

## Data Mining (BIKDM5)

Dinsdag 20 december 2005, 13.30-16.30, zaal A.A

Mobiele telefoons zijn niet toegestaan en dienen uit het zicht opgeborgen te worden.

Dit is niet een open boek tentamen. Wel is het toegestaan een "spiekbrief" te gebruiken.

Toegestaan formaat: 1 vel A4, aan een 1 kant bedrukt of beschreven. U dient uw spiekbrief met uw antwoorden in te leveren.

U kunt in totaal 40 punten verdienen door de vragen hieronder te beantwoorden: 10 voor de waar/onwaar-vragen, 15 voor de weet-vragen, en 15 voor de open vragen.

Veel succes!

### 1 Waar/onwaar vragen

Beantwoord de volgende 10 vragen met waar of onwaar. Elk van de volgende vragen is 1 punt waard. Het totale aantal te verdienen punten voor dit onderdeel is 10 punten.

1. Regressie is het voorspellen van een continue waarde als uitkomst van een lineaire som van de attribuutwaarden met geschikt gekozen gewichten.
2. ID3 is een methode voor het induceren van een beslissingsboom die gebruik maakt van het "information gain"-criterium.
3. Covering algoritmes maken gebruik van de 'divide-and-conquer' strategie.
4. Voor het bepalen van de afstand tussen attribuutwaarden is de Manhattan metric ongeschikt.
5. Het verschil tussen prior en posterior probability bij Naive Bayes is dat bij prior probability geen voorkennis/bewijs over de classificeren instantie, alleen geketen naar bestaande instanties; bij posterior probability wordt informatie over te classificeren instantie (= bewijs/voorkennis) ook meegenomen.
6. De *resubstitution error* geeft meestal een te positieve schatting van het getrag van een classifier op ongeziene data.
7. 'Subtree replacement' en 'subtree raising' zijn voorbeelden van mogelijkheden tot prepruning.
8. Bij het definiëren van een beslissingsboom heeft de *information gain*-maat een voorkeur voor attributen met een groot aantal mogelijke waarden.
9. Iedere classifier kan 'kosten-bewijs' worden gemaakt door data samples te genereren met verschillende proporties *yes* en *no* instanties.
10. Bij de .632 bootstrap creëren we een nieuwe dataset door 63.20% van de trainings-instanties te nemen.

### 2 Weet-vragen

Beantwoord de volgende 6 vragen, in fatsoenlijk Nederlands, in 1 tot 3 zinnen. Elke vraag in dit onderdeel van het tentamen is 2,5 punt waard; het totale aantal te verdienen punten voor dit onderdeel is derhalve 15 punten.

1. Welke strategieën zijn er om om te gaan met een situatie waarin een instantie niet door een gegeven verzameling classificatie-regels wordt geclassificeerd?
2. Bespreek voor elk van de volgende activiteiten of het een data mining taak is of niet:
  - (a) Het monitoren van de hartslag van een patiënt met het oog op abnormaliteiten
  - (b) Het groeperen van de klanten van een bedrijf aan de hand van hun *profitability*
  - (c) Het bepalen van de totale verkoop van een bedrijf.
  - (d) Het voorspellen van de uitkomst van het werpen van twee zilvere dobbelstenen.
  - (e) Het voorspellen van de beurskoers van een bedrijf met gebruikmaking van historische gegevens.
3. Beschrijf de TAR2 treatment learner uit het artikel van Menzies en Hu.
4. Hoe gebruikt u numerieke attributen bij *Naive Bayes*?
5. *Instance-based learning* is tijdrovend voor data-sets van realistische afmetingen omdat voor iedere test instantie de gehele training set gescand moet worden. Beschrijf procedures die dit probleem aanpakken.
6. Neem aan dat onze data mining taak bestaat uit het clusteren van de volgende 8 punten (met  $(x, y)$ -coördinaten) in drie clusters:  
A1 (2,10), A2 (2,5), A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C1 (1,2), en C2 (4,9).  
De te gebruiken metriek is de Euclidische metriek. Neem aan dat we, om te beginnen, A1, B1, en C1 aanwijzen als centra van de clusters. Gebruik nu het *k*-means algoritme om te laten zien
  - (a) welke clusters we hebben na het algoritme één ronde te hebben uitgevoerd; en
  - (b) wat de uiteindelijke clusters zijn die het algoritme oplevert.

