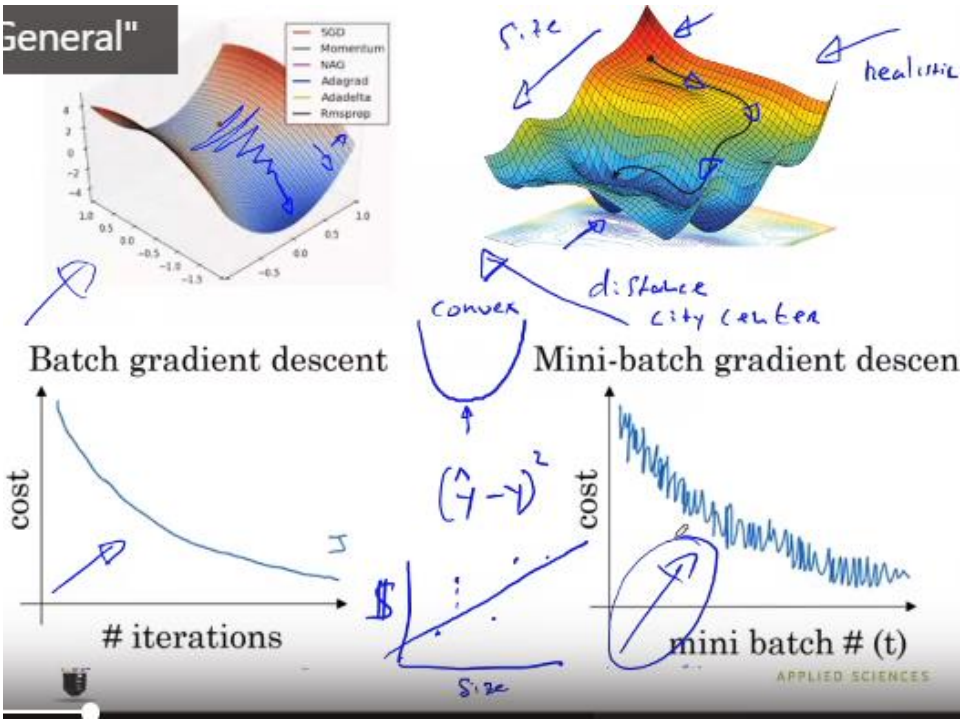
Cost function geeft ons een verschil tussen het predicted en daadwerkelijke gebeurtenis.

(y(predicted) – y(daadwerkelijk)) met een hoge 2 geeft ons een parabool met hen minimum.



Convex is een ander woord voor bol.

Ideaal is een parabool maar in realiteit is het de rechter bovenste foto , de ene x wat trager omlaag dan de andere x



We willen in een model linksboven en rechts onder voorkomen want is niet handig. Maar hoe kunnnen we dat voorkomen?

We kunnen de data scalen zodat het oppervlakt platter wordt.

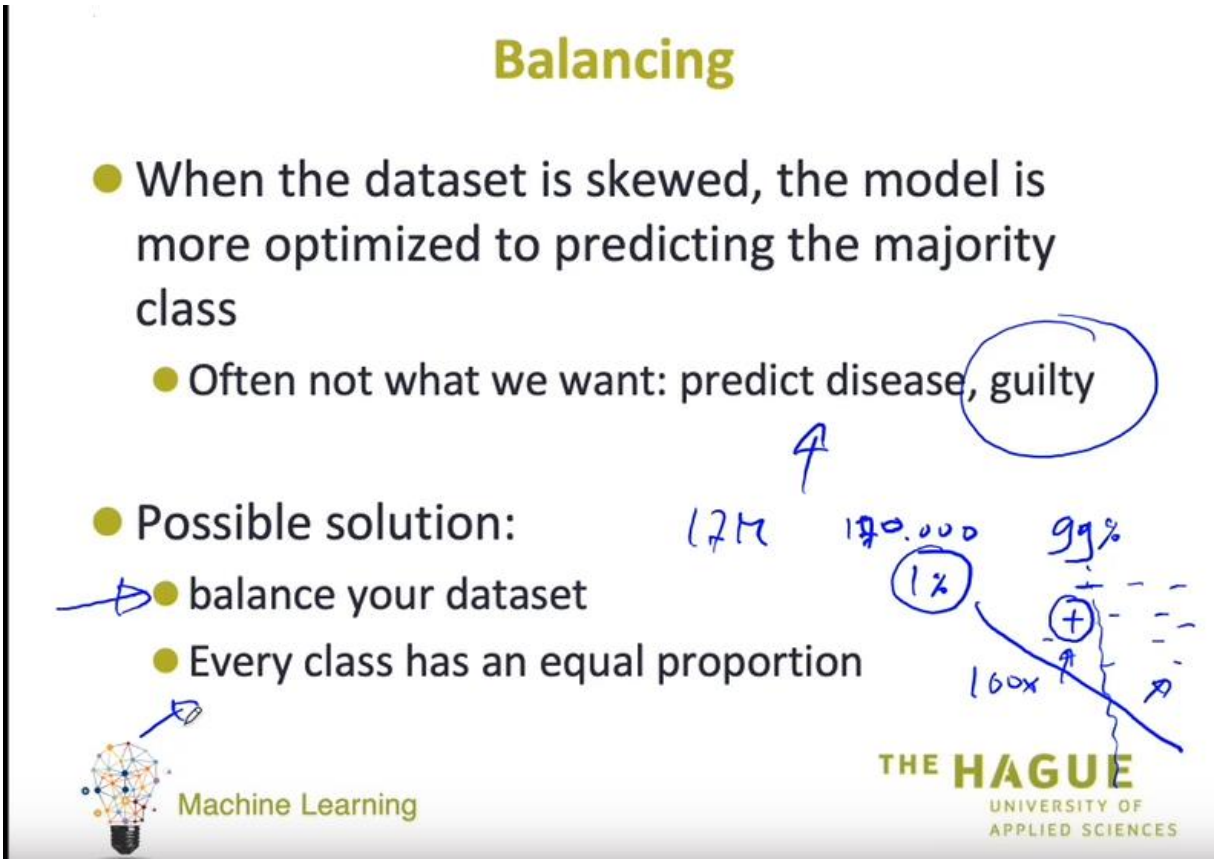
Wannner gebruik je scaling?

