Beslissingsbomen

1 The Simpsons

In deze oefening exploreren we de basistechnieken van beslissingsbomen aan de hand van de Simpsons.

1. lees het bestand “simpsons_categorical.csv” in. We willen de kolom ‘geslacht’ voorspellen.
2. bereken zelf de entropie van deze tabel. Schrijf ook de formule op
3. bereken de gain voor iedere kolom. Kan je ook de gain van de laatste kolom berekenen? Welke waarde vind je dan? Schrijf ook telkens de formule op
4. maak een nieuwe subtabel met enkel de rijen waarbij het gewicht altijd <=160 is.
5. wat is de entropie van deze subtabel?
6. bepaal de gain voor iedere kolom van deze subtabel. Welke kolom kies je dus voor de volgende stap?
7. maak de boomstructuur in Python met het gegeven ID3 algoritme
8. maak de boomstructuur ook eens met de Id3Estimator. Krijg je dezelfde structuur?
9. maak tenslotte de boomstructuur met het CART algoritme. Zet daartoe de data eerst om naar getallen met pd.get_dummies(). Krijg je hetzelfde resultaat als bij de voorgaande algoritmen?
10. hoe zou de volgende simpson geklasseerd worden?
    haarlengte: 8, gewicht: 290, leeftijd: 38

2 Bank

Een bank wil graag weten of een klant zijn lening zal kunnen betalen of niet. De bank beschikt over gegevens van klanten in het verleden en wil hieruit een beslissingsboom laten opstellen.

1. lees het bestand “bank-data.csv” in. Dit bevat allerlei gegevens van vorige klanten. De kolom "pep" geeft weer of de klant zijn lening kon afbetalen. Verwijder de kolom “id”
2. Maak nu een beslissingsboom met de Id3Estimator. Breek de boom af op een diepte van 3 en maak een plot.
3. Welke kolom heeft de hoogste information gain?
4. We willen nu ook het CART algoritme gebruiken om de boom op te stellen. Lees de data terug in en wis alle kolommen die geen getallen of booleans (yes/no) bevatten (id, region en sex). Vervang de waarden ‘yes’ door 1 en ‘no’ door 0. Gebruik deze data nu om met het CART algoritme een boomstructuur op te stellen (zet ook hier de maximum diepte op 3). Welke boom bekom je? Is die vergelijkbaar met de vorige?
5. In de nodes zitten er getallen tussen vierkante haakjes. Wat betekenen deze?
6. Welke 2 factoren zijn het belangrijkst om te weten of iemand zijn lening zal kunnen terug betalen?
7. Waar zou een klant met volgende gegevens uitkomen volgens de laatste boomstructuur (CART)?

age	sex	region	income	married	children	car	save_act	current_act	mortgage
44	FEMALE	TOWN	15735.8	YES	1	NO	YES	YES	YES

8. kan je iets zeggen over de betrouwbaarheid van vorige uitspraak?

3 Titanic

Op 15 april 15 1912 zonk de “onzinkbare” Titanic op haar eerste tocht. Er waren niet genoeg reddingsboten voor iedereen aan boord. Hierdoor stierven 1502 van de 2224 opvarenden.
In deze oefening proberen we te weten te komen welke soorten passagiers er meer kans hadden om te overleven. Hiervoor starten we van een bestand waarin de gegevens van 891 passagiers genoteerd staan. Daaruit gaan we de computer laten zoeken naar regels die bepalen of een passagier overleefde of niet. We kunnen het resulterende model dan gebruiken om te kijken welke van de andere passagiers overleefden.
Je vindt meer informatie op: https://www.kaggle.com/competitions/titanic

De volgende kolommen zijn beschikbaar:

- PassengerId: een volgnummer per passagier
- Survived: heeft deze passagier de ramp overleeft?
- Pclass: de klasse waarin de passagier vaarde
- Name: de naam van de passagier
- Sex: het geslacht van de passagier
- Age: de leeftijd van de passagier
- SibSp: het aantal broers of zussen die ook aan boord waren
- Parch: het aantal ouders of kinderen die ook aan boord waren
- Ticket: het nummer van het ticket
- Fare: het bedrag dat de passagier betaalde
- Cabin: het nummer van de kabine waarin de passagier verbleef
- Embarked: de plaats waar de passagier inscheepte (C = Cherbourg, Q = Queenstown, S = Southampton)

1. Lees de data in en bestudeer deze. Welke kolom bevat het meeste ontbrekende waarden?
2. De kolommen “PassengerId”, “Name”, “Ticket” en “Cabin” kunnen we moeilijk gebruiken in een beslissingsboom (kan je zien waaom?). Verwijder ze. De kolom “Survived” zet je best om naar strings (titanic.Survived = titanic.Survived.apply(str))
3. Haal nu alle lijnen met ontbrekende waarden uit de tabel. Hoeveel lijnen blijven er over?
4. Maak een beslissingsboom, gebruik makende van de Id3Estimator met een maximum diepte van 3. Maak een plot. Welke kolom heeft de hoogste information gain?
5. Zou een man van 20 jaar die in klasse 2 reisde overleefd hebben volgens deze beslissingsboom? Hoe zeker ben je van die uitspraak?
6. Wat zouden de kansen geweest zijn als deze man in klasse 1 had gereisd?
7. We willen ook een beslissingsboom maken met het CART algoritme. Daarvoor moeten alle kolommen getallen bevatten. We kunnen dit eenvoudig doen door het geslacht te vervangen door 0 en 1. Voor de kolom “Embarked” gebruik je one-hot encoding.
8. Maak nu een beslissingsboom met CART met een maximum diepte van 3 en een random_state van 42 en plot deze. Is deze vergelijkbaar met de vorige?
9. We gaan nu proberen te voorspellen welke van de andere passagiers overleefden. Lees daarvoor het bestand “titanic_test.csv” in. Haal ook hier de kolommen “PassengerId”, “Name”, “Ticket” en “Cabin” weg en verwijder daarna alle rijen met ontbrekende waarden.
10. Voorspel nu of de test-passagiers overleefd zouden hebben of niet. Gebruik beide algoritmes en vergelijk de resultaten. Zijn er verschillen?
11. Hoeveel procent van de test-passagiers zouden volgens deze beslissingsbomen overleefd hebben?

4 Music Genre Classification

We willen een applicatie maken die een muziekfragment kan klassificeren. De bedoeling is dat er een geluidsfragment van 30 seconden wordt herkend. Dit geluidsfragment wordt eerst helemaal ontleed door verschillende algoritmen. Zo komen we verschillende eigenschappen te weten over dat geluidsfragment. De vraag is nu of we uit die eigenschappen kunnen afleiden welk genre er in dat fragment gespeeld wordt (blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock). Voor meer informatie, zie: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

1. Lees de data vanuit het bestand “music.csv” en bestudeer het.
2. Verwijder de kolommen “filename” en “length”.
3. We willen de dataset in 2 delen splitsen: een training set en een test set. We kunnen de test set dan gebruiken om te zien hoe goed ons algoritme werkt. Zoek eens op internet op hoe je dat kan doen (hint: train_test_split). Splits de data nu op in 80% trainingsdata en 20% testdata (gebruik een random_state van 42).
4. Maak een beslissingsboom met het CART algoritme van de trainingsdata met een random_state van 42 en een maximum diepte van 8.
5. Laat de beslissingsboom nu voorspellingen doen voor de testdata. Hoeveel procent van de voorspelde waarden komen overeen en hoeveel niet? Is dit een goede voorspeller?