Tentamen Data Mining

Datum: 9 januari 2012
Tijd: 14:00 - 17:00

Algemene Opmerkingen

- Dit is geen open boek tentamen, noch mogen er aantekeningen gebruikt worden.
- Laat bij het uitvoeren van berekeningen zien hoe je aan een antwoord gekomen bent. Als je alleen een antwoord opschrijft en dat is fout, rest ons niets dan het geheel fout te rekenen.
- Een rekenmachine is toegestaan.
- De cijfers van de nagekeken tentamens zullen binnen 4 weken op de deur van de kamer 110 gepubliceerd worden.

Opgave 1. Korte vragen (16 punten)

Geef korte, ter zake doende antwoorden op de volgende vragen:

(a) Is het Nearest-Neighbour algoritme alleen toepasbaar op continue data?

(b) Leg uit waar cross-validation voor te gebruiken is.

(c) Een onderneming heeft met behulp van een clustering-algoritme zijn klanten in een aantal groepen ingedeeld. De variabelen waarop geclusterd is zijn o.a. leeftijd, inkomen, en huwelijkse staat. Direct na de conversie van de guldens naar euros werd het algoritme opnieuw gedraaid en vond men andere groepen dan voorheen. Wat kan hiervan de oorzaak zijn? Hoe had men dit kunnen voorkomen?

(d) Wat is het Apriori principe dat gebruikt wordt bij het berekenen van frequent item-sets?

(e) Waarom is Weka zo'n belangrijk tool voor de data mining commmunity.

(f) Een *parity function* is een Boolean functie waarvan de waarde 1 is als de input-vector een oneven aantal 1-en heeft (bijv: $(1, 0, 1, 1) \rightarrow 1$). Deze functie kan bijvoorbeeld gebruikt worden als checksum bij het verzenden van data: als de berekende parity van een verzonden byte niet overeenkomt met zijn checksum, dan is er iets mis gegaan in het verzenden. Leg uit waarom de parity function een probleem is voor data mining algoritmen, en geef een eenvoudig voorbeeld.

(g) Regressie en Associatie zijn twee data mining paradigma's. Noem er nog twee.

Opgave 2. Data Representatie (10 punten)

Bij het reviewen van wetenschappelijke artikelen wordt vaak de volgende indeling gebruikt om de kwaliteit aan te geven: '*strong accept*', '*weak accept*', '*weak reject*' en '*strong reject*'. Stel we hebben een dataset opgesteld naar aanleiding van reviews van alle papers voor een te houden conferentie, waarbij kwaliteit één van de attributen is.

(a) Hoe noem je het type van een attribuut ~~die~~ dat deze kwaliteit aanduidt?

(b) Geef een voorbeeld van een alternatieve representatie voor de kwaliteit. Noem ook twee voorbeelden van een algoritme dat baat zou hebben bij de alternatieve representatie (waarbij kwaliteit niet de target is).

(c) Stel we willen de kwaliteit wel als target gebruiken, en willen binaire classificatie gebruiken om de kwaliteit te 'voorspellen'. Op welke drie manieren kunnen we de representatie van kwaliteit aanpassen zodat dit mogelijk wordt? Wat is het nadeel van deze manieren van representatie?

Opgave 3. Clustering (13 punten)

Hieronder staan gegevens van 4 (fictieve) studenten van de data mining cursus. We noteerden respectievelijk het aantal bijgewoonde lessen, hun score (op 10), het aantal dagen examenvoorbereiding, en of ze al dan niet voor het examen kwamen opdagen:

$$s_1 : (0, 0, 0, 1)$$

$$s_2 : (7, 8, 2, 1)$$

$$s_3 : (4, 8, 2, 1)$$

$$s_4 : (3, 0, 0, 1)$$

We willen deze studenten clusteren, wat we kunnen doen door middel van de k-means methode. Kies $k = 2$, en als initiële random cluster centroids:

$$c_1 : (0, 6, 1, 0) \quad c_2 : (7, 2, 1, 0)$$

Pas de k-means methode toe op D tot een maximum van 3 iteraties. Noteer voor elke iteratie welke clusters gevormd worden en wat de cluster centroids zijn. Convergeert de methode? Zo ja, motiveer.

