Endterm Tentamen
"Data mining"

- Schrijf op elk blad je naam en studentnummer.
- Bij elke opgave staat hoeveel punten er te verdienen zijn (100 totaal).
- Succes!

1. Het evalueren van de kwaliteit van oplossingen, om zo verschillende oplossingen met elkaar te kunnen vergelijken, is een belangrijk onderdeel van het data mining proces.

   (a) [**8**] Het evalueren van oplossingen voor supervised problemen (bijv. classificatie) is in het algemeen beduidend eenvoudiger dan voor unsupervised problemen (bijv. clustering, detectie van uitbijters). Hoe komt dat?

   (b) [**8**] Wanneer we verschillende clusterings willen evalueren kunnen we onderscheid maken tussen het (meest voorkomende) geval waarbij er geen klasselabels beschikbaar zijn en het (minder voorkomende) geval waarbij er wel klasselabels beschikbaar zijn. Geef voor elk van deze twee gevallen een evaluatiecriterium met een korte beschrijving hiervan.

   (c) [**8**] Stel, nadat je het K-means algoritme hebt losgelaten op je data vind je voor de "sum of squared errors" (SSE) de waarde 0.12. Kun je hieruit concluderen of de data duidelijk geclusterd is of juist niet? Zo ja, geef aan wat je concludeert en waarom. Zo nee, geef aan wat je zou moeten doen om een dergelijke conclusie wel te kunnen trekken.

2. Het bekendste algoritme voor associatie analyse is het Apriori algoritme. Dit algoritme zoekt eerst naar frequente itemsets, d.w.z., itemsets met een "support" groter dan een ingestelde drempel. Vervolgens verzamelt het, gegeven de gevonden frequent itemsets, alle regels met een "confidence" hoger dan een ingestelde drempel.

   Voor het vinden van frequente itemsets maakt het Apriori algoritme gebruik van een genereer-en-tel strategie . Kandidaat itemsets van grootte $k+1$ worden gegenereerd door twee frequent itemsets van grootte k met elkaar te combineren. Een kandidaat itemset wordt verder genegeerd ("gepruned") als één of meer van zijn subsets niet frequent blijkt te zijn.

   Stel je voor dat het Apriori algoritme wordt losgelaten op de data set in onderstaande tabel met "minsup = 30%", d.w.z., elke itemset die in minder dan 3 transacties voorkomt wordt als niet frequent beschouwd.

Transaction ID	Items bought
1	{a,b,c}
2	{a,c,d}
3	{a,b,c}
4	{a,b,d}
5	{a,c,d}
6	{a,b,c}
7	{a,d}
8	{b,c,d}
9	{b}
10	{a,c}

   (a) [**12**] Teken het itemset rooster ("lattice") behorend bij deze data set. Geef elke knoop een label **NG**, **NP**, **NS** of **F**, aansluitend bij de volgende definities.

   - **NG**: de itemset wordt niet gegenereerd als mogelijke kandidaat itemset.
   - **NP**: de itemset wordt als mogelijke kandidaat itemset gegenereerd, maar vervolgens gepruned.
   - **NS**: de itemset wordt als mogelijk kandidaat itemset gegenereerd, vervolgens niet gepruned, maar blijkt na berekening van de "support" toch niet frequent.
   - **F**: de kandidaat itemset is frequent.

   (b) [**8**] Leg uit wat de monoticiteits-eigenschap van "support" inhoudt en hoe die in het vinden van frequente itemsets van pas komt, eventueel aan de hand van onderdeel (a).

   Vervolgens wordt gezocht naar regels met een voldoende "confidence". Ook hier kan weer gebruik worden gemaakt van een plezierige eigenschap, zoals verwoord in het volgende theorema.

Theorema 1. Als de regel $X \rightarrow Y - X$ (Y is de hele itemset, X het precedent, $Y - X$ het consequent) een "confidence" heeft onder de "confidence threshold", dan geldt dat ook voor de regel $X' \rightarrow Y - X'$, met $X' \subset X$ (X' een subset van X).

