Normaliseren

Bron - https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0

Normaliseren is het 'normaal' maken van de data als er geen normale distributie is. Dat betekent dat aan alle waarnemingsgetallen een nieuwe waarde toegekend moet worden zodat ze wel met elkaar vergeleken kunnen worden. Neem bijvoorbeeld de hartbewaking. Iedere patiënt heeft een eigen hart. Dus de apparatuur moet ingesteld worden op dat specifieke hart. Anders geeft de hartbewaking soms vals alarm.
Er zijn meerdere manieren om tot een 'normaal' getal te komen, hieronder worden enkele uitgelegd:

- **Min-Max normalisatie:** hier wordt de data geschaald tussen 0 en 1.
- **Standaardiseren**: hier wordt het gemiddelde op 0 gezet en de standaarddeviatie op 1 voor alle waardes. Trek het gemiddelde

Ensemble methods

Bron : https://pythoncursus.nl/ensemble-methods/

Dit is een onderdeel van machine learning waarbij je individuele modellen combineert om samen betere voorspellingen te doen. Ensemble betekent letterlijk "samen, op dezelfde tijd". Het wordt voornamelijk toegepast op simpelere modellen die niet de gewenste kwaliteit behalen met hun output.

Er zijn 3 ensemble methodes:

- **Stacking:** er wordt getraind op dezelfde data met verschillende algoritmes. De output van alle modellen samen zijn de input van het uiteindelijke model. Deze methode is minder nauwkeurig.
- **Bagging:** er wordt getraind met hetzelfde algoritme, maar met verschillende data. De data wordt opgesplitst in 'subsets' en per subset laat je het algoritme erop los. Het gemiddelde van alle resultaten is het uiteindelijke resultaat. Bagging staat voor bootstrap aggregating. Bootstrapping is een statistische methode om steekproeven te doen, aggregating staat voor verzamelen. Random forest is eigenlijk een decision tree met bagging.
- **Boosting**: er wordt getraind met de onjuiste voorspellingen van het algoritme wat eerder gebruikt is. Dus alle onjuiste voorspellingen worden meegenomen en zullen worden aangepast in het nieuwe model. Boosting betekent letterlijk stimuleren, versterken.

Performance measures

Bronnen:
https://www.kaggle.com/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model en
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
Om modellen te kunnen evalueren of ze goed werk verricht hebben, zijn er performance measures. Verschillende methodes worden gebruikt voor verschillende modellen/problemen.

Voorbeelden van performance measures:

- Confusion Matrix
- F1 Score
- Gain and Lift Charts
- Kolmogorov Smirnov Chart
- AUC – ROC
- Log Loss
- Gini Coefficient
- Root Mean Squared Error (RMSE)

Unsupervised learning

Bron-https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
Unsupervised learning is machine learning met niet-gelabelde data waar onderliggende patronen in gezocht worden. Bijvoorbeeld een webwinkel die gegevens verzameld van klanten en het in groepen opgedeeld en hiermee gepersonaliseerde suggesties kan doen. Omdat er geen controle data is, is het moeilijk om voorspellingen te meten zoals met performance measures bij supervised learning.

Clustering

https://www.geeksforgeeks.org/clustering-in-machine-learning/
Dit is een unsupervised learning methode. Hierbij verdeel je datapunten in meerdere groepen (het lijkt een beetje op classificatie) waarbij de data punten op elkaar lijken. Denk aan bijvoorbeeld een extra suggestie op Google als je iets opzoekt.
K-means is een voorbeeld van een unsupervised learning algoritme (clusteren). Hierbij wordt geclusterd op gelijkenis van de kenmerken. K betekent dat clustering tot doel heeft om gegevens voor verschillende groepen te zoeken.

Cross Validation

Bron - https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Dit is een manier om te valideren of een getraind machine learning model ook goed presteert op nieuwe data die het aangeboden krijgt. Het neemt verschillende train en test samples van de data.