Gebruik de standaard, Euclidische afstandmaat voor het berekenen van afstanden:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Merk op dat om 2 afstanden te vergelijken, de vierkantswortel weggelaten kan worden. Je kan jezelf ook wat rekenwerk besparen door goed naar de coordinaten van de gegeven punten te kijken.

Opgave 4. Naïve Bayes (13 punten)

We willen Naïve Bayes toepassen op de bekende Iris dataset. Aangezien dit een dataset met numerieke attributen is, moeten we eerst de data discretizeren. Na discretizatie en selectie van 10 voorbeelden ziet de data er als volgt uit:

Sepal length	Sepal width	Petal length	Petal width	Species
long	narrow	long	narrow	setosa
short	wide	short	narrow	versicolor
long	narrow	long	narrow	setosa
long	narrow	long	wide	versicolor
long	narrow	long	narrow	setosa
long	narrow	short	narrow	setosa
short	wide	long	narrow	versicolor
short	wide	short	narrow	setosa
long	narrow	long	narrow	versicolor
long	narrow	short	narrow	setosa

We willen irissen kunnen classificeren als zijnde een *Iris Setosa* of een *Iris Versicolor*.

(a) Bij deze dataset wordt echter een aanname van het Naïve Bayes algoritme geschaad. Wat is deze aanname en leg uit waarom deze geschaadt wordt.

(b) Laat zien hoe de Naïve Bayes methode deze dataset vertaalt tot een probabilistisch model dat gebruikt kan worden voor toekomstige classificaties.

(c) De eerste nieuwe iris die we tegenkomen heeft de volgende eigenschappen:

Sepal length	Sepal width	Petal length	Petal width	Species
long	narrow	short	wide	?

Wat is het meest waarschijnlijke classificatie voor de nieuwe iris? Toon je berekening.

(d) In dit geval treedt er een bekend probleem op ivm kansrekening. Wat is dit probleem, en leg uit wat de standaard manier is om dit probleem op te lossen. Gebruik deze methode om toch de nieuwe iris te classificeren.

Je krijgt de formule voor Bayes' rule: $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$.

Opgave 5. Frequent Pattern Mining (15 punten)

Een fruitsapproducent experimenteert met fruitsapcombinaties. Gegeven een transactionele database D met combinaties gesuggereerd door testpersonen:

tid	Items
1	{Sinaasappel, Appel, Kers}
2	{Sinaasappel, Ananas, Mango}
3	{Sinaasappel, Kers}
4	{Ananas, Mango}
5	{Appel, Kers}
6	{Sinaasappel, Kers, Ananas, Mango}
7	{Kers, Ananas, Mango}
8	{Appel, Kers}
9	{Sinaasappel, Kers, Ananas, Mango}
10	{Appel}

(a) Onze database D heeft 5 unieke items. Uit D kunnen we itemset-patronen en associatie-regels verkrijgen. Geef in dit geval voor zowel patronen als associatie-regels het theoretisch maximale aantal.

(b) Gegeven een minimal support $minsup = 0$, wat is het maximum aantal items in de frequente itemsets? Beantwoord deze vraag ook voor $minsup = 0.1$.

(c) Gegeven een minimal support $minsup = 0.3$, teken de itemset lattice en label elke knoop met minstens een van de volgende letters die van toepassing is: *I*= infrequent itemset, *F*=frequent itemset, *M*=maximal itemset, *C*=closed itemset.

(d) Gegeven de volgende associatieregels:

$$\{Appel\} \rightarrow \{Kers\}$$

$$\{Sinaasappel, Ananas\} \rightarrow \{Mango\}$$

$$\{Kers\} \rightarrow \{Ananas, Mango\}$$

Bereken de *support* en *confidence* voor deze regels.

Opgave 6. Subgroup Discovery (15 punten)

Ga uit van een dataset met meerdere attributen, waarvan 1 attribuut de target vormt. Dit attribuut heeft twee waarden, T en F. Stel dat we een subgroep S op de data geëvalueerd hebben, en het blijkt dat de subgroep een support heeft van 80%. Ook blijkt dat S positief geassocieerd is met de target.

