Tentamen Data Mining

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 XOR probleem oplosbaar met één enkele lineaire classifier?

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

(c) Is de volgende stelling waar? 'Alleen in het geval van ongebalanceerde data is het nuttig om cross-validation te verrichten'.

(d) 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?

(e) Wat is het A-priori principe dat gebruikt word bij het berekenen van frequent itemsets?

(f) Van de Hollandse Brug worden met behulp van sensoren over een bepaalde tijdspanne eigenschappen gemeten. Dit levert 100 signalen op met elk 10^{12} (genormaliseerde) meetpunten: $s_i = \{x_1, x_2, \ldots, x_{10^{12}}\}, i \in [1, 100]$.
Noem een nadeel op van het gebruik van een euclidische afstandsmaat tussen signalen (en dus sensoren) in deze context zoals gebruikelijk bij methoden zoals nearest neighbour. Noem daarnaast een aanpak die gekozen kan worden indien men deze signalen in groepen wil indelen, zodanig dat de signalen die het meest op elkaar lijken in dezelfde groep zitten.

(g) Om de juiste waarde van k te bepalen, het juiste aantal clusters zogezegd, doet een onderzoeker voor verschillende waarde van k een herhaling van het cluster algoritme. Om deze te vergelijken gebruikt hij een error maat:

$$E = \sum_{i=1}^{n} d^2(x_i, c(x_i)),$$

waarbij $c(x)$ de dichtsbijzijnde centroid voor datapunt x is, en d een afstandsmaat. Wat vindt u van deze werkwijze?

(h) Data Mining algoritmen kunnen in twee groepen opgedeeld worden, afhankelijk van hoe de modellen gebruikt gaan worden. De eerste groep bestaat uit black-box methoden. Hoe noemen we de andere groep?

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 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 (15 punten)

Gegeven is dataset D met $n = 8$ datapunten, die elk 2-dimensionaal zijn $\binom{x}{y}$:

$$D = \{\binom{0}{1}, \binom{1}{0}, \binom{0}{2}, \binom{1}{3}, \binom{4}{1}, \binom{5}{2}, \binom{4}{4}, \binom{5}{3}\}.$$

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

$$c_0 = \binom{3}{1}, c_1 = \binom{3}{4}.$$

Pas de k-means methode toe op D, en laat duidelijk zien welke cluster centroids resulteren na convergentie.

Opgave 4. Naïve Bayes (10 punten)

Omdat het maken van een tentamen data mining een waar feest is, hebben we onze kamer versierd met veel mooie ballonnen. Tijdens al dat ballonnen blazen hielden we voor de volgende attributen iedere waargenomen waarde bij: *kleur*, *grootte*, *vorm*, *functie*, *vol*.

KLEUR	GROOTTE	VORM	OPDRUK	VOL
geel	klein	bol	ja	waar
blauw	groot	plat	ja	onwaar
geel	klein	bol	ja	waar
geel	klein	bol	nee	onwaar
geel	klein	bol	ja	waar
geel	klein	plat	ja	waar
blauw	groot	bol	ja	onwaar
blauw	groot	plat	ja	waar
geel	klein	bol	ja	onwaar
geel	klein	plat	ja	waar

We willen kijken wat de eigenschappen zijn van een volle ballon t.o.v. een niet volle ballon.

(a) Voor deze classicatie taak gebruiken we Naïve Bayes. Bij deze database wordt echter een aanname geschaad. Wat is deze aanname en leg uit hoe deze geschaadt wordt.

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

(c) Later vindt iemand nog een ballon, de gemeten waarden zijn:

KLEUR	GROOTTE	VORM	OPDRUK	VOL
geel	klein	plat	nee	?

Als we willen voorspellen of dit een opgeblazen ballon op gaat leveren lopen we tegen een probleem op. Wat is dit probleem, en leg uit wat de standaard manier is om dit probleem op te lossen. Gebruik deze methode om deze nieuwe instantie te classificeren.

Opgave 5. Frequent Pattern Mining (16 punten)

Gegeven is een transactionele database D waar elke transactie een itemset heeft:

tid	Items
1	{a, b, c}
2	{a, d, e}
3	{a, c}
4	{d, e}
5	{b, c}
6	{a, c, d, e}
7	{c, d, e}
8	{b, c}
9	{a, c, d, e}
10	{b}

(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) Wat is de maximum grootte (aantal items in X, of $|X|$) van de frequente itemsets. Beantwoord deze vraag voor zowel voor $minsup = 0$ als $minsup = 0.1$.

(c) Vind een paar itemsets (A, B) waarvoor geldt: conf($A \rightarrow B$) = conf($B \rightarrow A$), en $|A| > 1, |B| > 1$

(d) 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. Hierbij geldt: $minsup = 0.3$.

(e) Stel voor de volgende regels de contingency tabel op, en rank de regels a.d.h.v. evaluatie maten. De regels zijn:

$$\{b\} \rightarrow \{c\}, \{a, d\} \rightarrow \{e\}, \text{ en } \{c\} \rightarrow \{d, e\}.$$

De evaluatie maten zijn:

- Support
- Confidence

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.

(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.

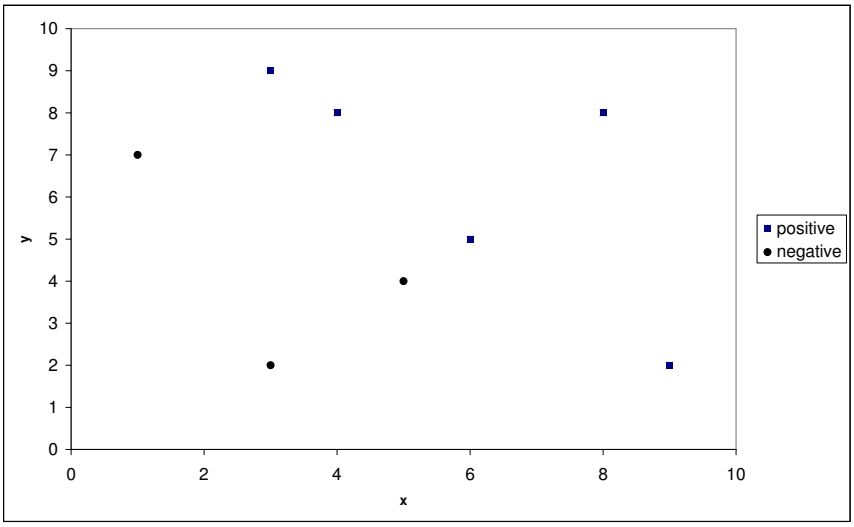
(c) Stel dat de we alleen geïntereseerd zijn in patronen met een minimum support van 10%, en C1 en C2 voldoen aan deze eis. Kan "C1 $\wedge$ C2" een te lage support hebben? Leg uit.

(d) Is er een C1 en C2 denkbaar, volgens de genoemde eisen, zodat "C1 $\wedge$ C2" een hogere Weighted Relative Accuracy heeft dan zowel C1 als C2 afzonderlijk? Verklaar.

(e) Leg uit wat een isometric is, met betrekking tot kwaliteitsmaten in de ROC ruimte.

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) 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.

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

(c) Geef voor attribuut x aan welke drempelwaarden relevant zijn voor het bepalen van de maximale information gain voor een splitsing op x.

(d) Bereken de information gain van de split $x < 4.5$. 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