- Als je te maken hebt met oscillation(dus dat heen en weer gaande en hiervoor al uitgelegd)
- Als je de globale minimum niet kan vinden
- Als de nummers die er zijn onstabiel zijn(dus als je bij de ene kant praat over vierkante meter(soms is dat wel 1000m2), en het aantal kamers in een oso(maximaal zijn dat meestal 1-7 kamers) hier ontstaat onstabiliteit dus scaling gebruiken.

Scaling werkt tegen overfitting en extrapolation

Scaling maakt van huisvierkante meter van 400 tot 2000(400=-5,2000=5)

Bij gebruik van scaling is het belangrijk om dezelfde mean en variance te gebruiken bij validation&test en training



wannner de dataset scheef is, dan kiest de model in de meeste gevallen vooor de grote groep.

Dit is niet handig voor het voorspellen van guilty en ziektes.(als 1% ziek is zegt het model juist 0% is ziek want hij kist voor het overgrote groep)

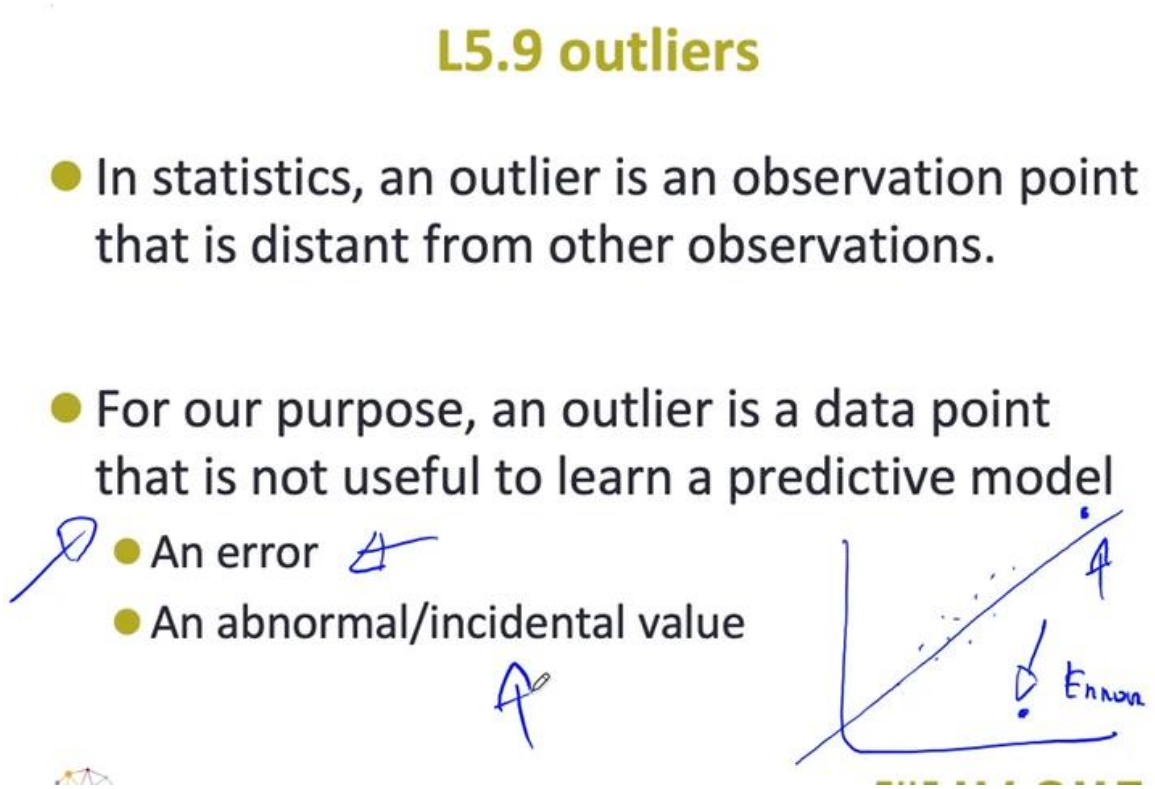
Hoe kan je dit oplossen?

