Vergelijkende studie van voorspellingsmodellen voor tijdreeksen

Onderzoeksvoorstel Bachelorproef 2019-2020

Emiel Declercq[1]

Samenvatting

Artificiele intelligentie heeft zich de laatste jaren ontwikkeld als een van de belangerijkste domeinen in de hedendaagse technologie. Maar om ervoor te zorgen dat de mogelijkheden van artificiele intelligentie optimaal benut worden moet er ook rekening gehouden worden met de opkomst van nieuwere modellen die voor bepaalde toepassingen misschien veel beter zouden kunnen zijn dan de klassiekere modellen. Dit zal onderzocht worden voor tijdsafhankelijke data.

Sleutelwoorden
Machineleertechnieken en kunstmatige intelligentie — Machine learning — AI

Co-promotor
Stijn Lievens[2]

Contact: [1] emiel.declercq@student.hogent.be; [2] stijn.lievens@hogent.be;

Inhoudsopgave

1. Introductie

Artificiele intelligentie wordt steeds meer toegepast dus de optimale methodes bepalen om voorspellingen te maken is van vitaal belang. Verschillende datasets kunnen er volledig anders uitzien en deze kunnen dan ook op verschillende manieren ingedeeld worden. Voor dit onderzoek zal gefocust worden op data die tijdsgebonden is. Door het gebruik van dit type data zullen de modellen rekening moeten houden met de tijdsafhankelijkheid tussen de verschillende waarden.

2. Stand van zaken

Er zijn heel wat methoden die kunnen toegepast worden om een voorspelling te maken van tijdsgebonden data. Voor deze paper zullen enkel polynomiale vergelijkingen, ARIMA en LSTM getest worden.
De meest primitieve manier om een trend te voorspellen is het fitten van een polynomiale vergelijking op de trainingsdata en deze nadien toe te passen op de testdata. Daarnaast kan ook de ARIMA-methode (Brownlee, 2018) gebruikt worden ofwel het Autoregressive Integrated Moving Average. Deze methode combineert autoregressie en voortschrijdend gemiddelde. Autoregressie modeleert de volgende stap in een sequentie als een lineaire functie van de waarden uit voorgaande tijdspannes. De methode van het voortschrijdend gemiddelde modeleert de volgende stap in de sequentie als een lineaire functie van de resterende fouten van een gemiddeld proces bij voorgaande tijdspannes. Er moet ook opgemerkt worden dat er een verschil is tussen een model met een voortschrijdend gemiddelde en het voortschrijdend gemiddelde van de dataset zelf.
Ook neurale netwerken kunnen toegepast worden bij het maken van voorspellingen van tijdreeksen. LSTM (Long Short Term Memory) is een vaak gebruikt modeltype om tijdreeksen te voorspellen. Dit model zal het verloop van de volgende waarden voorspellen op basis van de ingevoerde waarden rekening houdend met de chronologie waarin ze voorkomen. Hierbij zal de invloed van oudere waarden minder relevant worden naargelang er meer waarden ingevoerd worden.
Op zowel de ARIMA als de LSTM modellen bestaan er varianten om multivariate tijdreeksen te voorspellen, bij ARIMA worden deze benoemd als VARMAX modellen. Ook bij polynomiale regressie kunnen multivariate times series voorspeld worden. Ook voor tijdreeksdata waar een duidelijk seizoenseffect zichtbaar is bestaat er een variant op het ARIMA model genaamd SARIMA.

3. Methodologie

Om na te gaan welke methodes de beste resultaten behalen zullen zowel polynomiale regressie, ARIMA en LSTM toegepast worden op 2 datasets, 1 waarbij een duidelijk seizoenseffect zichtbaar is en 1 waar geen duidelijke sei-

zoensgebonden invloed aanwezig is. Daarnaast zullen ook al deze methodes of gespecialiseerdere varianten van deze methodes toegepast worden op datasets met en zonder seizoenseffect waarbij meerdere invoerparameters gebruikt zullen worden.
Om deze methodes te scoren zullen de laatste waarden weggelaten en voorspeld worden waardoor uit de foutmarge tussen de voorspellingen en de werkelijke waarden afgeleid zal kunnen worden welke methode de meest accurate voorspelling zal kunnen maken. Om deze methodes te quoteren zullen de r^2 en de RMPSE (Root Mean Square Percentage Error) scoringsmethodes benut worden.

4. Verwachte resultaten

Er valt te verwachten dat polynomiale regressie het zwakste resultaat zal behalen aangezien polynomiale technieken, door de aard van een veelterm, doorgaans minder goed zijn voor extrapolatie waarvoor ze in deze context benut zullen worden. Ik verwacht dat LSTM best zal scoren aangezien dit type model specifiek voor tijdreeksen is opgesteld gevolgd door ARIMA.

5. Verwachte conclusies

Er valt te verwachten dat de voorgestelde technieken goede resultaten zullen behalen. Vooral voor LSTM liggen mijn verwachtingen vrij hoog omdat ik reeds een artikel (Siami-Namini, Tavakoli & Siami Namin, 2018) heb gelezen waarbij de voorspellingen voor LSTM accurater zijn. Ik heb een pak minder vertrouwen in polynomiale regressie aangezien deze techniek minder goed presteert bij extrapolatie en maar beter tot zijn recht komt bij interpolatie. ARIMA zal waarschijnlijk ook goede resultaten behalen.

Referenties

Brownlee, J. (2018). 11 Classical Time Series Forecasting Methods in Python (Cheat Sheet). Verkregen 29 juni 2020, van https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/

Siami-Namini, S., Tavakoli, N. & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1394–1401).