Mensen <5 skills tellen niet

min 1 functie en education

segmenten maken met NN MLP... neuronen die het sterkst reageren op data uitlezen

AE MLP(GNG) AE

Lijstje algoritmiek:

## Correlaties

### Omschrijving

Met dit algoritme worden correlaties berekend tussen verschillende attributen. Zo worden er correlaties berekend tussen Vaardigheid & Werkervaring, Vaardigheid & Educatie, en Vaardigheid & Industrie.

### Voorbeeld

Bijvoorbeeld, hoe vaker een Vaardigheid A voorkomt bij Werkervaring F1, hoe sterker de correlatie oftewel kans is dat deze vaardigheid voorkomt bij de werkervaring. Door de kansen van de correlaties tussen deze drie correlatievlakken te combineren, kan een sterkere voorspelling worden gedaan voor de vaardigheden die een persoon bezit. Uiteindelijk moet besloten worden welke correlatie sterker geldt dan andere. Bijvoorbeeld, vaardigheden die horen bij een industrie zijn vrij algemeen. Vaardigheden bij een educatie dienen als een fundering. Vaardigheden bij een werkervaring dienen als een verdere specialisatie.

### Voordelen

Aan de hand van de data waartussen veelvuldig correlaties aanwezig zijn kun je met zekerheid zeggen dat een vaardigheid daadwerkelijk aan iets is gerelateerd.

### Nadelen

De berekening (trainen) van alle correlaties is erg duur.

Simpele correlaties tussen Vaardigheid en 1 ander attribuut kan te weinig zeggen. Stel, persoon P1 zit in industrie I1 (vaardigheid D komt met name voor) en heeft educatie E1 en vaardigheden A, B en D. Persoon P2 zit in industrie I2 (vaardigheid E komt met name voor) en heeft educatie E1 en vaardigheden A, C en E. Persoon P3 zit in industrie I1 en heeft educatie E1, maar de vaardigheden zijn onbekend. Volgens de industrie correlatie heeft deze persoon vaardigheid D en volgens de educatie correlatie vaardigheid A. Vaardigheid B mist nu omdat de industrie en educatie niet zijn samengenomen. (Oplossing: neem Vaardigheid & Industrie en Werkervaring, Vaardigheid & Industrie en Educatie als correlaties?)

Er zullen 'gewichten' mee moeten spelen omdat een persoon meerdere educaties en werkervaringen kan hebben, maar dat industrie en vaardigheden als algemeen attribuut worden beschouwd. Bij persoon P1 die in industrie I1 zit en educaties E1 en E2 heeft gedaan en vaardigheden A en B heeft, kan niet met zekerheid worden gezegd welke vaardigheid tot welke educatie relateert.

## Decision trees

Een supervised learning methode. Het doel hiervan is om de vaardigheden te voorspellen door simpele tussenvragen te stellen om de informatie dusdanig op te splitsen dat het meest kernmerkend is voor de vaardigheid.

### Voorbeeld

In principe kan hetzelfde hier worden gedaan als in het vorige algoritme. Zowel de industrie als de werkervaringen of educaties data worden in het algoritme ingevoerd waaruit predicties van vaardigheden zullen komen.

### Voordelen

Eenvoudig te begrijpen (en te visualiseren).

Logaritmische kost in het aantal datapunten die gebruikt is voor het trainen.

Ondersteunt multi-labeling

Het berekenen van correlaties wordt door het algoritme gedaan.

Snelle performance bij training en predictie.

### Nadelen

Overfitting kan snel gebeuren. Parameters als minimum number of samples bij een leaf node of maximum depth van de tree kunnen dit voorkomen indien correct ingevoerd.

Kleine variaties in data kunnen totaal andere tree structuren opleveren. Een ensemble kan dit oplossen. Het trainen van meerdere trees kan een beter optimaal resultaat opleveren.

Biased trees kunnen gecreëerd worden als sommige classes domineren.

## ML-KNN

Het doel van KNN is om de meest soortgelijke data te verdelen over een gegeven aantal clusters. Een cluster is het middelpunt van een collectie van soortgelijke data.

### Voorbeeld

Wederom kan dezelfde strategie worden gehanteerd als bij de voorgaande algoritmen. K wordt gelijk gezet met het aantal vaardigheden en elke sample bij het trainen heeft meerdere labels (vaardigheden). In principe moet het mogelijk zijn om zowel de industrie, educatie en werkervaring als features te nemen voor het trainen en predicten.

### Voordelen

Eenvoudig te begrijpen.

Lage kosten in het trainen. Predictie is over het algemeen ook snel, maar een groot aantal samples kan voor performanceverlies zorgen.

Het algoritme berekent zelf de afstanden tot clusters waartoe een sample hoort.

Ondersteunt multi-labeling.

### Nadelen

Een groot aantal samples kan voor performanceverlies zorgen.

Onrelevante data hebben grote effecten op de accuratie.

## Auto encoder

Het doel van een auto encoder is om features te herstructureren. Dit wordt gebruikt om data zonder of bijna zonder enige correlatie weg te laten. Dit is met name handig voor het reduceren van de dimensie van de data. De auto encoder wordt onder andere gebruikt voor het voortrainen van de data.

mlp ~ gng ~ som ~ lvq

micholov reinforcement learning

expectation maximization