-balancing

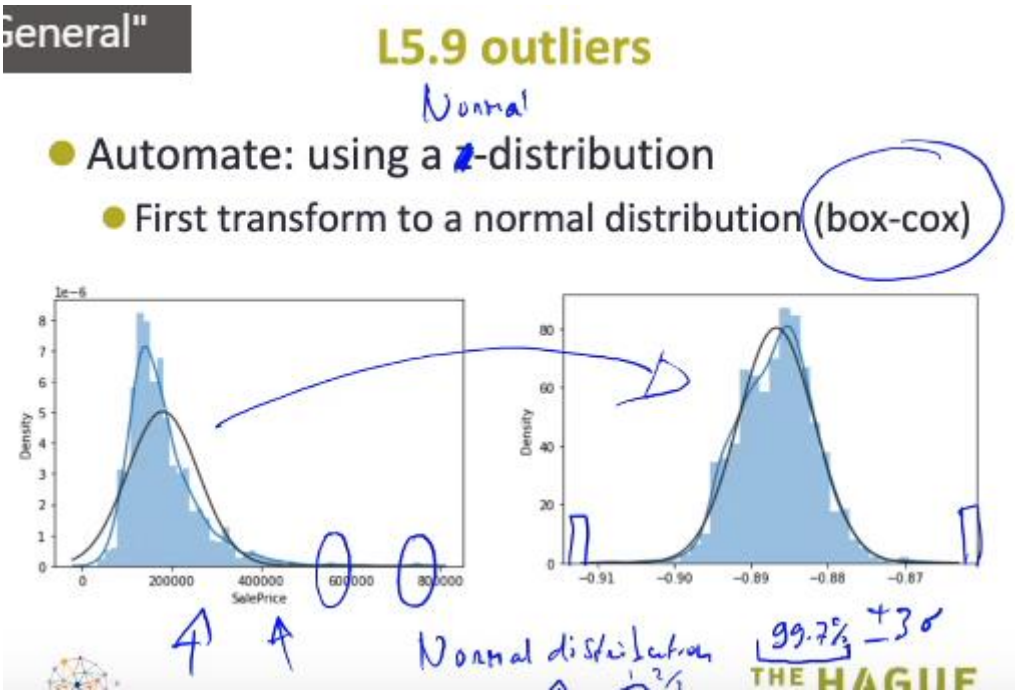
- elke groep heeft zn eigen proportion

Wanneer dien je je model te balanceren?>>
wanneer het verschil tussen 2 klassen te groot is

Outliers



Een outliers is gewoon een uitschieter die niet handig is voor voorspellingen.



Hoe maak je outliers bekend?
bij links zie je outliers niet goed

Je transformeert dat dmv normal distribution en krijg je een wat helder grafiek(rechts)

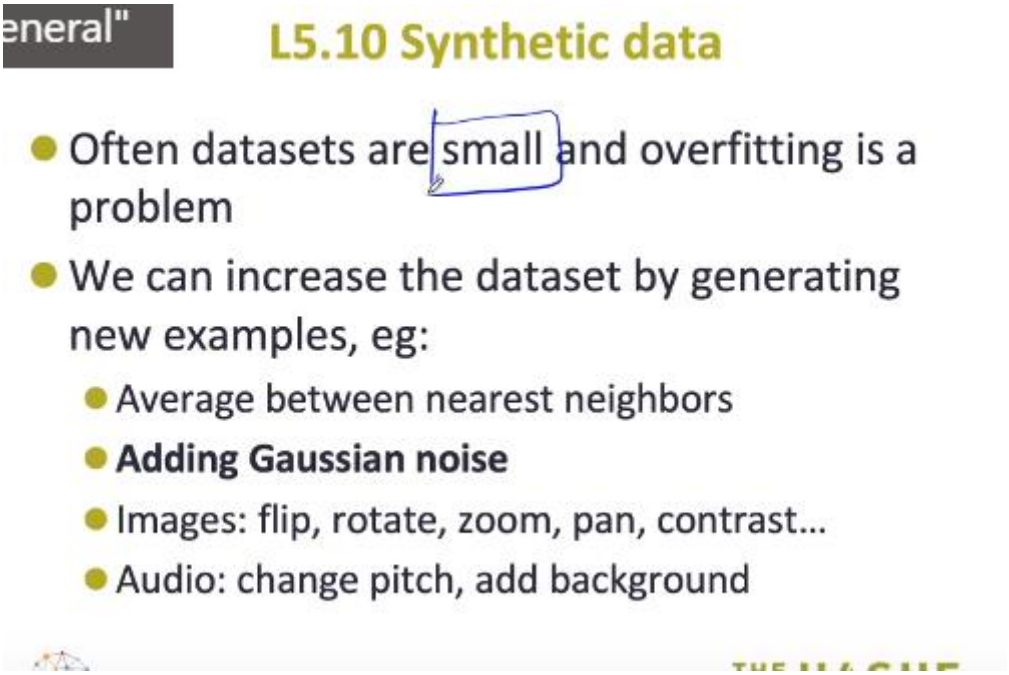
Wanneer moe tje outliers weghalen?

Je moet sws oppassen dat je het model niet gaat extrapoleren.

Synthetic

Syntethic data is data wat niet bestaat eigenlijk maar om het data optevullen doe je dat

Meestal kunnen datasets heel klein zijn of is overfitting een probleem. We kunnen deze problemen oplossen dmv……



- Gemmidddelde tussen nearest neighbors
- Gaussian noise: is extra punten toevoegen in de buurt van een bestaande punt
- Images
- Audio

Week 7 features

Feature is de informatie om voorspellingen te creeren.

- Een kleine deel is critical van de features
- Niet critical kunne leiden naar een aantal problemen
    - Problemen met learning
    - Overfitting
    - Non transparant models
    - Onnodig toewijzing van middelen

Het vinden van een minimale subgroep van features dat leid naar de optimale resultaten

- Het verwijderen van features met lage variatie(manual)
- Filter methode (zet op volgorde kijkend naar verschillende gegevens)(manual)
- Wrapper method(selecteert automatisch wat de optimale features zijn)(automatic)

Het verwijderen van features met lage varatie

Als het boolean waarde heeel scheef is, is er weinig informatie

Dat kan leiden tot overfitting

Oplossing hiervoor is het weghalen van booleans emt lage variatie

Je dient je threshold altijd te tunen bij feature elimination

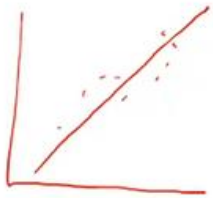
Filter methode

Feature Selection: Filter method

- **Filter method**: rank the features according to distance, information gain, dependency, consistency

