Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Computerwetenschappen

Technieken binnen de Machine Learning voor Text Mining

Yannick Merckx

Voorbereiding op de bachelorproef

Rolnummer: 500294

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

Februari 2015

Inhoud

1 Introductie **2**

2 Machine Learning **3**
2.1 Wat is Machine Learning 3
2.2 Supervised Learning 3
2.2.1 Regressie Probleem 4
2.2.2 Classificatie Probleem 7
2.3 Unsupervised Learning 7

3 Text Mining **8**
3.1 Document Pre-processing 8
3.2 Methoden 10
3.2.1 Vector Space Methode 10
3.2.2 Probablistic methode 10
3.3 LSA Experiment 10

4 Conclusie **11**

Chapter 1

Introductie

In deze voorbereiding gaat men technieken binnen de machine learning bespreken die men kan gebruiken voor text mining. Eerst gaat men een introductie gegeven over wat machine learning juist inhoudt, welke algemene technieken er worden gebruikt en waar men rekening mee moet houden bij deze technieken. Vervolgens gaat men specifiekere technieken bespreken, met de focus op text mining. Als laatste gaat men kijken hoe men deze technieken kan koppelen aan de eigelijke bachelorproef namelijk gevoelsanalyse op sociale media.

Chapter 2

Machine Learning

Machine learning is een welgekend begrip in de informatica wereld, maar wat het juist omvat, welke algmene technieken er bestaan en met welke factoren men moet rekening houden, wordt besproken in dit hoofdstuk.

2.1 Wat is Machine Learning

Over Machine Learnig vindt men nergens een eenduidige definitie. Vele 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 algoritmes.Hoofdzakelijk bestaat machine learning uit drie stappen namelijk data verzamelen, verwerken en analyseren.
Binnen machine learning kan men verschillende groepen van lerende algoritmes 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 algoritmen omvatten specifiekere technieken die zich lenen tot het gebruik bij text mining.

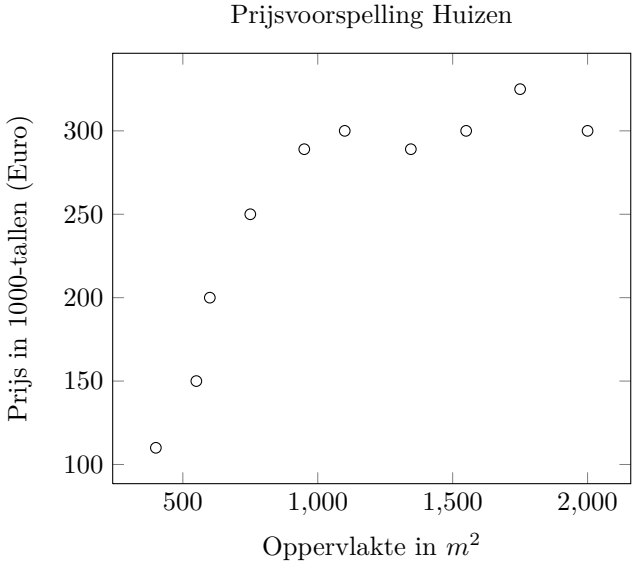
2.2 Supervised Learning

Wanneer men een algoritme wil trainen, heeft men informatie nodig om het algoritme te trainen. Dergelijke informatie noemt men de trainingset.
Laat men als voorbeeld een trainingsset met positieve en negatieve artikels nemen. Men weet welke artikels positief en negatief zijn en ieder artikel bevat deze kennis aan de hand van een label. Het algoritme kan de informatie van de

labels gebruiken om een zeker kennis te vergaren over artikels in het algemeen. Na het doorlopen van de trainingset, kan het algoritme de vergarde kennis gebruiken om een ongelabelde artikel te situeren als een positief of negatief artikel.
Deze techniek waarbij men een algoritme traint met een data waarvan men de antwoorden al weet noemt men supervised learning.
Het zelfstandig beslissingen maken over ongekende data is niet altijd even gemakkelijk en kan voor problemen zorgen. Bij supervised learning zijn er twee soorten problemen die kunnen optreden: een regressie probleem of een classification problem.

2.2.1 Regressie Probleem

Het doel dat men wil bereiken met supervised learning is dat het algoritme na een training antwoorden kan bezorgen over ongekende data. 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 algoritme traint zich met een trainigset en bekomt volgend resultaat als men zijn bevindingen zou plotten.



Stel nu dat men aan het algoritme 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 regressie***

gebruik maken.

