In deze paragraaf ga je leren een simpel machine learning programma te maken. We zullen het zo eenvoudig mogelijk houden, want het wordt al snel erg lastig en zou niet meer geschikt zijn voor de middelbare school.

Voordat we kunnen beginnen is het nodig om anaconda geïnstalleerd te hebben (zie hoofdstuk 3 paragraaf 2). We zullen alle code gaan schrijven in jupyter notebook. Het notebook voor deze paragraaf is te krijgen via … en de dataset die erbij hoort via … . Het notebook geeft een voorbeeld van lineaire regressie, dat is het opstellen van een rechte lijn aan de hand van een aantal gegeven waardes. Het lees prijzen in uit een soort excel bestand waar een huis met een bepaalde grote bij hoort. Het model gaat vervolgens uitrekenen welke rechte lijn het dichtbij alle gegeven punten ligt. Je zou dan kunnen voorspellen welke waarde bij een gegeven prijs of grote van een huis past door het af te lezen uit de grafiek.

Je kunt dezelfde code vrij gemakkelijk aanpassen voor een eigen dataset. De dataset moet dan wel voldoen aan een lineaire relatie, oftewel er moet een rechte lijn zijn die dicht bij de waardes in de dataset moet kunnen komen. Om het model aan te passen hoef je dan alleen maar de variabele ‘data’ gelijk te zetten jouw dataset als je deze een andere naam wilt geven dan ‘dataset.csv’. Let op dat het wel een .csv (comma seperated values) bestand moet zijn, je kunt ook een excel bestand omzetten naar een csv bestand. Je moet dan de waardes voor de x- en y-as kiezen en de naam ervan veranderen op de plekken waar nu x en y staat, denk ook de namen voor de assen in de grafieken. Vervolgens moet je dan nog de waardes voor de coëfficiënt van const en jouw gekozen variabele aflezen en deze invoeren in de formule yhat (yhat betekent de voorspelde y waarde).

Vragen

1. Wat is een dataset?
2. Wat is lineaire regressie?
3. Hoeveel zou een huis met grote 800 waarschijnlijk kosten?
4. Hoe groot zal een huis met als prijs 400000 waarschijnlijk zijn?
5. Dan nu een uitdaging. Maak zelf een dataset (verzin hiervoor wat waardes of zoek er een op) en pas het model aan zodat het werkt voor jouw nieuwe data.

Antwoorden

1. Een aantal gegeven waardes.
2. Lineaire regressie is een methode om een rechte lijn op te stellen dat zo dicht mogelijk langs alle gegeven punten komt.
3. 300000
4. 1300
5. Tip: pas regel 9 (optioneel), 12, 13, 17, 18, 32 (blok 2 regel 3), 34 en 35 aan.