Criterium

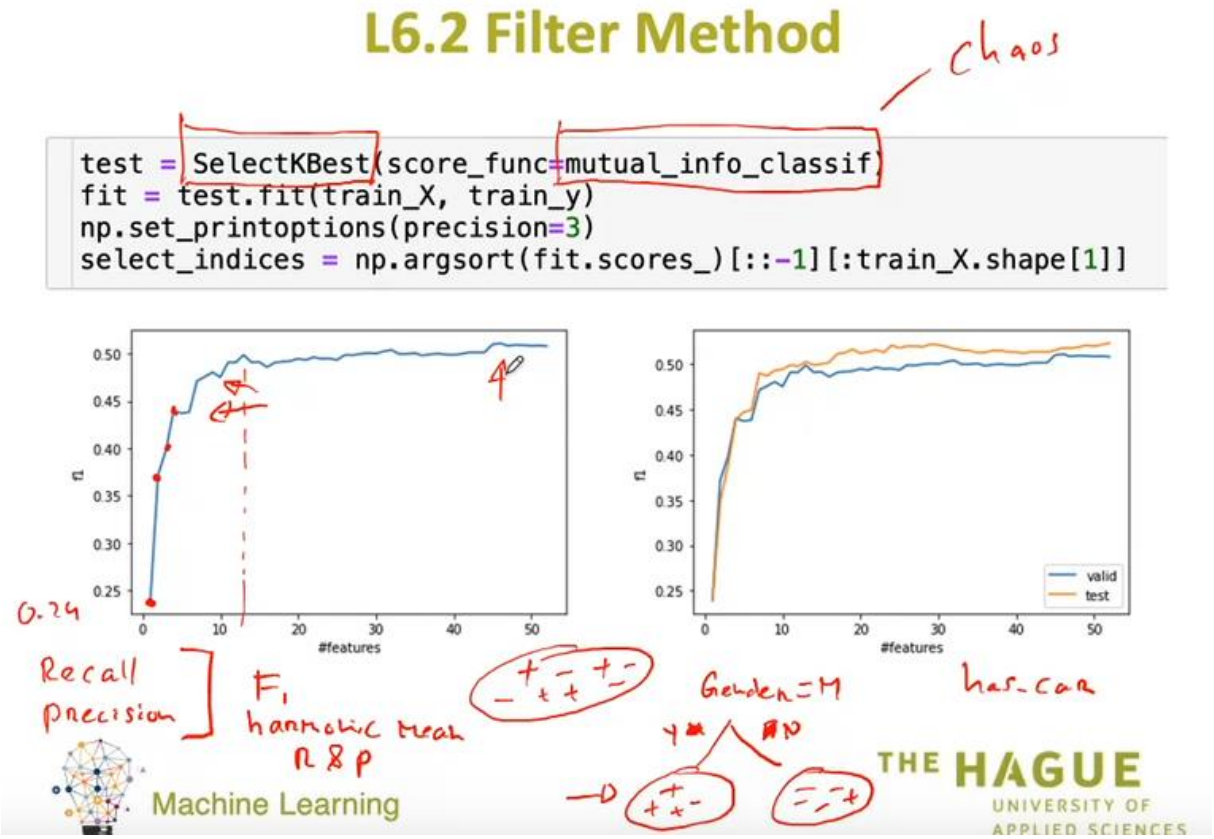
Feature/response	Continuous	Discrete
Continuous	Pearson's corr	Mutual information
Discrete	Anova F-Test	χ^2