(a) Teken de ROC ruimte voor dit binaire attribuut, en teken daarin de locatie van een subgroep die S zou kunnen voorstellen. Geef ook aan op welke lijn alle subgroepen met support van 80% liggen.

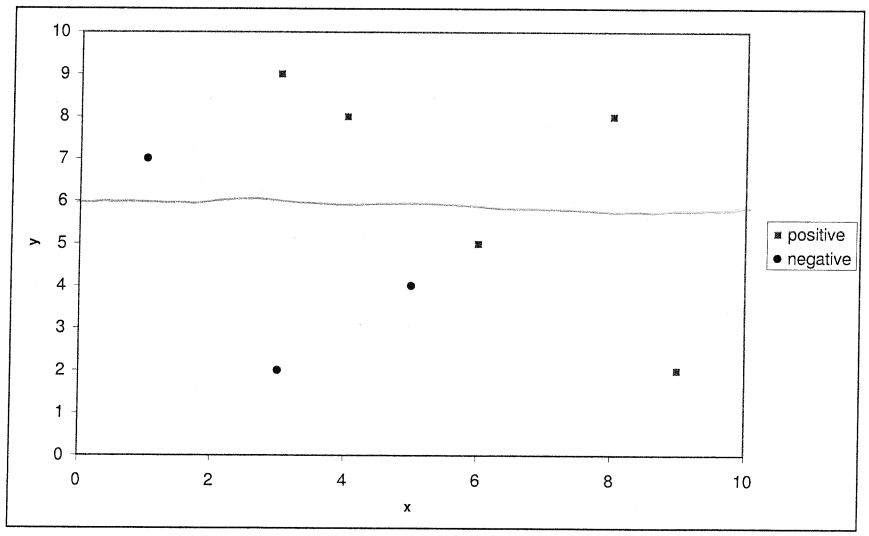
(b) Stel nu dat we S uitbreiden met een conditie c, zodat $S' = S \wedge c$. Geef in het schema duidelijk aan in welk gebied de subgroep S' moet liggen. Geef ook schematische aan hoe dit gebied bepaald kan worden. Als je een voorkeur hebt voor een tekstuele beschrijving, kan dit ook.

(c) Stel dat de we alleen geïntereseerd zijn in patronen met een minimum support van 10%, en S' voldoet aan deze eis. Geef aan in welk gedeelte S' niet kan liggen.

(d) Weighted Relative Accuracy ($WRAcc(S)$) is een bekende kwaliteitsmaat binnen SD. Stel dat we een nieuw maat introduceren die als volgt gedefinieerd is: $Q(S) = WRAcc(S)^2$. We evalueren een groot aantal subgroepen met deze nieuwe maat, en vergelijken de ranking van resultaten met die van $WRAcc$. Beschrijf de grootste verschillen, en de eventuele voor- en nadelen hiervan.

Opgave 7. Decision Trees (18 punten)

Ga uit van een dataset met twee numerieke attributen (x en y) en een binair target. De dataset bevat 8 voorbeelden, en ziet er als volgt uit. Alleen integer waarden komen voor:



(a) Pruning is een manier om een gebouwde beslisboom weer te snoeien, zodat overfitting voorkomen kan worden. Leg uit waarom in de meeste beslisboomalgoritmen eerst een te grote boom gebouwd wordt, terwijl er daarna toch vaak weer gesnoeid gaat worden.

(b) Geef een voorbeeld van een beslisboom van diepte 2 (dus maximaal twee splitsingen per pad van de wortel naar een blad) zoals die waarschijnlijk door een algoritme op basis van information gain geproduceerd wordt.

(c) Geef een suggestie hoe deze data met behulp van een beslisboom makkelijker en preciezer gemodeleerd kan worden.

(d) Leg uit waarom $x < 7$ nooit een optimal splitsing zal zijn.

(e) Bereken de information gain van de splitsing $y < 6$. Een lage precisie van je berekening volstaat. Gebruik eventueel de onderstaande tabel met benaderde waarden voor de entropie $H(p)$.

p	$H(p)$
0	0
1/8	0.54
2/8	0.81
3/8	0.95
4/8	1
5/8	0.95
6/8	0.81
7/8	0.54
1	0