### 3 Open vragen

Voor elk van de drie open vragen kunt u 5 punten verdienen. Het totale aantal te verdienen punten voor dit onderdeel van het tentamen bedraagt derhalve 15. Schrijf duidelijk en geef tekst en uitleg, maar vermijd overbodige uitweidingen.

1. Beschouw een fictieve dataset met een nominale doelklasse. Neem aan dat de dataset  $n$  instanties heeft. Bekijk de  $n$  *nearest neighbors classifier* voor deze dataset. Dat wil zeggen: de classifier zal de  $n$  *nearest neighbors* van een gegeven test-instantie gebruiken om aan die test-instantie een classificatie toe te kennen.

- (a) Welke classificatie zal aan de test-instantie worden toegekend?  
 (b) Welke andere classifier die u kent geeft in de hierboven beschreven situatie hetzelfde resultaat als de  $n$  *nearest neighbors*?

Bekijk nu een dataset met twee numerieke attributen  $X$  en  $KLASSE$ . Neem aan dat de dataset de volgende instanties bevat, waarbij de eerste component de  $X$ -waarde weergeeft, en de tweede de  $KLASSE$ -waarde: (2,10), (4,7), (7,3), (12,4), (15,8).

(c) Representer deze data in een diagram, met langs de  $X$ -as waarden van  $X$  (uiteenlopend van 0 to 20), en langs de  $Y$ -as de waarden van  $KLASSE$ . Geef de bovenstaande instanties weer. En geef in uw grafiek eveneens aan welke  $KLASSE$ -waarde de  $1$ -*nearest neighbor classifier* toekent aan de waarden van  $X$  die niet expliciet in de dataset genoemd worden (0, 1, 3, 5, 6, 8, etc).

2. Bekijk de volgende trainingsdata, waar iemand "ja" scoort op *achterstand* als hij/zij betalingsachterstand heeft of heeft gehad:

instantie	huiseigenaar	status	inkomen (*1000)	achterstand
1	ja	ongetrouwd	125	nee
2	nee	getrouwd	100	nee
3	nee	ongetrouwd	70	nee
4	ja	getrouwd	120	nee
5	nee	gescheiden	95	ja
6	nee	getrouwd	60	nee
7	ja	gescheiden	220	nee
8	nee	ongetrouwd	85	ja
9	nee	getrouwd	75	nee
10	nee	ongetrouwd	90	ja

Verder is het volgende gegeven voor het *inkomen*

- als achterstand = nee: sample mean = 110  
 sample variance = 2975  
 als achterstand = ja: sample mean = 90  
 sample variance = 25

Gebruik Naive Bayes om de waarde van het *achterstand*-attribuut te voorspellen voor de volgende instantie:

*huiseigenaar*: ja, *getrouwd*: getrouwd, *inkomen*: 120

3. Een abstract probleem rondom de *.632 bootstrap*.

Bekijk een probleem waar geclassificeerd moet worden op een attribuut met twee mogelijke waarden, met een gelijk aantal positieve en negatieve voorbeelden in de data. Neem aan dat de klasse-labels voor de voorbeelden random gegenereerd zijn. De gebruikte classifier is een beslissingsboom (die niet *gepruned* is, etc, en die, met andere woorden, "alles onthoudt"). Bepaal nu de nauwkeurigheid van de classifier aan de hand van de volgende methoden:

- (a) De *holdout* methode, waarbij tweederde-deel van de data gebruikt wordt om te trainen en het resterende eenderde-deel gebruikt wordt om te testen.  
 (b) Ten-fold cross-validation.  
 (c) De *.632 bootstrap* methode.  
 (d) Welke methode geeft de meest betrouwbare evaluatie van de nauwkeurigheid van de nauwkeurigheid van de classifier?