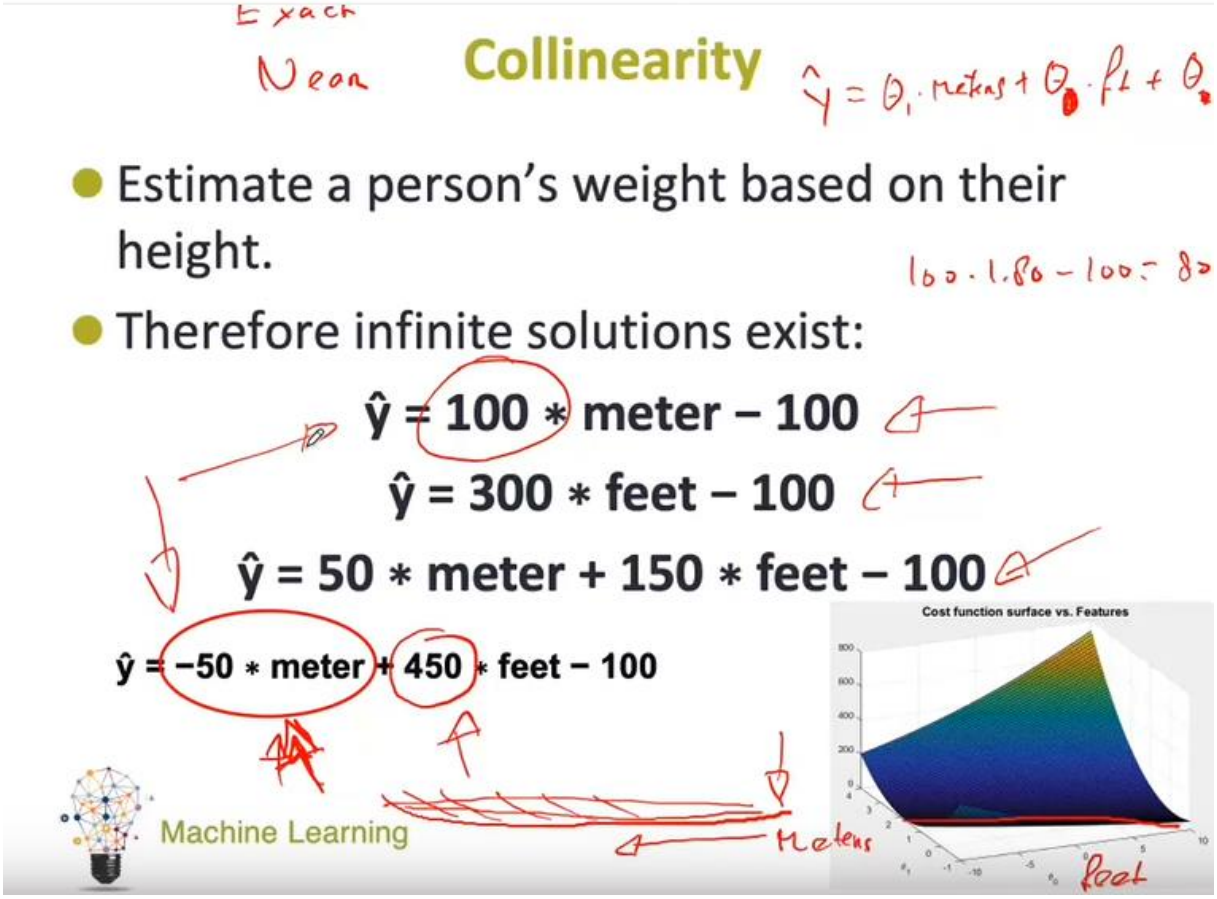
$p \approx 1$

$p \approx 0$

F1 score



Je ziet dat het stijgt maar ongeveer vanaf 15 features blijft de f1 score ongeveer gelijk , hoe komt dat?
de features die er telkesne bij komen zitten al een beetje in features ervoor dus er zal niet veel veranderen. Het is tevens niet handig zoveel features erbij te zetten want dat zorgt voor ‘door de bomen zie ik het bos niet meer’



Collinear zijn features die het meest op elkaar lijken met waardes

Collinearity problemen

- Interpreteren van coefficients is moeilijk, welke feature draagt positeif of negatief bij.
- Convergentie kan een probleem veroorzaken
- Kleine aanpassing in dataset kan resulteren in een hele andere dataset

Er zijn ook uitgebreidde werk bij het collinearity problemen

Oplossingen hierbij zijjn;

- Bij 2 grote aghankelijke features gebruik er maar 1 met behulp van correlation matrx
- Verwijder features met een hoge multicollinearity door gebruik te maken van variable inflation factors
- Verwijder features waarbij een feature teveel lijkt op de andere met zijn waardes
- Chek t test waardes om te zien of een features een verschil aantoont

Met correlation matrix kan je correlatie vinden van 2 features

Conclusie collinearity

-soms is het moeilijk om intezien of een featuer collinear is

- er is geen fixed approach to get the best result

- cross validation is je beste matie

-suggestie : leer moddelen van de vermoedelijke features, bereken de correlation tussen de uitkomsten, is het in de buurt van 1 dan is het collinear

Wrapper methode is methode collinearity automatisch uit te filteren

Is automatisch

- Forward selection = je kiest het allerbeste feature daarna het 2de beste dan 3de etc etc
- Backward elimination = je hebt alle features daarna kijk je wat is het slechtste dus de ene met minste invloed en zo dooor en haal je die allemaal weg

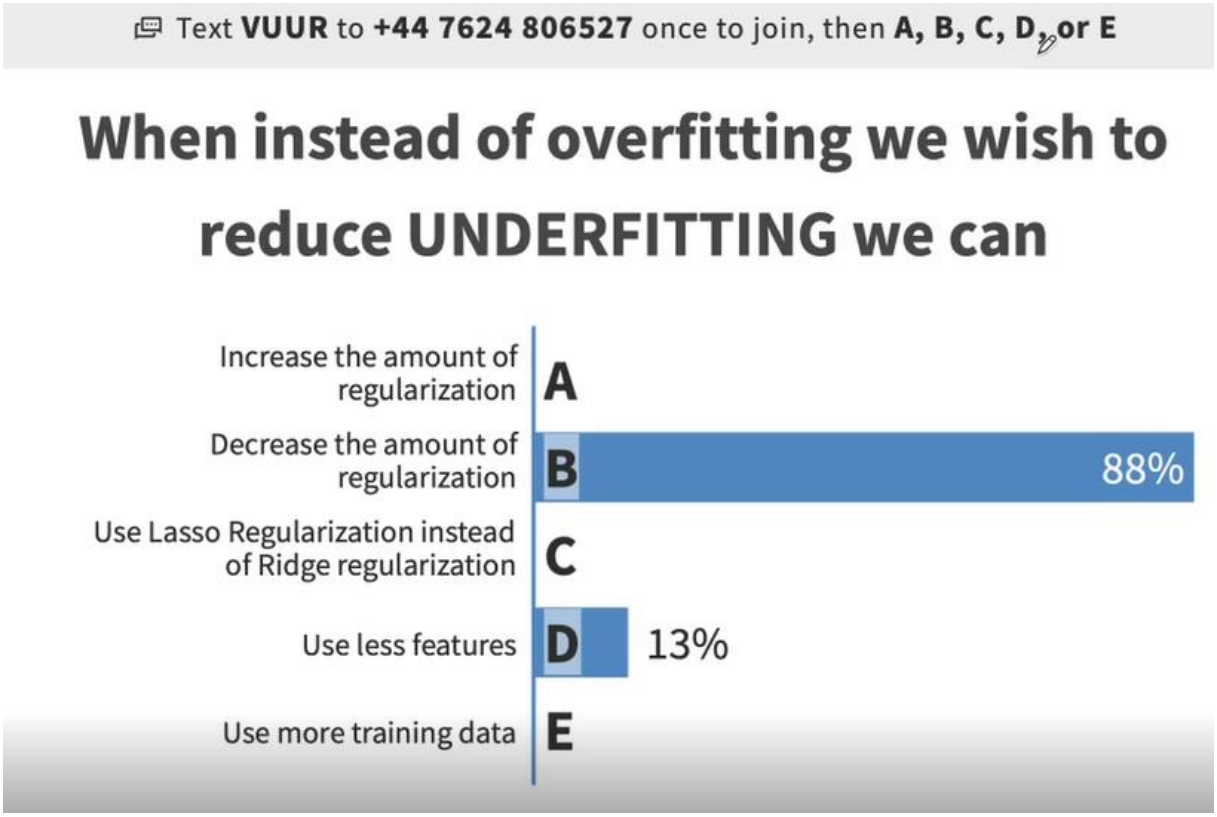
Wanneer we recursive feature elimination doen, kiezen we de feature die het meest effect heeft op…… validation set. Want hiermee zie je in hoe effectief het model is.

