Tentamen Data Mining

Datum: 4 januari 2016
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) Leg uit hoe 10-fold cross-validation werkt.

(b) Zowel k-means als SOM maken gebruik van clustercentra om een distributie te benaderen, en de algoritmen lijken dus op elkaar. Leg uit welke extra eigenschap SOMs hebben ten opzichte van k-means.

(c) Data Mining algoritmen zijn in te delen in twee categorieën: *black box* en *mining for insight*. Geef van de volgende algoritmen aan in welke categorie ze vallen, en onderbouw je antwoord:

- C4.5
- Regression trees (met constante waarden aan de bladeren)
- k-Nearest Neighbour
- Subgroup Discovery

(d) Geef twee voordelen van hiërarchische clustering ten opzichte van een meer standaard methode zoals k-means.

(e) Is het XOR probleem oplosbaar met één enkele lineaire classifier?

(f) Geef drie voorbeelden van regressie-algoritmen.

(g) Geef de meer gebruikelijke namen voor de volgende implementaties van algoritmen in Weka:

- IBK
- j48

Opgave 2. Frequent Pattern Mining (15 punten)

Gegeven een transactionele database met de volgende itemsets over $\{A, \ldots, E\}$:

tid	Items
1	$\{D\}$
2	$\{A, C\}$
3	$\{B, E\}$
4	$\{D, C\}$
5	$\{D, C\}$
6	$\{A, D, C\}$
7	$\{A, B, E\}$
8	$\{C, B, E\}$
9	$\{A, C, B, E\}$
10	$\{A, C, B, E\}$

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

(b) Leg uit hoe de termen *maximal itemset* en *closed itemset* gedefinieerd zijn.

(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 zijn: *I*= infrequent itemset, *F*=frequent itemset, *M*=maximal itemset, *C*=closed itemset.

Opgave 3. Maximally Informative k-Itemsets (15 punten)

Gegeven de volgende dataset van binaire attributen:

A	B	C	D
0	1	0	1
0	1	0	1
1	1	1	1
1	1	1	1
0	1	1	0
0	0	1	0
1	0	0	0
1	0	0	0

(a) Geef de entropie van elk van de 4 attributen (over de hele dataset). 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

(b) Bewijs dat $\{A, C\}$ een miki is voor $k = 2$ en geef de joint entropy. Is het de enige miki van 2 attributen?

(c) Geef een bovengrens van $H(\{A, C, D\})$ op basis van de hierboven berekende joint entropy van $\{A, C\}$. Bepaal daarna de exacte waarde van $H(\{A, C, D\})$.

(d) Geef een miki voor $k = 3$.

(e) Leg uit hoe een taart in acht even grote delen te snijden is, met slechts drie sneden.

Opgave 4. Data Representatie (11 punten)

Skipistes in de Alpen worden vaak ingedeeld in verschillende moeilijkheidsklassen, door middel van de volgende kleuren: '*groen*', '*blauw*', '*rood*', en '*zwart*', in volgorde van oplopende moeilijkheid. Stel, er wordt een dataset van alle skipistes in de Alpen samengesteld, waarin de moeilijkheid, de lengte, het hoogteverschil en verscheidene andere attributen worden opgeslagen.

(a) Hoe noem je het type van een attribuut dat deze moeilijkheid aanduidt?

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

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

(d) Geef een voorbeeld van een binair en van een nominaal attribuut zoals die in de dataset van pistes kan voorkomen. Welke waarden kunnen deze twee attributen aannemen?

Opgave 5. Decision Trees (15 punten)

(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 van de volgende statements aan of ze waar zijn

1. Information gain van een gegeven attribuut neemt altijd toe naarmate je dieper in de boom komt.
2. Information gain van een gegeven attribuut neemt altijd af naarmate je dieper in de boom komt.
3. Information gain van een gegeven attribuut kan toenemen naarmate je dieper in de boom komt.
4. Information gain van een gegeven attribuut is altijd ≥ 0.
5. Information gain van een gegeven attribuut is altijd ≤ 1.

Ga uit van een dataset met drie binaire attributen (A, B en C) en een binair target D. De dataset is identiek aan die in opgave 3.

(c) Geef een beslisboom van diepte 2 (dus maximaal twee splitsingen per pad van de wortel naar een blad) zoals die door C4.5 op basis van information gain geproduceerd zal worden.

(d) Geef aan hoe via pre-processing van de dataset een beter model gebouwd kan worden voor de target D.

(e) Stel dat we een nieuwe numeriek target E introduceren, dat functioneel afhankelijk is van $A, \ldots, D$, namelijk $E = A + B + C + D$. Leg uit hoe een model tree geïnduceerd op deze dataset er uit zal zien.

Opgave 6. Clustering (15 punten)

Zie de onderstaande database van vier voorbeelden, elk bestaande uit een vector van vier waarden:

$$s_1 : (2, 2, 2, 4)$$
$$s_2 : (16, 18, 6, 4)$$
$$s_3 : (10, 18, 6, 4)$$
$$s_4 : (8, 2, 2, 4)$$

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

$$c_1 : (2, 14, 4, 2) \quad c_2 : (16, 6, 4, 2)$$

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.

Opgave 7. Data Mining casus (13 punten)

Een fietsenmaker houdt data bij over fietsen die gerepareerd worden. Voor elke reparatie wordt genoteerd om welk soort fiets en eigenaar het ging, en welke onderdelen gebruikt werden bij de reparatie. In totaal zijn er 50 mogelijke onderdelen, allemaal met een nette code. Het doel van de fietsenmaker is om uit de historische data te leren, zodat voor een geplande reparatie voorspeld kan worden welke onderdelen nodig zijn. Met behulp van deze voorspelling is de fietsenmaker in staat om onderdelen van tevoren te bestellen, of om ze eventueel langdurig op voorraad te hebben. De fietsenmaker heeft een database opgebouwd van 1000 reparaties.

(a) Leg uit hoe de fietsenmaker Data Mining kan inzetten om deze voorspelling te doen, en om welk DM paradigma het hier gaat. Geef een voorbeeld van een specifiek algoritme dat hier gebruikt kan worden.

(b) Geef vier voorbeelden van attributen die de fietsenmaker zou moeten bijhouden in zijn database om een goede voorspelling mogelijk te maken. Geef ten minste voor elk van de volgende attribuut-typen een voorbeeld: numeriek, nominaal, binair.

(c) De fietsenmaker heeft slechte ervaringen met DM, omdat de lokale DM-leverancier ook het telefoonnummer van de klant had meegenomen als attribuut. De voorspellingen bleken niet betrouwbaar. Geef aan wat er mis is gegaan.

(d) De datum van de reparatie is één van de attributen die mogelijk een invloed hebben. Leg uit hoe dit attribuut via preprocessing het best gebruikt kan worden, en geef enkele voorbeelden.