(c) [**12**] Beschouw de itemset $\{a, b, c\}$ en een "confidence threshold" ("minconf") 0.7. Teken een rooster ("lattice") met alle regels die je kunt maken, waarbij elke volgende laag in het rooster 1 variabele meer in het consequent heeft staan. Geef elke knoop (met minimaal 1 item in het antecedent en 1 in het consequent) een label **NG**, **NC** of **C**, aansluitend bij de volgende definities.

- **NG**: de regel hoeft niet beschouwd te worden omdat Theorema 1 van toepassing is.
- **NC**: Theorema 1 is niet van toepassing, maar na berekening blijkt de regel een "confidence" kleiner dan 0.7 te hebben.
- **C**: de regel heeft een "confidence" groter dan of gelijk aan 0.7.

(d) [**6**] Bewijs Theorema 1. Hint: bereken de "confidence" van de regels waarop Theorema 1 van toepassing is om inspiratie op te doen.

3. Lineair separeerbare classificatieproblemen zijn beduidend eenvoudiger op te lossen dan niet-lineair separeerbare classificatieproblemen. Zo zijn lineair separeerbare problemen gegarandeerd binnen eindige tijd op te lossen met het perceptron algoritme.

(a) [**5**] Geef voor elk van onderstaande Boolse ("Boolean") functies aan of het probleem lineair separeerbaar is.

i. NOT A AND B
ii. A AND B AND C
iii. (B OR C) AND (A OR C)
iv. (A XOR B) AND (A OR B)
v. (A XOR B) OR (A OR B)

Neurale netwerken hebben, in vergelijking tot het standaard perceptron model, één of meer extra lagen met zogenaamde verborgen neuronen. Dit maakt dat ze ook niet-lineair separeerbare problemen aan kunnen. Voor de activatiefunctie van deze verborgen neuronen wordt een sigmoide (S-curve) of een tanh gekozen.

(b) [4] Waarom wordt een sigmoide of tanh en niet, zoals bij het perceptron, een sign functie ($\text{sign}(x) = 1$ als $x > 0$ en $\text{sign}(x) = -1$ als $x < 0$) gekozen?

(c) [4] Waarom heeft het geen zin om in plaats van de sigmoide of tanh een lineaire activatiefunctie te kiezen?

Wanneer een probleem lineair separeerbaar is, is de oplossing meestal niet uniek. De oplossing waarnaar het perceptron algoritme convergeert hangt dan ook ondermeer af van de initialisatie van de parameters (gewichten). Support vector machines geven wel een unieke oplossing.

(d) [6] Welke oplossing is dit? Wat is de motivatie voor deze oplossing?

(e) [6] Hoe kan een support vector machine toch een niet-lineair separeerbaar probleem oplossen?

Je wordt gevraagd voor een bank een classificatieprobleem op te lossen m.b.t. de loyaliteit van klanten. Beschikbaar is een database met 20 inputs, zoals o.a. leeftijd, geslacht, verschillende variabelen die het betaalgedrag van de klant samenvatten. De classificator variabele is binair en geeft aan of de klant wel of niet is overgestapt naar een andere bank. Het aantal objecten (klanten) is 100.000. Je hebt de beschikking over 4 verschillende classificatiemethodes: een beslisboom, een ensemble van beslisbomen, een neuraal netwerk en een support vector machine.

(f) [3] De opdrachtgever hecht veel waarde aan de interpreteerbaarheid van de uitkomst en inzicht in het domein. Voor welke classificatiemethode kies je en waarom?

(g) [6] De opdrachtgever wil het classificatiemodel met name in gaan zetten om de loyaliteit van nieuwe klanten in te schatten. Prestaties worden hierbij belangrijker gevonden dan interpreteerbaarheid. Voor welke classificatiemethode kies je en waarom? Beschouw in je argumentatie de mogelijke aanwezigheid van uitbijters, de heterogeniteit van de inputs en de efficiëntie (zowel hoeveel tijd het kost het model te trainen als ook hoeveel tijd het kost een nieuwe klant te classificeren). Geef niet alleen expliciet aan waarom je een bepaalde methode wel kiest, maar ook waarom een andere niet.

(h) [4] Leg uit wat "Customer Lifetime Value modeling" is en wat data mining hiervoor kan betekenen.