Embedded methodes

Helpt bij de bijdragen aan een training

- Regularization(deze helpen bij het voorkomen van overfittin)
    - Lasso
    - Ridge
    - Elastic Net

Regularization helpt het voorkomen van overfitting, feature selection is maar 1 aspect binnen regularization



Hier is antwoord B maar Als je overfitting wilt moet je a kiezen

Feature engineering

Transformeert data naar bruibare signalen

- Speech recognition-> word segmention
- Image recognition ->light correction

Random forest

-decision trees hebben de neiging om te gaan overfitten, tenzij ze ondiep zijn. hierdoor is de power gelmiteerd daarom kan je het volgende gebruiken om het te verbeterern

- Creer vele decission trees(forest(bos))
    - Allemaal getraint op een random subset van een training set
    - Met elke tree met zn eigen subset aan features
- Neem het gemmiddelde van alle trees

Random forest is een ensamble(combinatie van verschillende decision trees)

Adaboost is een linieare mix van verschillende modellen

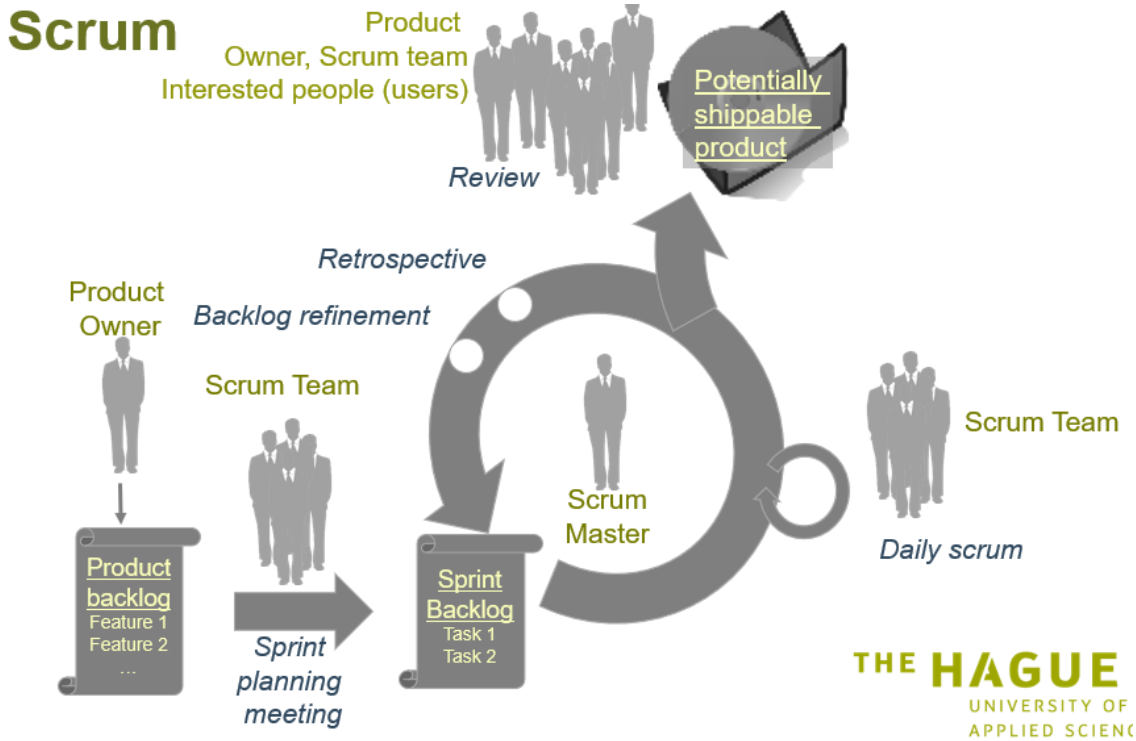
Tekst classification

Van docu naar vector space representation

Je hebt een docu en een woordenlijst, bij die woordelijst staat hoevaak een woord voorkomt in een docu

Waar is dit handig voor ? zo kan er voorspeld worden waar het docu over kan gaan.