Neem bijvoorbeeld: Leave One Out Cross-CV. Hier train je het model met alle datapunten, behalve 1. Met dat ene overgebleven datapunt valideer je het model. Dit doe je net zo vaak tot je met elk datapunt het model gevalideerd hebt.

Of in K-fold-CV splits je de train-set op in K-aantal folds. Je traint met het K-aantal folds, min 1, zodat je met de laatste fold kan testen/valideren. Op deze manier beperk je de rekenkracht die nodig is en zorg je er ook voor dat er wat meer bias in zit doordat je een grotere test-set hebt.

Overfitting, underfitting

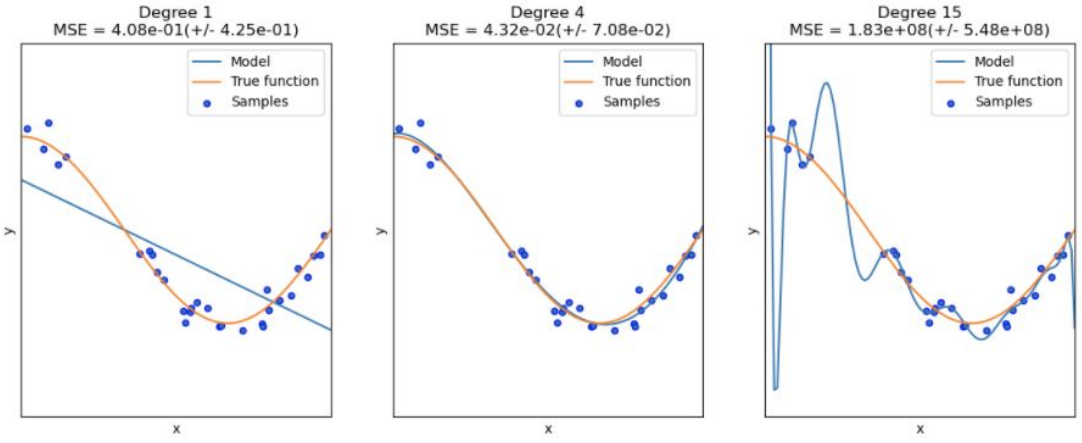
https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Overfitting

Dit is wanneer het model zichzelf de 'noise' aanleert en het het te goed doet. Dit is te zien in de rechter grafiek hieronder.

Underfitting

Underfitting is het tegenovergestelde van overfitting, dus dat het zichzelf te weinig aanleert. Denk aan bijvoorbeeld te weinig train-data waardoor het model niet genoeg informatie heeft en moeilijk voorspellingen kan doen. Dit is te zien in de linker grafiek hieronder.



Parametrische en non parametrische toetsen

Bronnen - https://www.worldsupporter.org/nl/chapter/72594-h15-wat-zijn-non-parametrische-toetsen en

Een parametrische toets is een statistische toets waarbij aangenomen wordt dat de onderliggende verdeling op een of meer parameters na bekend is. De nulhypothese die getoetst wordt, veronderstelt bepaalde waarden voor een of meer van deze parameters.

Wanneer de aanname van de onderliggende verdeling de juiste is, biedt een parametrische toets meestal een hoger onderscheidingsvermogen dan een verdelingsvrije toets. Is deze veronderstelling echter onjuist, dan wordt een systematische fout gemaakt. Het hangt van de context af of een parametrische keuze verantwoord is.
Denk aan: t-toets, f-toets of z-toets

Non-parametrische toetsen worden gebruikt als de scores niet numeriek maar ordinaal zijn of als er niet aan de parametrische toets-eisen voldoet. Een voorbeeld is chi-score of de binomiaal toets.

Standaard afwijking

Bron - https://wiskundeacademie.nl/vwo-a/standaardafwijking
De standaardafwijking wordt gebruikt om de spreiding – de mate waarin de waarden onderling verschillen – van een verdeling aan te geven. De standaardafwijking wordt in dezelfde eenheid uitgedrukt als de verwachtingswaarde of het gemiddelde.