Lineaire regressie

Lineaire regressie is een techniek waarbij het algoritme 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 algoritme volgende hypothese opstellen.

$$H_\theta(x) = \theta_0 + \theta_1 x$$

Gegeven hypothese is een lineaire functie met als parameters θ_0 de nulconditie en θ_1 de richtingscoefficient. Een hypothese met één functie noemt men ook wel een ééndimensionale 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 Graduele afdaling

Men herneemt het voorbeeld van de prijsvoorspelling van huizen. Men moest een zo precies mogelijke prijs voorspellen voor een oppervlakte van 1225 m^2. Om dit probleem op te lossen gaat het algoritme gaat voor zowel rechten als 2de orde polynomen de kostfuncties berekenen. Dit gebeurd door de prijzen van de gekende oppervlaktes te vergelijken met de prijzen van de hypothese. Door telkens het verschil in prijs voor een bepaalde oppervlakte te nemen, deze op te tellen en het gemiddelde te nemen, verkrijgt men de gemiddelde afwijking van de prijs van de hypothese ten op zichte van de echte prijs. Hierdoor krijgt men een beeld over hoe de prijzen van de hypothese zich verhouden tegenover de eigelijke prijzen. De kost functie voor dit voorbeeld is de functie met als functiewaarden de gemiddelde afwijkingen voor telkense een andere hypothese. Om dan een zo precies mogelijke voorspelling te kunnen doen voor de oppervlakte van 1225 m^2 moet men er voor zorgen dat men een hypothese kiest waarbij de gemiddelde afwijking zo laag mogelijk is.
Algemeen kan men de kost functie definieren als 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:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (H_\theta(x_i) - Y_i)^2$$

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 ***graduele 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 $J(\theta_0, \theta_1, \theta_2, \theta_3, ..., \theta_n)$ 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 algoritme valt het best uit te leggen aan de hand van een voorbeeld. We nemen als voorbeeld onze eerder opgestelde hypothese met twee parameters θ_0 en θ_1. Als men de kost functie $J(\theta_0, \theta_1)$ berekent en deze weergeeft in een driedimensionale weergave, krijgt men onderstaande plot.

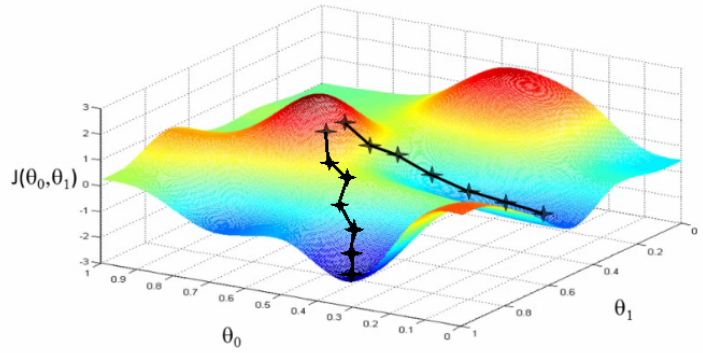


Figure 2.1: Driedimensionale weergave van de kostfunctie en zijn parameters

Het stapgewijs dalen kan als volgende formeel neergeschreven worden:

$$\theta_j := \theta_j - \alpha \frac{d}{d\theta_j} J(\theta_0, \theta_1) \text{ (voor j = 0 en j = 1)}$$

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 graduele afdalingsalgoritme. Als men deze te groot neemt kan men locale minima overslagen en convergeert het algoritme niet. Als men alfa te klein neemt kan het algoritme heel lang duren.
Een belangrijk en subtiel detail bij de formule en het algoritme is het simultaan

updaten van de twee parameters (zowel θ_0 als θ_1). Als men dit niet doet, spreekt men niet van graduele afdaling.
Een bedenking die men moet maken bij graduele afdaling is het bestaan van meerdere lokale minima. Dit kan men echter eenvoudig oplossen door meerdere keren het algoritme 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 en reder elemenent maar tot één klasse mag behoren. De classificatie kan gebaseerd zijn op één attribuut, maar ook op meerdere.
In het algemeen wordt het classificatie probleem het meest opgelost door ***logistische regressie***

2.3 Unsupervised Learning

Unsupervised learning is een techniek waarbij het algoritme 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.
Als men het zou vergelijken met het voorbeeld van supervised learning, zou men bij unsupervised learning als voorbeeld de situatie kunnen nemen waarbij het algoritme een ongelabelde trainingsset van artikels krijgt en na het verwerken van deze artikels zelfstandig keuzes kan maken over welke artikels positief en negatief zijn.
Echter kan een algoritme de structuren en patronen herkennen, maar kan het niet de data concreet identificeren. Dit probleem kan men oplossen door gebruik te maken van cluster algoritmes. Concreet gaat een cluster algoritme de data groeperen of ***clusteren*** in groepen en zo de data concreet identificeren.