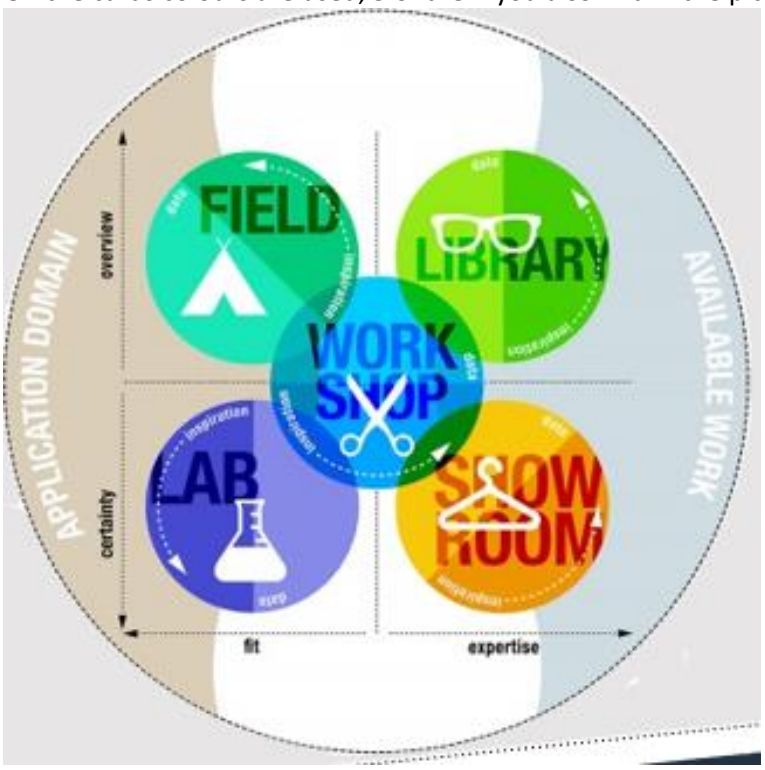
Scrum



Workshop Research 1

The DOT frameworks aim is to help structure research

On the cards colours are used, 5 of them you also find in the picture below



The 5 colours above are called research strategies

- **Library research** is done to explore what is already done and what guidelines and theories exist that could help you
    - The term 'library' could be misleading, interviewing experts, for example, is also a valid activity in 'the library'
- **Field research** is done to explore the **application context**. You apply a field strategy to get to know your end users, their needs, desires and limitations as organizational and physical contexts in which they will use your product
- **Workshop research** is done to explore opportunities. Prototyping, designing and co-creation activities are always to gain insights in what is possible and how things could work
- **Lab research** is done to test parts or concepts of your product, of the final product. You use lab research to learn if things work out the way you intended them, or to test different scenarios
- **Showroom research** is done to test your ideas in relation to existing work. Showing your prototype to experts can be a form of showroom research or spelling out how your product is different from the competition. Also testing your product to general guidelines is a form of showroom research.

More dimensions:
Fit – Expertise

LEFT	RIGHT
"Application domain"	"Available work"
Daly practice	Theoretical fundaments
Optimize the fit between product and the application context	Optimize to contemporary quality standards

Methods rarely optimize both sides, so you may need to combine methods, to get a good mix

The white space between the available work and the application domain is the 'gap between theory and practice'. This is the innovation space

Overview – Certainty

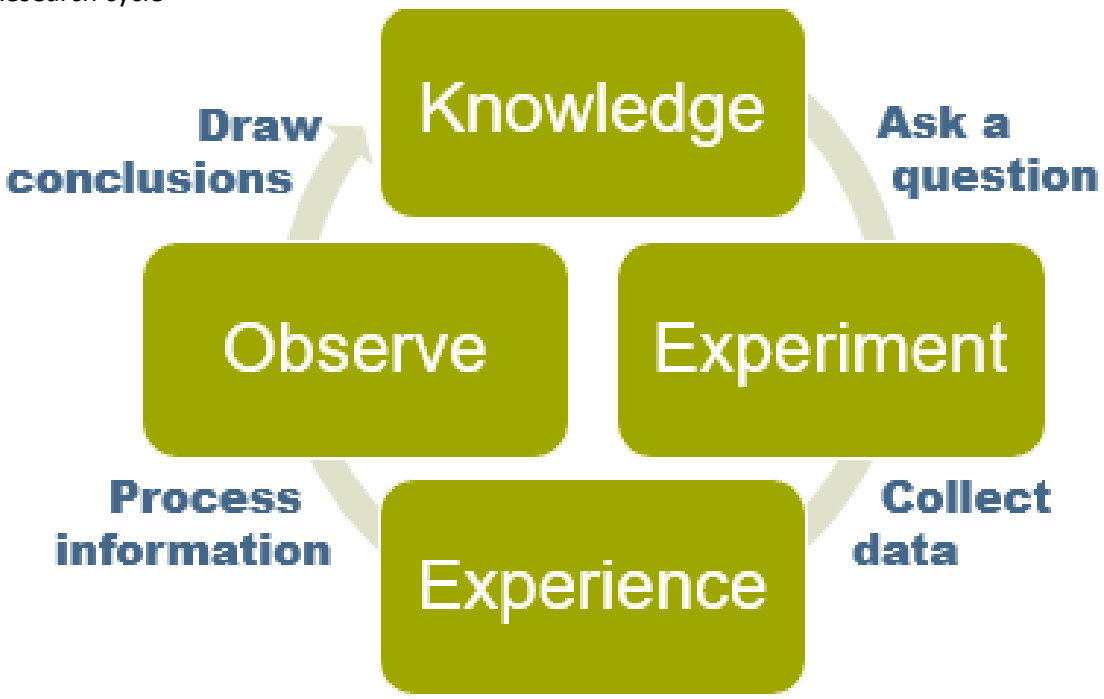
TOP	BOTTOM
Overview	Certainty
Orienting, discovering, exploring	Testing, evaluating, validating
Chances, opportunities	Are your conclusions correct??

