Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Computerwetenschappen

Toepassen van Nederlandse Gevoelsanalyse via sociale media

Yannick Merckx

Voorbereiding op de bachelorproef

Rolnummer: 500294

Promotor: Yann-MichaelÌĹl De Hauwere
Begeleiders: Maarten Deville
Peter Vrancx

Februari 2015

Inhoud

Chapter 1

Introductie

Vandaag de dag is sociale media een alledaags gegeven. Het is niet alleen onderdeel van het dagelijkse leven, maar ook een enorme bron aan informatie. Door deze bron van data te analyseren kan men andere informatie afleiden. Zo kan men gevoelens afleiden uit de data. Met behulp van computers en meerbepaald Machine Learning kan men de data verzamelen, verwerken en analyseren zodanig dat men bepaalde informatie kan afleiden uit de data. Hoe het verzamelen,verwerken en analyseren juist in elkaar zit, wordt beschreven in deze voorbereiding.

Chapter 2

Machine Learning

Machine learning is een welgekend begrip in de informatica wereld, maar wat het juist omvat, zijn toepassingen en hoe het helpt om de juiste verbanden te achterhalen uit enorme datasets wordt uitgelegd in dit hoofdstuk.

2.1 Wat is Machine Learning

Over Machine Learnig vindt men nergens een eenduidige definitie. Vele hebben hebben geprobeerd om een eenduidige definitie te definieren. Arthur Samuel(1959) definieerde machine learning als "'Field of study that gives computers the ability to learn without being explicitly programmed". Later stelde Tom Michel(1999) een well-posed learning problem als "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Als men machine learnig wil omvatten, kan men het best omschrijven als een onderzoeksdomein dat zich bezighoudt met het onderzoeken en de ontwikkeling van zelflerende algorithmes.
Binnen machine learning kan men verschillende groepen van lerende algorithmes onderscheiden. Zo heeft men Supervised learning, Unsupervised learning, Reinforcement learning en Recommender systems. In deze voorbereiding gaat men zich enkel opleggen op supervised en unsupervised learning. Deze soorten algorithmen gaan de nodige antwoorden bezorgen om later gevoelsanalyses te kunnen uitvoeren.

2.2 Supervised Learning

Supervised learning is een techniek, waarbij men het algorithme traint met data waarvan men de antwoorden al weet. Algemeen noemt men een dataset waarmee men een algorithme traint een trainingsset. Nadat dit algorithme zijn training heeft ondergaan, kan het zelfstandig keuzes maken aan de hand van de vergaarde kennis.

Bij supervised learning zijn er twee soorten problemen die kunnen optreden: een regressie probleem of een classificatie probleem.

2.2.1 Regressie Probleem

Het doel dat men wilt bereiken met supervised learning is dat het algorithme na een training antwoorden kan bezorgen. Bij het voorspellen van die antwoorden kan men te maken hebben met een regressie probleem. Dit probleem valt het best uit te leggen aan de hand van een voorbeeld.

Neem nu dat men de prijs van een huis wilt voorspellen. Het algorithme traint zich met een trainigset en bekomt volgend resultaat als men zijn bevindingen zou plotten.

TEKENENING HIER EEN GRAFIEK MET DATAPUNTEN

Stel nu dat men aan het algorithme de prijs van een huis met 1225 vierkante meter vraagt. Deze waarde zat niet in de dataset en moet dus voorspeld worden. Maar welke trend moet men volgen om de waarden te voorspellen. Men kan zowel kiezen voor een rechte of een 2de orde polynoom. Beiden zijn een mogelijkheid, maar geven een verschillend antwoord. De situatie, waarbij men een continue waarde moet bepalen en geen echte discrete afbakening bestaat, noemt men een regressie probleem.

Om dit probleem op te lossen, kan men van de techniek “Lineaire regeressie” gebruik maken.

Lineaire regressie

Lineaire regressie is een techniek waarbij het algorithme een hypothese probeert te vormen. De hypothese is een functie die opgesteld is aan de hand van de trainigsset en de gekende en ongekende outputwaarden zo goed mogelijk benaderd.

Als we terug kijken naar het voorbeeld van het huis. Kan het algorithme volgende hypothese opstellen.

FORMULE VAN HYPOTHESE / IS EEN RECHTE MET 1 VARIABLE
H(x) = THETA1 X + THETA 0

Gegeven hypothese is een lineaire functie met als parameters Theta 0 de nulconditie en Theta 1 de richtingscoefficient. Een hypothese met 1 functie noemt men ook wel een ÃĬÃĬndimensionale lineaire regressie.

Het opstellen van de hypothese introduceert op zijn beurt een “Minimalisatie probleem”. Men moet de hypothese zo goed mogelijk opstellen, zodat de afwijking ten op zichte van de gekende resultaten minimaal is. Als de hypothese minimaal is, kan men er van uit gaan dat de afwijking op ongekende resultaten ook minimaal is.

Het minimalisatie probleem kan opgelost worden met een kost functie en graduele afdaling.