Chapter 3

Text Mining

Nu men een algemeen begrip heeft van wat machine learning juist is en welke algemene technieken het omvat, kan men overgaan naar text mining en zijn geschikte technieken. In dit hoofdstuk gaat men bespreken welke technieken men kan gebruken voor text mining en wat deze juist inhouden. Als laatste gaat men de theorie toepassen op een voorbeeld en gaat men de resultaten van dit experiment bespreken.

Text mining of text data mining is een techniek waarbij men aan tekstanalyse doet om zo trends en patronen te kunnn vaststelllen. Neem opnieuw als voorbeeld onze artikels. Met text mining wil men de artikels zodanig analyseren zodanig dat men kan uitmaken welk artikel positief en welk negatief is. Een probleem dat zich onmiddelijk bij text mining voordoet is het ontbreken van een één-op-én relatie van woorden en een concept. Woorden verwijzen zelfden eenduidig naar ÃľÃľn concept. Zo het voorkomen van het woord "bank" in een tekst zowel verwijzen naar de finaciele instelling als naar een doodgewone zitbank in het park. Dergelijke dubbele betekenis van woorden maakt het moeilijk om de woorden, met als gevolg ook de tekst, te mappen op een bepaald concept.

Verder heeft men ook woorden in een tekst die weinig bijdragen tot de bepaling van het concept van de tekst bijvoorbeeld: ik,en,want... Deze woorden kan men uit de tekst filteren door een database aan te leggen met woorden die moeten men moet negeren. Deze techniek en nog soortgelijke alternatieven vereisen dat er al een voorverwerking plaatsvindt voordat men de dataset echt gaat analyseren op patronenen trends. Algemeen kan men zeggen als men de resultaten van de text mining wil optimaliseren, men aan ***document pre-processing*** moet doen.

3.1 Document Pre-processing

Document pre-processing is een optionele, maar zeker nuttig stap in het text mining proces. Document pre-processing bestaat eruit om je dataset al eens te verwerken, zodanig je extra informatie hebt, die je kan gebruiken bij de eigelijke analyse van de dataset. Zo kan je bijvoorbeeld alle stopwoorden verwijderen uit

de dataset. Wanneer men dan op deze gewijzigde dataset een analyze uitvoert, geeft men indirect de informatie mee dat stopwoorden er niet toe doen. Uiteraard is het verwijderen van stopwoorden één van de technieken. Er bestaan nog andere technieken die nuttig zijn als voorverwerking van een dataset. Zo kan men tekst en stucturen afleiden. Bijvoorbeeld het omzetten van Microsoft Word of Latex documenten naar XML maakt het parsen en analyseren van de documenten voor het algoritme veel gemakkelijker.Verder kan men ook ***stemming*** toepassen. Stemming is een techniek waarbij men tracht om de stam van het woord te achterhalen. Bijvoorbeeld uit het woord *katachtig* kan men het woord *kat* afleiden. De techniek kan gebaseerd zijn op een woordenboek bijvoorbeeld *Mmorph* is zo'n stemming woordenboek ontwikkeld door de Universiteit van Genève. Verder kan men de stemming ook baseren op een set van regels, bepaald door taalkundige. Het onderstaande voorbeeld illustreert een set van stemming regels voor het Frans:

$$
\begin{aligned}
&(m > 0)\ aux \rightarrow al \\
&(m > 0)\ ouse \rightarrow ou \\
&(m > 0)\ eille \rightarrow eil \\
&(m > 0)\ nne \rightarrow n \\
&(m > 0)\ fs \rightarrow v
\end{aligned}
$$

Figure 3.1: Voorbeeld van stemming regels in het Frans

Tenslotte is ***named entity recognition*** (NER) ook een techniek die men kan gebruiken bij document pre-processing. Hierbij gaat men entiteiten proberen detecteren in de tekst en deze labelen. Neem bijvoorbeeld de zin *Yannick heeft zich ingeschreven de richting Computerwetenschappen aan de Vrije Universiteit Brussel in 2012.* Men kan met NER de entiteiten eruit halen, labelen en volgend resultaat verkrijgen: $[Yannick]_{persoon}$ *heeft zich ingeschreven de richting Computerwetenschappen aan de* $[Vrije\ Universiteit\ Brussel]_{organisatie}$ *in* $[2012]_{tijdsaanduiding}$
Algemeen ziet men dat al deze technieken samen worden gecombineerd, wat alleen maar de uiteindelijke resultaten ten goede komt. Hoe deze gecombineerd kunnen wordt in het onderstaande voorbeeld geïllustreerd.

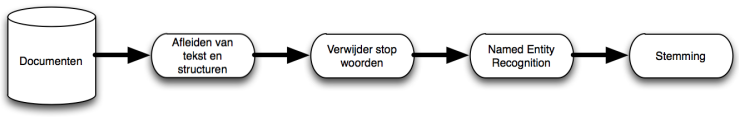


Figure 3.2: Combinatie van technieken bij document pre-processing

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