Inspiration – Data

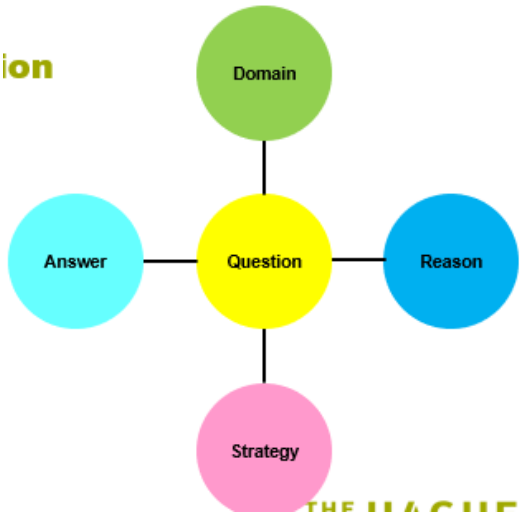
Available in all strategies	
Inspiration	Data
Creativity, personal involvement, intuition	Facts, rational, objective

Workshop Research 2

Research cycle



Ask a question:



Domain → Anchored
Reason → Relevant
Strategy → Functional / efficient
Answer → Precise

Quality

- Controllable
  - Be open on what you do and why
- Skilfully
  - Effective, efficient and permissible
    - Acceptable in the domain
    - Legal
    - Ethical
- Logical
  - Reason correctly
- Valid
  - Systematic fault
    - Measuring something else
  - Most cars hit the break before an accident. Breaking is dangerous!
- Reliable
  - It cannot be a coincidence
    - You can have a day off
    - Your watch is broken, you are measuring the time wrong
- Adequate
  - You're answering the wrong question

Do's and Don'ts

You are bridging the gap between a theory and a (new) application domain. You need to be convincing, and reliable!

- Be open about your research
- Explain what you did (exactly) and why
- Don't hide negative results

Data visualisation

- The human brain is good at the interpretation of images. Its more easy to see patterns in images than by looking at numbers
- The 'trick' of making a good visualisation is:
    - To convey the data correctly
    - To make it pleasing to look at
    - Not to mislead
    - To let it support a clear message

Uses of data visualisation

- Exploratory
  Gaining insights in the data before you decide how you can process the data. (distribution, outliers, …)
- Explanatory
  Presenting and explaining results to stakeholders / public

Scales

- A start, when making a visualisation, is to look at the types of variables, since not all variables should be represented the same way.
  (and for statistics, not all statistical values can be calculated on all types of variables)

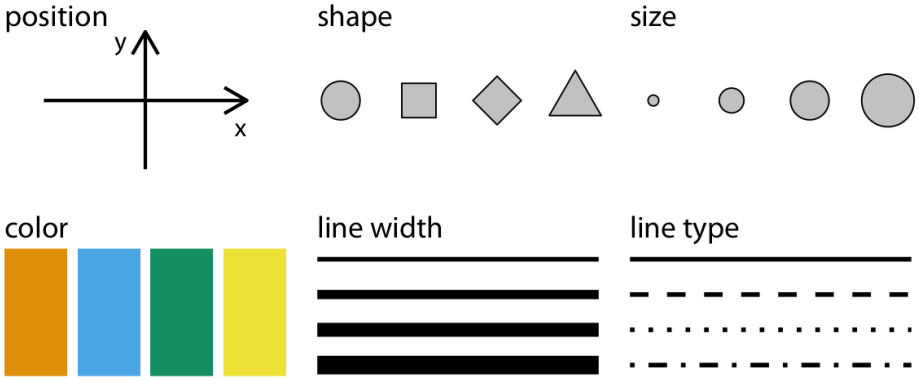
Type of variables	Example	Description
Nominal scale	Male / Female Village / Town / Suburbs	Named labels for classification. No order. You can calculate the mode. (but not the median or mean)
Ordinal scale	Very Unsatisfied / Unsatisfied / Neutral / Satisfied / Very Satisfied	Named labels like nominal, but these have an order. You can calculate the median (but no mean) The difference between two classes has no meaning.
Interval scale	The year of a date Time Temperature in Celsius or Fahrenheit	Named + ordered + the difference has a meaning, is proportional. You can calculate the mean.
Ratio scale	Length The number of correctly answered question	Named + ordered + proportional + a 'natural' value 0.

'0'is bij interval geen afwezigheid je hebt 0 graden maar die is er wel
bij ratio wel number of correct answer 0 is afwezigheid

Aesthetics

- There are many aesthetics elements that can make up a visualisation. See some of them below

- You can see that some types of aesthetics match some scales better than others



Although there are no ‘laws’…

- Using a shape to distinguish between values in a nominal scale, seems logical
- To distinguish values in an ordinal scale, size seems a more appropriate choice
- Lines are typically useful for interval and ratio scales

We can choose how these aesthetics map on the variables we want to visualise. For example, colour seems perhaps a bit more ‘nominal-like’. But if we use a colour gradient, we could use it to visualise an interval scale such as temperature

Dataviz Ducks

- Use the aesthetic elements to make your visualisation clear and pleasing for the eye
- But… make sure these aesthetics have a meaning in the data representation

Types of graphs

- First you decide on what you want to show, the function on the visualisation
- Then you can search for a type of graph

Visualise amounts

- Bar charts
  - Grouped bars
  - Stacked bars
- Heatmap
- Dots / lollipop

Visualise distributions

- Histogram
- Violin chart
- Boxplot
- Choropleth map

Visualise proportion

- Pie chart
- Bar chart
- Tree map
- Alluvial Diagram

Visualise x-y relations

- Scatter plot
- Contour plot

- Bubble chart
- Arc diagram

Visualise geospatial data
- Pin map
- Choropleth map

Elements of a visualisation
- A visualisation should speak for itself. The visualisation must tell a story, you're missing a change when you have to explain the visualisation!
- To make a visualisation clear you
    - Can add a title
    - Should add axis with tick-labels (values)
    - Can add axis titles
    - Add a legend
    - Choose the right type of chart
    - Add elements like trend lines or error bars when appropriate
    - Choose the correct axis
    - Use colours, shapes and other aesthetics correct

Complex visualisations
- With visualisation tools you can create very complex figures but keep in mind that complex figures are also hard to understand
- Keep it as simple as possible, to make a clear statement

Multi panel figures
When combining multiple figures in one visualisation:
- Use the same colour / shape for the same variables
- Make sure that axis are scaled the same if they represent the same kind of value
- Align the figures nicely
- Be consistent, but do not use the same type of graph over and over again. A dashboard with just bar graphs is boring!