Kost Functie en Gradule afdaling

Een kost functie is een functie die voor een bepaalde waarden van de parameters de gemiddelde afwijking van de hypothese ten opzichten van de resultaten gaat berekenen. Volgende formule kan men opstellen voor de kost functie:

FORMULE DE KOST FUNCTIE

Deze kost functie noemt men ook wel de “squared error cost function” en wordt over het algemeen het meest gebruikt. Merk op dat men niet zomaar telkens de som van het verschil tussen het resultaat van de hypothese neemt en de eigelijke waarden. Het kwadraat van het verschil wordt genomen vanwege de negatieve verschillen die ook moeten worden opgenomen als afwijking. Verder vereenvoudigt men het rekenwerk door te delen door twee (De helft van de kleinste waarde, blijft de kleinste waarde).

Zoals eerder gezegd is het de bedoeling om de afwijking zo klein mogelijk te houden. Om het minimum van de kost functie te vinden, kan men de techniek“gradiuele afdeling” gebruiken. Omwille van verschillende redenen is Graduele afdaling een van de meest gebruikte techniek binnen machine learning voor minimalisatie. Zo werkt de techniek voor een algemeen kost functie met n parameters J(theta0,theta2, theta 3, ... ,theta n) en kan het altijd uitgevoert worden aangezien de lineaire regressie kost functie altijd convex is.
De techniek start met een random start punt te nemen. Vervolgens gaat men stapsgewijs proberen te dalen tot je convergeert naar een lokaal minimum.

De preciese werking van het algorithme valt het best uit te leggen aan de hand van een voorbeeld. We nemen als voorbeeld onze eerder opgestelde hypothese met twee parameters theta 0 en theta 1. Als men de kost functie J(theta 0, theta 1) berekent en deze weergeeft in een driedimensionale weergave, krijgt men onderstaande plot.

AFBEELDING VAN EEN 3D-PLOT beschrijving: van axissen
Het stapgewijs dalen verloopt dalen verloopt met volgende formule:

FORMULE STAPGEWIJS DALEN

Alfa noemt men hier de learning rate. Dit is de grote van de stappen die men neemt bij het afdalen. De learning rate is een belangrijk element in het gradiuele afdalingsalgorithme. Als men deze te groot neemt kan men locale minima overslagen en convergeert het algorithme niet. Als men alfa te klein neemt kan het algorithme heel lang duren.
Een belangrijk en subtiel detail bij de formule en het algorithme is het simultaan updaten van de twee parameters. Als men dit niet doet, spreekt men niet van gradiuele afdaling.

Een bedenking die men moet maken bij gradiuele afdaling is het bestaan van meerdere lokale minima. Dit kan men echter eenvoudig oplossen door meerdere keren het algorithme uit te voeren met een ander startpunt.

Het gegeven voorbeeld noemt men specifieker "Batch graduele afdaling"

waarbij men telkens bij iedere stap de hele trainingsset vergelijkt. Er bestaan ook niet batch versies van graduele afdaling.

2.2.2 Classificatie Probleem

Een Classificatie probleem is een ander soort van probleem dat zich voordoet bij supervised training. Een classificatie probleem doet zich voor wanneer men data moeten verdelen over verschillende discrete klassen. Ieder elemenent mag maar tot 1 klasse behoren. De classificatie kan gebaseerd zijn op ÃĺÃľn attribuut, maar ook meerdere.

In het algemeen wordt het classificatie probleem het meest opgelost door "Logistieke Regressie "

Logistieke Regressie

2.3 Unsupervised Learning

Unsupervised learning is een techniek waarbij het algorithme zelfstandig moet leren hoe het juist moet en deze kennis gebruikt om later patronen en structuren in data te herkennen. De trainingsset bevat niet de antwoorden.

Het herkennen van structuren en patronen is niet voldoende, men moet de data concreet kunnen identiceren. Dit kan men doen door gebruik te maken van cluster algorithmes. Concreet gaat een cluster algorithme de data groeperen of "clusteren" in groepen.

Chapter 3

Text Mining

3.1 Document Preprocessing

3.2 Methoden

3.2.1 Vector Space Methode

Term weigthing

Latent Semantic Models

3.2.2 Probablistic methode

3.3 LSA Experiment

Chapter 4

Conclusie

Men kan besluiten dat Machine learning de oplossing biedt. Het bevat alle middelen om de gevoelsanalyse op sociale media toe te passen. Om succevol een gevoelsanalyse toe te passen moet men volgende stappen ondernemen. Ten eerst moet men de data verzamelen. Aangezien men te maken heeft met een dataset waarvan men geen informatie heeft, moet men technieken gebruiken van unsupervised learning. Vervolgens moet met de data preprocessen met het LSA algorithme. Het expirment heeft uitgewezen dat dit een zeer krachtig algorithme is, dat men meer inzicht geeft in de data. Ten slotte moet men de data verwerken met een cluster algorithme. De sequentie van verzamelen, verwerken en analyseren is een zeer belangrijk gegeven bij onze gevoelsanalyse.
CONCLUSIE KAN NOG BETER EN MISSCHIEN UITGEBREIDER, MAAR DIT ZIJN DE OUTLINES VAN DE CONCLUSIE